Advances in Water Resources **15** (1992) 75–83

Ground-water models cannot be validated

Leonard F. Konikow

US Geological Survey, 431 National Center, Reston, Virginia 22092 USA

&

John D. Bredehoeft

US Geological Survey, 345 Middlefield Road, MS 439, Menlo Park, California 94025, USA

Ground-water models are embodiments of scientific hypotheses. As such, the models cannot be proven or validated, but only tested and invalidated. However, model testing and the evaluation of predictive errors lead to improved models and a better understanding of the problem at hand. In applying ground-water models to field problems, errors arise from conceptual deficiencies, numerical errors, and inadequate parameter estimation. Case histories of model applications to the Dakota Aquifer, South Dakota, to bedded salts in New Mexico, and to the upper Coachella Valley, California, illustrate that calibration produces a nonunique solution and that validation, *per se*, is a futile objective. Although models are definitely valuable tools for analyzing ground-water systems, their predictive accuracy is limited. The terms *validation* and *verification* are misleading and their use in ground-water science should be abandoned in favor of more meaningful model-assessment descriptors.

INTRODUCTION

The need to calibrate ground-water models has existed as long as ground-water models. In recent years, there has been an increased emphasis on the need to validate ground-water models, driven largely by those engaged in radioactive waste disposal. This has led to institutionalized and publicized programs for verification or validation of hydrogeological models, such as the INTRACOIN, HYDROCOIN, INTRAVAL, and GEOVAL projects. For example, two of the three stated objectives of the HYDROCOIN project are code verification and model validation (Swedish Nuclear Power Inspectorate[21]). The INTRAVAL project was established to evaluate the validity of mathematical models for predicting the potential transport of radioactive substances in the geosphere (Swedish Nuclear Power Inspectorate[22]). It is natural for people who apply ground-water models, as well as those who make decisions based on model results, to want assurance that the model is valid.

It is our intent to approach the question of validation on two levels: (1) the philosophical level, and (2) the practical level of validating a site-specific model. We will argue that, at both levels, validation has no place in hydrology. Although we focus on ground-water flow and transport models, the discussion is applicable to other types of ground-water models, such as geochemical models.

MODELS

The word *model* has so many meanings and is so overused that it is sometimes difficult to know what one is referring to. For this discussion, we define a *model* as a *representation of a real system or process.* To help clarify our discussion, we briefly discuss several types of ground-water models.

We define a *conceptual model* as a *hypothesis for how a system or process operates.* The idea can be expressed quantitatively as a mathematical model. *Mathematical models* are abstractions that replace objects, forces, and events by expressions that contain mathematical variables, parameters, and constants (Krumbein and Graybill,[13] p. 15).

Most ground-water models in use today are *deterministic* mathematical models. Deterministic models are based on conservation of mass, momentum, and energy — that is, on a balance of the various fluxes of these quantities. Originally, the description of the governing processes was the result of great individual insight coupled with experimentation; one good example is Darcy's Law. Experimental laws, such as Darcy's Law, Fourier's Law of thermal diffusion, and Fick's Law of chemical species diffusion, are mathematical statements (or constitutive equations) relating fluxes of mass, momentum, and energy to measurable state variables, such as hydraulic head, temperature, and solute concentration. Deterministic models describe cause and effect relations.

Deterministic ground-water models generally require the solution of partial differential equations. Exact solutions can often be obtained analytically, but analytical models require that the parameters and boundaries be highly idealized. Numerical methods yield approximate solutions to the governing equation (or equations); they require discretizaton of space and time. Within the discretized format one approximates the variable internal properties, boundaries, and stresses of the system. Numerical models relax the idealized conditions of analytical models and are therefore more realistic and flexible; one should remember, however, that numerical methods provide only approximate solutions.

When a numerical algorithm is implemented in a computer code to solve one or more partial differential equations, the resulting computer code can be considered a *generic* model. When the parameters (such as hydraulic conductivity and storativity), boundary conditions, and grid dimensions of the generic model are specified to represent a particular geographical area, the resulting computer program is a *site-specific* model. Generic models are not so robust as to preclude the generation of significant numerical errors when applied to a field problem. If the user of a model is unaware of or ignores the details of the numerical method, including the derivative approximations, the scale of discretization, and the matrix solution techniques, significant errors can be introduced and remain undetected.

ERRORS

The philosophy underlying the use of deterministic ground-water models is: 'given a high degree of understanding of the processes by which stresses produce subsequent responses in a system, the system's response to any set of stresses can be predicted even if the magnitude of the new stresses is outside the range of those historically observed' (Konikow and Patten[10]). Discrepancies between observed and predicted responses of a system are the manifestation of errors in the mathematical model. In applying ground-water models to field problems, there are three sources of error.

One source is conceptual errors — that is, theoretical misconceptions about the basic processes that are incorporated in the model. Conceptual errors include both neglecting relevant processes as well as representing inappropriate processes. Examples of such errors include the application of a model based upon Darcy's Law to materials where Darcy's Law is inappropriate, or the use of a two-dimensional model where significant flow or transport occurs in the third dimension. A second source of error involves numerical errors arising in the equation-solving algorithm. These include truncation errors and numerical dispersion. A third source of error arises from uncertainties and inadequacies in the input data that reflect our inability to describe the aquifer properties, stresses, and boundaries. In most model applications conceptualization problems and uncertainty concerning the data are the most common sources of error.

VERIFICATION, CALIBRATION, VALIDATION, AND PREDICTION

One of the questions in modeling is: *Does the computer code provide an accurate solution to the governing partial differential equation for various boundary value problems?* This is checked by demonstrating that the code gives good results for problems having known solutions. This test is usually done by comparing the numerical model results to that of an analytical solution. Because numerical solutions are sensitive to spatial and temporal discretization, even a perfect agreement only proves that the numerical code *can* accurately solve the governing equations, not that it *will* under any and all circumstances.

Analytical solutions generally require simple geometry, uniform properties, and idealized boundary and initial conditions. The power of the numerical methods is that they relax the simplification imposed by analytical methods and allow the introduction of non-homogeneous, anisotropic parameter sets, irregular geometry, mixed boundary conditions, and even non-linearities into the boundary value problems. Usually, analytical solutions approximating these complexities are unavailable for comparison. The problem is: *Once these complexities are introduced, how does one know the computer code is calculating an accurate solution to the governing equations?* The answer is: *One cannot be sure!* You can do simple tests, such as checking mass conservation and evaluating the global mass-balance error, but in the final analysis you cannot be sure.

Another question in modeling is: Can we adequately describe the internal properties and boundaries of the ground-water system? To determine uniquely the parameter distribution for a field problem would require so much expensive field testing that it is seldom feasible either economically or technically. Therefore, typically we attempt, in effect, to solve a large set of simultaneous equations having more unknowns than equations. It is inherently impossible to obtain a unique solution to such a problem. One attempts to select a set of parameter estimates that yields the *best* solution through model *calibration*. This is done by comparing observations of head or solute concentration to corresponding values calculated by the model. The calibration procedure involves varying parameter values within reasonable ranges until the differences between observed and computed values are minimized. This minimization can be attempted through trial and error adjustments or through some automated inverse or parameter estimation procedure.

The model is considered calibrated when it reproduces historical data within some subjectively acceptable level of coherence — there are no rules other than one's judgement. One does not obtain a unique set of parameters. A poor match suggests (1) an error in the conceptual model, (2) an error in the numerical solution, or (3) a poor set of parameter values. One may not be able to distinguish among the several sources of error. A good match does not prove the validity of the model; because the solution is nonunique, the model can include compensating errors. The model may adequately reproduce historical data, but fail to predict future responses under a new or extended set of stresses. In discussing fundamental problems associated with physically-based hydrologic models, Beven[3] argues that comparisons of predicted and observed hydrographs are a necessary test, but cannot be considered a sufficient test.

In the petroleum industry, model calibration is called *history matching*, which Crichlow[5] (p. 248) defines as the process whereby the existing model data are modified until a reasonable comparison is made with observed data. The term *history matching* more clearly conveys the essence of the modeling process than does the term *calibration*. Petroleum reservoir engineers generally do not attempt to predict reservoir performance for more than one or two times the period of the history match. Thomas[23] (p. 9), after noting the problem of nonuniqueness, made the precautionary observations that (1) generally, 'the longer the matched history period, the more reliable the predicted performance will be'; and (2) the predicted and actual performance must be monitored, and the physical picture of the reservoir must be updated periodically.

Some debate about validation can be attributed to semantics. The terms *verification* and *validation* are often used interchangeably in hydrology. However, there are some who define *verification* as demonstrating the ability of a generic model to solve the governing equation and *validation* as demonstrating the ability of a site-specific model to represent cause and effect relations at a particular field area. Regardless, both words imply the authentication of both the truth and accuracy of the model. (Based on definitions in dictionaries and synonyms in thesauruses, this meaning is inferred by laymen as well as by scientists.) If a model is validated, it follows that the model is valid. A logical inference is that a model certified as valid can make reliable predictions, without qualifications. Yet, accepting that one needs to calibrate a site-specific ground-water model is tantamount to acknowledging the impossibility of validating such a model.

One purported goal of validation is to produce confidence in the ability of a model to make reliable predictions. We pose the following additional question: *Will a supposedly site-specific ground-water model provide accurate predictions?* The ability of a model to reproduce what has been observed (which is the outcome of model calibration or history matching) enables the analyst to understand the ground-water system being analyzed. Examples can certainly be cited of cases where better understanding leads to better management decisions. However, Winograd[27] states that strong philosophical arguments exist for believing that explanation and prediction in the natural sciences are not symmetrical — that is, understanding a process, and being able to model it, does not mean that prediction is attainable. There are cases in hydrology, and in science in general, where our understanding of processes may be great, but predictive ability low, and other cases where understanding is minimal, but predictive accuracy very high. In any event, the accuracy of the prediction cannot be assessed until after the predicted period of time has passed.

PHILOSOPHY OF VALIDATION

Validation in science is a question of great interest to philosophers of science, and central to how we, as scientists, view what we do. The question of validation in science must ultimately be asked at a philosophical level.

There are two principal schools of philosophical thought on this issue. One school, called positivism, holds that '. . . theories are confirmed or refuted on the basis of critical experiments designed to verify the consequences of the theories' (Matalas *et al.*[17]) One of the principal proponents of positivism is Thomas Kuhn.[14] A second school, espoused by Karl Popper,[18] argues that 'as scientists we can never validate a hypothesis, only invalidate it.' Popper elegantly points out the incompatibility of these two schools. In the end one must make a personal choice which school one believes to be correct.

We believe that many, if not most, present-day scientists who have considered these issues find themselves in Popper's camp; we do. The following is a quotation from a recent book, *A Brief History of Time*, by the noted physicist Stephen Hawking[7] (p. 10; reproduced with permission of publisher, Bantam Books):

> Any physical theory is always provisional, in the sense that it is only a hypothesis: You can never prove it. No matter how many times the results of experiments agree with some theory, you can never be sure that the next time the result will not contradict the theory. On the other hand, you can disprove a theory by finding even a single observation that disagrees with the predictions of the theory. As philosopher of science

> Karl Popper has emphasized, a good theory is characterized by the fact that it makes a number of predictions that could in principle be disproved or falsified by observation. Each time new experiments are observed to agree with the predictions the theory survives, and our confidence in it is increased; but if ever a new observation is found to disagree, we have to abandon or modify the theory. At least that is what is supposed to happen, but you can always question the competence of the person who carried out the observation.
>
> In practice, what often happens is that a new theory is devised that is really an extension of the previous theory. For example, very accurate observations of the planet Mercury revealed a small difference between its motion and the predictions of Newton's theory of gravity. Einstein's general theory of relativity predicted a slightly different motion from Newton's theory. The fact that Einstein's predictions matched what was seen while Newton's did not, was one of the crucial confirmations of the new theory. However, we still used Newton's theory for all practical purposes because the difference between its predictions and those of general relativity is very small in the situations that we normally deal with. (Newton's theory also has the great advantage that it is much simpler to work with than Einstein's!)

All science is a progress report: Einstein's general theory of relativity replaced Newton's theory of gravity, plate tectonics has made enormous changes in how we interpret geology, and research into ground-water transport may well change some basic hydrogeologic thinking in the next several years, making some previous work obsolete. Site-specific ground-water models are elements of applied earth science — in effect, an agglomeration of multiple hydrogeologic theories. As such, they are subject to improvement via invalidation, but cannot be proven valid. Validation cannot add to the fund of knowledge.

OPERATIONAL DEFINITIONS

In discussing validation of models used for performance assessment of high-level nuclear waste repositories, Davis *et al.*[6] note that the US Nuclear Regulatory Commission[26] defined validation as the process of obtaining 'assurance that a model, as embodied in a computer code, is a correct representation of the process or system for which it is intended.' They also note that the US Department of Energy[25] defines validation as 'a process whose objective is to ascertain that the code or model indeed reflects the behavior of the real world.' The International Atomic Energy Agency[8] states that models are validated when it is confirmed that the models 'provide a good representation of the actual processes occurring in the real system.' These definitions are concerned with providing assurances or building confidence that the model represents reality.

Davis *et al.*[6] point out that these definitions are inconsistent with Popper's view of scientific validation. Davis *et al.*[6] argue that the process of model validation is useful in the decision making or regulatory process. They go on to point out 'the definitions indicate only that assurance be provided that the models are *adequate* representations of the real system. Defining what is "adequate" will, in the end, be a subjective decision made by the regulator.'

We believe these definitions were a poor choice, even for operational purposes. They are certainly contrary to the prevailing scientific and layman's view of validation. They tend to lend undue credibility to a process that even Davis *et al.*[6] point out is, in the end, inherently subjective. Petroleum engineers' terminology of 'history matching' is a much more realistic and accurate description of what is done during this so-called 'validation' process.

GROUND-WATER MODEL VALIDATION: A PRACTICAL PERSPECTIVE

In practice, validation is attempted through the same process that is typically identified as calibration. The International Atomic Energy Agency[8] states, 'Validation is thus carried out by comparison of calculation with observations and experimental measurements.' However, as previously discussed, the non-uniqueness of model solutions means that a good comparison can be achieved with an inadequate or erroneous model. Also, because the definition of 'good' is subjective, under the common operational definitions of validation, one competent and reasonable scientist may declare a model as validated while another may use the same data to demonstrate that the model is invalid. In science and engineering, such an operational definition would not appear to be meaningful.

Some attempts to render this comparative approach to model validation more rigorous are based on split sampling. This approach in ground-water studies is patterned after a verification approach used in watershed modeling. With this procedure a model is calibrated using only one part of the historical record that contains one or more events that characterize the response of the system. The model is then used to reproduce another part of the historical record that is independent of the data used in the calibration; sometimes this second phase is called a *verification* phase.

Split sampling is usually a weak procedure when applied to ground water. The time scale on which ground-water systems respond is much longer than that of surface-water systems; it is rare in ground-water analysis to have a historical record long enough to be broken into independent data sets. If split sampling is used, it is necessary to show that stresses during the calibration period do not influence the system response during the verification period. Rarely can such independence be demonstrated for a large-scale aquifer system.

CASE HISTORIES

Matalas *et al.*[17] (p. 122) note that 'the positivistic view continues to pervade the conduct of hydrologic research — strong emphasis is placed on such pursuits as "model verification" on the basis of the data, i.e., the empirical evidence alone (even though "model verification" is not pursued strictly in accordance with the notion of 'criticality' or formulated in terms of critical experiments), and advances in hydrology are judged primarily in terms of predictive capability.' It is reasonable to ask: *is there evidence that the current practice of calibration and verification lead to a reliable predictive capability?* Several authors have examined this question; there is little evidence to support high confidence in long-term model predictions (see, for example, Lewis and Goldstein,[15] Konikow and Patten,[10] Konikow and Person,[11] Alley and Emery[1] and Konikow[9]).

In order to illustrate our points we would like to discuss several real examples of model application to field problems. In these examples we will emphasize the difficulty of selecting the appropriate conceptual model. Selecting the right conceptual model is an *a priori* decision by the analyst, and is usually based upon his understanding of the system. Often several conceptual models are possible; the empirical data can be fitted equally well to the several models. Using the operational definitions one would consider the model 'validated'. However, the long-term consequences of the choice of conceptual model are quite different. Selecting the appropriate one is a critical decision.

The Dakota Aquifer: South Dakota

The Dakota Aquifer in South Dakota is sometimes viewed as the prototype artesian aquifer.[4] It was studied by Darton at the turn of this century. In analyzing the Dakota Aquifer the question arises as to whether the Cretaceous Shale confining layer that overlies the aquifer is sufficiently permeable to provide a significant component of the total flow through this confining layer.

A 40-hour pumping test was run at Wall, South Dakota. The data from this test fitted quite well to the so-called 'Theis' solution for the response of a well to pumping. The Theis solution assumes impermeable confining layers both above and below the aquifer of interest. The data are fitted equally well to the Hantush 'modified leaky aquifer' solution. The Hantush solution allows for transient flow through the confining layers. In order to see a significant departure from the Theis solution, the pumping test would have to be run more than 1000 years. The choice of which solution to use to analyze the Wall data is an *a priori* decision by the analyst and depends upon his conceptual model of the system. Experienced hydrologists have analyzed the Wall data using both solutions. The Wall data are insufficient to invalidate one or the other model. In terms of the operational definitions given above, both models are 'validated' by the Wall data.

The long-term response of the Dakota Aquifer is to a large extent controlled by flow and storage in the confining layers. If the Dakota system is analyzed at a regional scale, then one sees that approximately 50 per cent of the recharge occurs through the confining layers, and almost 80 per cent of the discharge from the system is through the confining layer.[4] If one's objective is only to predict the short-term response of the well at Wall, then the Theis solution is adequate. If on the other hand one's objective is to predict the long-term response of the Dakota system, then neglecting the flow through the confining layers makes a big difference in the predicted system response.

Bedded salt at WIPP site: New Mexico

The Waste Isolation Pilot Plant (WIPP) is designed to be a mined repository in bedded salt in southeastern New Mexico for the disposal of transuranic wastes. It is now recognized that the Salado salt at the WIPP facility has approximately 1–3 per cent intergranular porosity. This pore space is filled with brine. The brine is observed to move into the underground workings.

A number of experiments at WIPP have been undertaken to describe the brine movement phenomena. There are two proposed hypotheses to describe the brine movement:

1. The first hypothesis is that brine in the pore space is released by deformation that accompanies the creation of any opening in the salt.
2. The second hypothesis is that the salt, even though it is plastic, has continuous filaments of connected brine, and that flow in the system can be decribed by Darcy's law.

Most of the experiments, even experiments well into the far field, are adequately fitted by the Darcy flow model. The data are insufficient to invalidate this model. The permeabilities one derives from the Darcy model are very low — 10^{-21} to less than $10^{-22}\,m^2$. The long-term response of the repository may be different depending upon which the of the two hypotheses is correct. Unfortunately the data may not be sufficient to invalidate one or the other of the models.

Coachella Valley: California

To illustrate how ground-water models are typically calibrated, validated, and used to make predictions, we present a summary of one case for which a postaudit that examined the accuracy of predictions was published.[12]

The upper Coachella Valley is the northern part of the Imperial Valley of California, and includes an area of approximately 300 mi^2 in Riverside County. Ground-water development in the valley was small before the late 1930s, when withdrawals were about 5,000 acre-ft/yr. Since 1940, development has increased; during the period 1968–73, pumpage averaged about 49,000 acre-ft/yr.[20] In 1973, artificial recharge was started in the basin to counteract declining ground-water levels; in some places, the declines exceeded 100 ft. Water for artificial recharge was available then from the Colorado River Aqueduct.

Swain[20] applied deterministic, two-dimensional, finite-element, ground-water flow and solute-transport models to predict the effects of artificial recharge on both water levels and water quality. Swain[20] presents the details of both the numerical methods and the site-specific models. We herein restrict our remarks to results of the flow model.

A steady-state flow model was calibrated first to estimate those hydrogeologic factors controlling heads before any significant man-induced stresses occurred. Tyley[24] had suggested 1936 as a period that best represented natural equilibrium conditions for which sufficient data were available to permit a reasonable simulation of the system. Swain[20] estimated steady-state recharge, discharge, underflow, and the areal distribution of transmissivity based upon the results of this calibration. The calculated ground-water levels were within 4 ft of the 1936 measured water levels throughout the model area — a good steady-state calibration.

Next, a transient-flow model was calibrated for the period 1936–68, which used as inputs (1) estimated annual net pumpage, (2) estimated annual recharge, and (3) initial water levels and transmissivity values from the calibrated steady-state flow model. The areal distribution of the storage coefficient was adjusted during the transient calibration. This adjusted set of storage coefficient values yielded the best fit to the transient water-level changes. Water-level declines during this period exceeded 50 ft over two-thirds of the study area. The differences between the measured and calculated water-level declines were less than 10 ft in more that 90 per cent of the area and less than 5 ft in more than two-thirds of the area; this would usually be considered a good calibration in this type of study. Figure 1 shows a comparison of measured water-level declines in selected wells with output from the calibrated model. Swain[20] stated that the similarity in the shape of measured and calculated hydrographs for the 1936–68 period suggests that the model closely represents the response of the ground-water system.

The accuracy of the calibrated model was then tested by simulating the period 1968–74 using annual pumpage estimates for that period, but without additional adjustments of the model parameters. The agreement between observed and calculated water levels during this *verification* period was close, as shown in Fig. 1. Swain[20] concluded: 'The results of this run verified the model and the chosen parameter values within the acceptable limits of cost and time. . . . Having verified that the parameters chosen were reasonable and that the model was capable of duplicating the response of the aquifer, it is now possible to use the model to predict water levels from projected pumpage and (or) projected artificial recharge.'

Ground-water pumping rates for the modeled area were projected for the period 1974–80. The extrapolations suggested that pumping would increase throughout the valley, with the largest increases of up to 45 per cent in the area of Palm Springs. The model predicted that, in 1980, the water level would rise by more than 60 ft near the area of artificial recharge while declining in more than two-thirds of the rest of the valley. Predicted declines for the period would exceed 20 ft in part of the area.

A comparison of observed and predicted water-level changes for 92 wells in the valley for the period 1974–80 is presented as a scatter diagram in Fig. 2. The diagonal solid line represents a perfect fit between observed and predicted changes. For most of the valley where there are observation wells water-level declines were overpredicted. The observed and predicted changes used to plot Fig. 2 have a correlation coefficient of 0.75.

Figure 3 presents a frequency distribution (histogram) of the errors, defined as predicted minus observed water-level change. The frequency distribution is skewed; water levels did not decline as much as predicted especially near ephemeral streams that enter the valley. The errors for all wells have a mean of -8.8 ft, a median of -5.6 ft, a standard deviation of 17.7 ft, and a range of -96 to $+15$ ft. (For comparison, the observed water-level changes have a mean of -5.3 ft, a median of -12.8 ft, a standard deviation of 25.0 ft, and a range of -30 to 92 ft.) The mean error suggests a bias toward overpredicting the magnitude of water-level declines; the wide range in errors indicates some lack of precision (in the sense of a large spread of errors about the mean).

The greatest errors are near tributary canyons that enter the main valley. Ephemeral streams in these canyons have floods that recharge the aquifer in the main valley. The flow and recharge from these creeks during the 1974–80 period turned out to be significantly larger than the long-term average; in particular, 1978–80 were especially wet years. In this desert area, the streamflow in these tributary valleys is highly variable. For example, during this period, the annual discharge in Palm Canyon Creek ranged from 15 acre-ft in 1975 to 35,000 acre-ft in 1980. The data plotted in Fig. 2 show distinctly separate clustering for those wells in or near the three major tributary valleys.

The dominant source of predictive error in the model is attributable to an erroneous assumption about the magnitude of the recharge from the tributary streams in the area. From a different perspective, one might argue that the calibration and verification periods were simply too short to capture the variability in

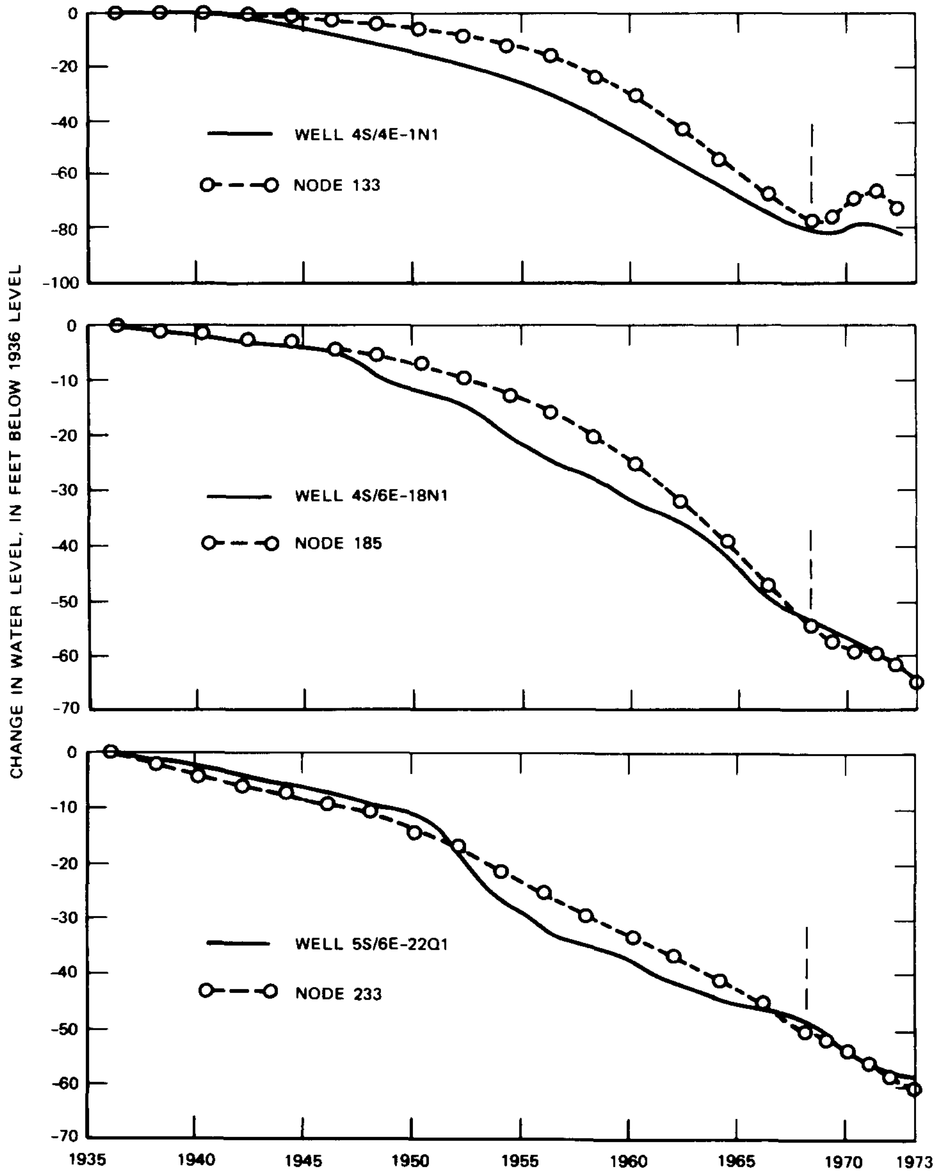

Fig. 1. Comparisons of hydrographs generated by the flow model (dashed lines) versus historical measured water levels (solid lines). (From Swain[20], Fig. 13.) Vertical dashed lines indicate start of verification phase of calibration period (1968–73), as indicated by Swain.[20]

the natural recharge. One can suggest either (1) that the conceptual model was inadequate in not adequately encompassing the highly variable nature of the recharge in this desert environment, or (2) that the calibration period was simply too short (i.e. the data were inadequate). Either view of the model error can be argued. The result is the same — *the model did not provide accurate predictions.*

This case history points out the deficiencies of the usual single-valued prediction. There is always uncertainty in making a prediction; this is especially true with ground-water models where the parameter estimates are non-unique. Predictions should be cast in a probabilistic framework with confidence limits bounding the predicted response. For example, the need for such a probabilistic framework is recognized in US Environmental Protection Agency regulations for radioactive waste disposal and is being implemented in the model analyses being conducted for performance assessment at the WIPP site.[2] However, this probabilistic approach is rarely followed in the multitude of ground-model applications to other sites and problems. This is an area in ground-water analysis that needs additional research into alternative approaches and more encouragement for application.

SYNTHESIS AND CONCLUSIONS

Given the uncertainty in conceptualization and parameter estimation that is inherent in ground-water models, *how can we validate that the model is correct?* Our view is that of Popper's:[18] *We cannot validate, we can only invalidate.* This, we believe, obligates us as scientists to perform a

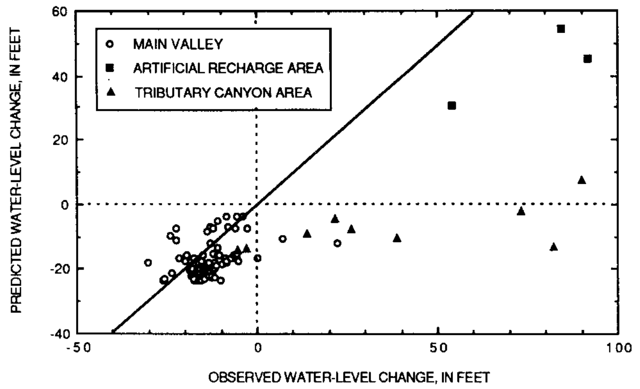

Fig. 2. Relation between predicted and observed water-level changes in the upper Coachella Valley, California, 1974–80. Solid diagonal line shows where predicted equals observed values. (From Konikow and Swain,[12] Fig. 5.)

critical set of experiments in an attempt to test, or invalidate, our model (or hypothesis). Our understanding only increases when we falsify a hypothesis (model) and advance to a new, more encompassing, hypothesis (model). The modeling study of the upper Coachella Valley, California, provides an example of a case for which a test of a model's predictions (in the form of a postaudit) led to an increased understanding of the hydrologic system.

A close examination of many model verification and validation studies reported in the literature indicates that what the investigators have done is to history match their models and in the process estimate parameter distributions, stresses on the system, and boundary and initial conditions. They imply that the resulting good fit constitutes validation of the model, which further implies its utility as a predictive tool. Such logic incorporates circular reasoning and begs the philosophical questions of model validation. If this were merely an argument over semantics, one would conclude that it is not a serious problem. However, many people, especially the public, will put too much faith in models that have the label *verified* or *validated.* Much professional effort is being devoted to validation; it is costing more than a semantic ambiguity is worth. The effort spent on model validation would be better spent on developing a more complete understanding of the particular hydrogeologic system or problem of interest. Expanding on the concern of Rogers:[19] *In focusing on model validation, the analyst is likely to learn more about the model than about the system being modeled or about useful policy implications.*

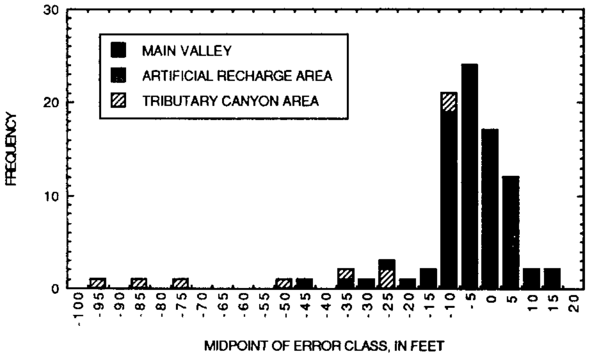

Fig. 3. Histogram showing frequency distribution of errors in water-level prediction for the ground-water flow model of the upper Coachella Valley, California, 1974–80. (From Konikow and Swain,[12] Fig. 6).

The nuclear-waste industry drives much of the present effort devoted to model validation. One problem in using hydrogeological models for decisions in radioactive waste disposal is that the period for prediction (10,000 years is commonly cited) is far beyond any period of observation; history matching of the sort done in petroleum engineering is impossible. In designing a nuclear-waste repository we will need to know the basic processes operating as well as possible; only our fundamental understanding of these processes will make possible defensible long-term predictions. Malone[16] states that the 'lack of validated models for predicting geologic and hydrologic processes over 10,000 years' is a major liability of the Yucca Mountain, Nevada, proposed site for a high-level nuclear-waste repository. Winograd[28] counters that the required models 'though essential for guiding research, testing of worst-case scenarios, and eliminating marginal waste-disposal sites — cannot readily be validated or, perhaps, even calibrated.' Winograd[28] believes that decisions such as selection of waste-disposal sites 'must rest on technical judgement, not solely on the availability of "validated models." '

Our view of the selection of a waste repository is much like Winograd's. It is naive to believe that we will somehow validate a computer model so that it will make accurate predictions of system responses far into the future. In a sense, emphasizing validation deceives society with the impression that, by expending sufficient effort, uncertainty can be eliminated and absolute knowledge be attained. Society continually makes operational decisions in the face of uncertainty. These descisions are based upon judgements about future risks and consequences. Nuclear waste disposal is no different; in the final analysis, society will make a judgement concerning the prudence of what is proposed. We believe society will demand a consensus from the responsible scientific community that the actions being proposed are reasonable. This does not mean that our models were somehow validated; rather, the relevant problems have been investigated and we have assured ourselves that they do not pose unreasonable risks.

If the models cannot be validated, why are they useful? Models provide a tool for critical analysis. They are a means to organize our thinking, test ideas for their reasonableness, and indicate which are the sensitive parameters. They point the way for further investigation. They help formulate critical experiments with which to test hypotheses. Often the systems we deal with are complex, sometimes so complex that our intuition concerning how a particular system will respond to stress is not very good. We are commonly surprised by model outputs; they provide new insights that we would not get otherwise. They serve to sharpen our professional judgement. In the end, action concerning waste disposal will be a judgement; a professional judgement by the scientific community and a judgement by society.

What is usually done in testing the predictive capability of a model is best characterized as calibration or history matching; it is only a limited demonstration of the reliability of the model. We believe the terms *validation* and *verification* have little or no place in ground-water science; these terms lead to a false impression of model capability. More meaningful descriptors of the process include *model testing, model evaluation, model calibration, sensitivity testing, benchmarking, history matching*, and *parameter estimation.* Use of these terms will help to shift emphasis towards understanding complex hydrogeological systems and away from building false confidence into model predictions.

ACKNOWLEDGEMENTS

Discussions with N.C. Matalas, I.J. Winograd, A.M. Shapiro, and R.C. Ewing were very helpful and their review comments are appreciated.

REFERENCES

1. Alley, W.M. & Emery, P.A., Groundwater model of the Blue River Basin, Nebraska — Twenty years later. *Jour. Hydrology*, **85** (1986) 225–49.
2. Bertram-Howery, S.G., Marietta, M.G., Rechard, R.P., Swift, P.N., Anderson, D.R., Baker, B.L., Bean, J.E. Jr, Beyeler, W., Brinster, K.F., Guzowski, R.V., Helton, J.C., McCurley, R.D., Rudeen, D.K., Schreiber, J.D., & Vaughn, P., *Preliminary comparison with 40 CFR Part 191, Subpart B for the Waste Isolation Pilot Plant, December 1990.* Sandia National Laboratories Report SAND90-2347, Albuquerque, New Mexico, 1990.
3. Beven, K., Changing ideas in hydrology — The case of physically-based models, *Jour. Hydrology*, **105** (1989) 157–72.
4. Bredehoeft, J.D., Neuzil, C.E., & Milley, P.C.D., Regional flow in the Dakota Aquifer: A study of the role of confining layers, U.S. Geol. Survey Water-Supply Paper 2237, 1983.
5. Crichlow, H.B., *Modern reservoir engineering — A simulation approach.* Prentice-Hall, Inc., Englewood Cliffs, NJ, 1977.
6. Davis, P.A., Olague, N.E., & Goodrich, M.T., *Approaches for the validation of models used for performance assessment of high-level nuclear waste repositories.* Sandia National Laboratories SAND90–0575, Albuquerque, New Mexico, 1991.
7. Hawking, S.W., *A brief history of time: From the big bang to black holes.* Bantam Books, New York, 1988.
8. International Atomic Energy Agency, *Radioactive waste management glossary*, IAEA-TECDOC-264, International Atomic Energy Agency, Vienna, 1982.
9. Konikow, L.F., Predictive accuracy of a ground-water model — Lessons from a postaudit, *Ground Water*, **24**(2) (1986) 173–84.
10. Konikow, L.F. & Patten, E.P. Jr., Groundwater forecasting. In *Hydrological Forecasting*, ed. M.G. Anderson & T.P. Burt. John Wiley and Sons Ltd, New York, 1985.
11. Konikow, L.F. & Person, M.A., Assessment of long-term salinity changes in an irrigated stream-aquifer system. *Water Resources Research*, **21**(11) (1985) 1611–24.
12. Konikow, L.F. & Swain, L.A., Assessment of predictive accuracy of a model of artificial recharge effects in the upper Coachella Valley, California. In *Selected Papers on Hydrogeology from the 28th International Geological Congress*, vol. 1, ed. E.S. Simpson & J.M. Sharp, Jr. Internat. Assoc. of Hydrogeologists, Verlag Heinz Heise, Hannover, Germany, 1990.
13. Krumbein, W.C. & Graybill, F.A., *An introduction to statistical models in geology.* McGraw-Hill, New York, 1965.
14. Kuhn, T.S., *The structure of scientific revolutions.* University of Chicago Press, Chicago, Illinois, 1962.
15. Lewis, B.D. & Goldstein, F.J., *Evaluation of a predictive ground-water solute-transport model at the Idaho National Engineering Laboratory, Idaho.* U.S. Geol. Survey Water-Resour. Inv. 82-25, 1982.
16. Malone, C.R., The Yucca Mountain Project: Storage problems of high-level radioactive wastes, *Environ. Sci. Technol.*, **23**(12) (1989) 1452–3.
17. Matalas, N.C., Landwehr, J.M. & Wolman, M.G., Prediction in water management. In *Scientific basis of water-resource management.* National Research Council, National Academy Press, Washington, DC, 1982.
18. Popper, Sir Karl, *The logic of scientific discovery.* Harper and Row, New York, 1959.
19. Rogers, P., On the choice of the 'appropriate model' for water resources planning and management, *Water Resources Research*, **14**(6) (1978) 1003–10.
20. Swain, L.A., *Predicted water-level and water-quality effects of artificial recharge in the upper Coachella Valley, California, using a finite-element digital model.* U.S. Geol. Survey Water-Resour. Inv. 77–29, 1978.
21. Swedish Nuclear Power Inspectorate, *The International HYDROCOIN Project — Background and Results.* OECD, Paris, 1987.
22. Swedish Nuclear Power Inspectorate, *The International INTRAVAL Project — Background and Results.* OECD, Paris, 1990.
23. Thomas, G.W., *Principles of hydrocarbon reservoir simulation.* IHRDC Publ., Boston, 1982.
24. Tyley, S.J., *Analog model study of the ground-water basin of the upper Coachella Valley, California.* U.S. Geol. Survey Water-Supply Paper 2027, 1974.
25. U.S. Department of Energy, *Environmental assessment — Yucca Mountain site, Nevada Research and Development Area, Nevada.* DOE/RW-0073, Vol. 2, U.S. Department of Energy, Office of Civilian Radioactive Waste Management, Washington, DC.
26. U.S. Nuclear Regulatory Commission, *A revised modelling strategy document for high-level waste performance assessment.* U.S. Nuclear Regulatory Commission, Washington, DC.
27. Winograd, I.J., *Archaeology and public perception of a transscientific problem — disposal of toxic wastes in the unsaturated zone.* U.S. Geol. Survey Circular 990, 1986.
28. Winograd, I.J., The Yucca Mountain project: Another perspective, *Environ. Sci. Technol.*, **24** (9) (1990) 1291–3