United States
Environmental Protection
Agency

Robert S. Kerr Environmental
Research Laboratory
Ada OK 74820

Research and Development

EPA/600/S2-89/040 Sept. 1989

EPA

Project Summary

The Establishment of a Groundwater Research Data Center for Validation of Subsurface Flow and Transport Models

Paul K. M van der Heijde, Wilbert I. M. Elderhorst, Rachel A. Miller, and Manjit F. Trehan

The International Ground Water Modeling Center has established a Groundwater Research Data Center that provides information on datasets resulting from publicly funded field experiments and related bench studies in soil and groundwater pollution and distributes datasets for testing and validation of models for flow and contaminant transport in the subsurface. To fulfill its advisory role, the Data Center analyzes information and documentation resulting from field and laboratory experiments in the saturated and unsaturated zones and evaluates the appropriate datasets for their suitability in model testing and validation. To assure consistency in the analysis and description of these datasets and to provide an efficient way to search, retrieve, and report information on these datasets, the Center has developed a computerized data directory, SATURN, programmed independently from any proprietary software. As secondary users of such data are highly interested in information about the assessment of data quality, a primary concern of the Center is the evaluation and documentation of the level of quality assurance applied during data acquisition, data handling, and data storage. In addition to providing referral services, the Data Center distributes on an "as-is" basis, selected, high-quality datasets described in the data directory. The datasets of concern represent different hydrological, geological, and geographic-climatic settings, pollutant compositions, and degrees of contamination.

This Project Summary was developed by EPA's Robert S. Kerr Environmental Research Laboratory, Ada, OK, to announce key findings of the research project that is fully documented in a separate report of the same title (see Project Report ordering information at back).

Introduction

The ability to predict accurately the transport and fate of potential contaminants is critical to the success of most groundwater regulations. Attempts at protecting the integrity of an aquifer or engineered facility through monitoring of groundwater quality alone often are ineffective alternatives to predictive modeling. Thus, development and adoption of methods for predicting pollutant transport and fate in the saturated and unsaturated zones of the subsurface are key elements of the EPA's groundwater research strategy. The development and accuracy of such predictive capabilities cannot take place without an equally significant effort in subsurface characterization.

With the growing availability and use of subsurface flow and transport models,

concerns regarding their validity and accuracy has increased. Model testing, or more specifically model validation, provides model users, decision makers, policy makers, and legal authorities with information on a model's performance characteristics—information needed to judge the usefulness of the model results for their problem assessments.

The Groundwater Review Committee of EPA's Science Advisory Board concluded that regardless of the type of model chosen, increased emphasis should be given to field testing and field validation. Data generated in association with remedial action and monitoring Superfund sites may be used to fulfill model validation requirements. The Review Committee commented that these data should be made available for use by other investigators. The Review Committee also found that the conclusions of many publicly funded research efforts are based on data not available for peer review. Therefore, the Committee recommended that databases from field research projects be made readily available to other groups.

No institution has existed for rapidly locating and searching soil water and groundwater research databases or for standardizing data integrity and documentation of research datasets. Existing centralized database facilities for groundwater resource management do not provide the detail and quality of data required to successfully complete research on contaminant transport and fate. In many research projects, the lack of rapid access to these data causes delays and money unnecessarily spent, resulting in many incomplete model validation initiatives. The groundwater research strategy prepared by the U.S. Environmental Protection Agency (USEPA) and the National Center for Ground Water Research states that the data accumulated through Agency-funded research will be made available to the Agency and to the user community through information transfer. A central data clearinghouse could acquire and distribute such data in error-free, machine-usable form, efficiently and economically.

In addressing this need, the Holcomb Research Institute of Butler University, with support from USEPA, has established the Ground Water Research Data Center within the framework of the International Ground Water Modeling Center (IGWMC). The new Data Center provides information and referral services regarding datasets resulting from publicly funded field and laboratory research on soil and groundwater pollution. In addition, the Data Center has established procedures for selecting, evaluating, documenting, and redistributing such datasets. Creation of the Data Center is expected to lead to additional protocols for error checking, documentation, accessing, and transferring this kind of research data, and for acknowledging the rights that researchers have vested in their data.

Project Approach

The project consisted of two phases: (1) determination of the scope and design of the Data Center, and (2) development of facilities and implementation of operational procedures and organizational framework.

The first phase consisted of five elements: analysis of data needs and potential users; survey and analysis of existing datasets; assessment of quality assurance (QA) requirements; determination of computer and other facilities for an operational data center; and operational design of the Data Center.

The analysis of soil and groundwater research data needs and the identification of potential users of high-quality, well-documented datasets provided guidance, justification, and motivation for the development of the Data Center. To determine the required level-of-effort and to obtain baseline information for the design of the Data Center facilities, the availability and status of a number of groundwater datasets resulting from publicly funded research have been evaluated. Current practices in collecting, handling, storing, documenting and distributing these datasets have been studied.

Other data centers utilizing high-quality environmental research and monitoring datasets have been contacted to benefit from their experience in such areas as dataset acquisition, data handling, and quality assurance procedures. Specifically, issues related to the invested rights of researchers involved in the data collection have been discussed.

Quality assurance (QA), an essential task for a central data distribution facility, must be incorporated on two levels: (1) the quality of the datasets of interest needs to be determined and documented; and (2) adequate quality assurance procedures need to be established for the operation of the Data Center in such areas as dataset evaluation, referral, management, and transfer.

To determine the level of detail required for the Data Center in the evaluation of the quality of prospective datasets, an inventory has been mac standards and current accepted prac as documented in the open literature technical guidance of regula agencies.

Based on the findings in phase 1 institutional structure for the Data C has been determined and the data framework created. Two types of base have been developed: (directory-type or referral database taining descriptive information on sets available from the Data Cent from other sources; and (2) a data containing the datasets selected for bution by the Data Center. Inform resulting from the dataset surve phase 1 has been incorporated i referral database.

Arrangements have been mac protect dataset integrity in their tr from their generators to the Data C and from the Data Center to seco users. Furthermore, quality assu procedures have been implemente data handling, storing, archiving backup. Different levels of implen tion are distinguished, dependent c quality and extent of the dataset level of documentation, anc importance of the data. Technical s for format and transfer medium, an limited extent for the analysis of the will be provided; the level of supp depend on the implementation selected. Policies have been dev regarding such issues as propr rights, conditional use, potential liat and other legal and ethical issues.

As a part of the IGWMC, the Center's activities will be subj annual review by the IGWMC Board and the International Tec Advisory Committee (ITAC).

Groundwater Research Dat

Data on groundwater qualit quantity are characterized in bo spatial and temporal domains. Tw types of data are distinguished specific data, and generic, site-ind ent data. It should be noted that report the term groundwater is u water in both the saturated ar saturated zones of the aquati surface.

Certain kinds of site-specific d constant for the time period und sideration, but may vary from loc location. Other site-specific data show a significant time-depend havior. Collection of such data erally aimed at identifying regio terns during a certain time peric

tudying the time variability at specific ocations. These objectives of site-pecific data collection may change luring the operation of the data collection etwork, due to changes in management eeds, technology, and institutional rrangements. Subsequently, the design nd operation (when and where to ample or measure, and which variable to neasure) may be altered. Such variability ertainly applies to research data net-vorks, which are often project-oriented nd of relatively short duration.

Because water in the underground often moves quite slowly, abiotic or biotic ransformations may represent significant ttenuation processes in the transport nd fate of pollutants. The presence of uch processes results in a significant ncrease in data requirements for the pre-lictive analysis of water quality. Much of his additional data is generic and can be stablished off-site in controlled labora-ory or field experiments in combination vith relevant site characteristics. Such eneric, site-independent data on specific hemicals are increasingly available from esearch on the basic processes that overn contaminant transport and fate, nd are crucial for successful application f computer-based prediction techniques in specific hydrogeologic environments.

At the beginning of many research projects requiring data acquisition, the establishment of efficient data management practices is often more difficult than anticipated. Traditionally, researchers have had almost total control over the form and documentation of their data; even contractual requirements for data in machine-processible form have had little effect on the ultimate availability and utility of most data. In addition, control by funding agencies over procedures and quality of data collection, storing, and distribution to a large number of institutions, requires extensive organizational arrangements and additional personnel. This is especially true when securing the collected data for distribution after the final research has been completed and the original research staff is no longer available or when no funding is available for continuing data management at each individual site.

Datasets for use in transport and fate modeling studies require a high level of detail concerning soil and aquifer properties, density of data points, contaminant behavior, and qualitative data descriptors. Specific data requirements for subsurface models include the need to define precisely the units of measure of each input value; for example, point versus averaged values.

Data quality is often critical in model validation due to the sensitivity of most models to changes in certain parameters. Although a given field investigation may result in a large amount of data, the usefulness of the study site for model validation is determined to a large extent by the quality of the data, as reported in the data documentation. However, often the data documentation is lacking in detail, especially with respect to data quality.

Secondary Use of Research Data

A recent EPA groundwater protection data-requirements study stressed the importance of improved access to existing soil water and groundwater data and of lowering the transaction costs associated with obtaining and using such data. The report indicates that knowledge about and access to the large volume of groundwater data being generated from federal programs and state initiatives is limited, because the data are managed by many organizations and are stored in many different locations, files and formats. In addition, relatively few of these soil water and groundwater datasets are computerized, and a central cataloging facility is lacking. Although the study's conclusions are concerned with all groundwater data useful in the protection of groundwater resources, they apply equally well to research data.

Sharing Research Data

Availability and accessibility of environmental research data are discussed in a wide variety of environmental literature. Reviews of data availability indicate that many researchers give little thought to the use of their data other than for immediate research purposes. The appraisal by researchers of the importance of data accessibility is reflected in their approach to data management. Many consider it an administrative chore to be handled separately from the research, usually at the end of the study. Other investigators show a keen awareness of the importance of data management both for their own use and the use of others.

Sharing data from detailed groundwater monitoring studies and laboratory bench studies is a subject of concern both economically and with respect to the advancement of scientific research. Due to the ever-increasing cost of field studies and the extensive sampling periods required for transport and fate studies, it has become essential to share groundwater data so that unnecessary duplication can be avoided. Sharing data not only produces cost benefits; it "reinforces open, scientific inquiry; permits verification, refutation, or refinement of original research results; stimulates improvements in measurement and data collection methods; allows more efficient use of resources spent on data collection, encourages interdisciplinary use of data; and strongly discourages the uncommon, but nevertheless serious, problem of fraudulent research."

A comprehensive referral center as represented by the IGWMC Groundwater Research Data Center, focusing on selected datasets for groundwater model validation and testing, will help to avoid situations where datasets of value to many potential users go unrecognized and therefore unused.

Paul K. M van der Heijde, Wilbert I. M. Elderhorst, Rachel A. Miller, and Manjit F. Trehan are with Butler University, Indianapolis, IN 46208.
***Joe R. Williams** is the EPA Project Officer (see below).*
The complete report, entitled "The Establishment of a Groundwater Research Data Center for Validation of Subsurface Flow and Transport Models," (Order No. PB 89-224 455/AS; Cost: $28.95, subject to change) will be available only from:
National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
Telephone: 703-487-4650
The EPA Project Officer can be contacted at:
Robert S. Kerr Environmental Research Laboratory
U.S. Environmental Protection Agency
Ada, OK 74820

United States
Environmental Protection
Agency

Center for Environmental Research
Information
Cincinnati OH 45268

Official Business
Penalty for Private Use $300

EPA/600/S2-89/040