Techniques of Water-Resources Investigations
of the United States Geological Survey

A MODULAR FINITE-ELEMENT MODEL (MODFE) FOR AREAL AND AXISYMMETRIC GROUND-WATER FLOW PROBLEMS, PART 2: DERIVATION OF FINITE-ELEMENT EQUATIONS AND COMPARISONS WITH ANALYTICAL SOLUTIONS

BOOK 6
CHAPTER A4

Techniques of Water-Resources Investigations
of the United States Geological Survey

Chapter A4

A MODULAR FINITE-ELEMENT MODEL (MODFE) FOR AREAL AND AXISYMMETRIC GROUND-WATER FLOW PROBLEMS, PART 2: DERIVATION OF FINITE-ELEMENT EQUATIONS AND COMPARISONS WITH ANALYTICAL SOLUTIONS

By Richard L. Cooley

Book 6
Chapter A4

U.S. DEPARTMENT OF THE INTERIOR
MANUEL LUJAN, Jr., *Secretary*

U.S. GEOLOGICAL SURVEY
Dallas L. Peck, *Director*

PREFACE

The series of manuals on techniques describes procedures for planning and executing specialized work in water-resources investigations. The material is grouped under major subject headings called "Books" and further subdivided into sections and chapters. Section A of Book 6 is on ground-water modeling.

The unit of publication, the chapter, is limited to a narrow field of subject matters. This format allows flexibility in revision and publication as the need arises. Chapters 6A3, 6A4, and 6A5 are on the use of a particular transient finite-element numerical method for two-dimensional ground-water flow problems. These Chapters (6A3, 6A4, and 6A5) correspond to reports prepared on the finite-element model given the acronym MODFE and designated as parts 1, 2, and 3, respectively. Part 1 is on "model description and user's manual," part 2 is on "derivation of finite-element equations and comparisons with analytical solutions," and part 3 is on "design philosophy and programming details." Parts 1 and 3 have been released as Open-File Reports (see References, Torak (1992 a, b)) pending publication as Chapters 6A3 and 6A5 respectively.

Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

TECHNIQUES OF WATER-RESOURCES INVESTIGATIONS OF THE U.S. GEOLOGICAL SURVEY

The U.S. Geological Survey publishes a series of manuals describing procedures for planning and conducting specialized work in water-resources investigations. The manuals published to date are listed below and may be ordered by mail from the **U.S. Geological Survey, Books and Open-File Reports Section, Federal Center, Box 25425, Denver, Colorado 80225** (an authorized agent of the Superintendent of Documents, Government Printing Office).

Prepayment is required. Remittance should be sent by check or money order payable to U.S. Geological Survey. Prices are not included in the listing below as they are subject to change. **Current prices can be obtained** by writing to the USGS address shown above. Prices include cost of domestic surface transportation. For transmittal outside the U.S.A. (except to Canada and Mexico) a surcharge of 25 percent of the net bill should be included to cover surface transportation. When ordering any of these publications, please give the title, book number, chapter number, and "U.S. Geological Survey Techniques of Water-Resources Investigations."

---

[1]This manual is a revision of "Measurement of Time of Travel and Dispersion in Streams by Dye Tracing," by E.F. Hubbard, F.A. Kilpatrick, L.A. Martens, and J.F. Wilson, Jr., Book 3, Chapter A9, published in 1982.

[2]Spanish translation also available.

---

[1]This manual is a revision of TWRI 5–A3, "Methods of Analysis of Organic Substances in Water," by Donald F. Goerlitz and Eugene Brown, published in 1972.

[2]This manual supersedes TWRI 5–A4, "Methods for collection and analysis of aquatic biological and microbiological samples," edited by P.E. Greeson and others, published in 1977.

CONTENTS

FIGURES

A MODULAR FINITE-ELEMENT MODEL (MODFE) FOR AREAL AND
AXISYMMETRIC GROUND-WATER FLOW PROBLEMS,
PART 2: DERIVATION OF FINITE-ELEMENT EQUATIONS AND
COMPARISONS WITH ANALYTICAL SOLUTIONS

By Richard L. Cooley

ABSTRACT

MODFE, a modular finite-element model for simulating steady- or
unsteady-state, areal or axisymmetric flow of ground water in a hetero-
geneous anisotropic aquifer is documented in a three-part series of reports.
In this report, part 2, the finite-element equations are derived by minimiz-
ing a functional of the difference between the true and approximate hydrau-
lic head, which produces equations that are equivalent to those obtained by
either classical variational or Galerkin techniques. Spatial finite
elements are triangular with linear basis functions, and temporal finite
elements are one dimensional with linear basis functions. Physical
processes that can be represented by the model include (1) confined flow,
unconfined flow (using the Dupuit approximation), or a combination of both;
(2) leakage through either rigid or elastic confining units; (3) specified
recharge or discharge at points, along lines, or areally; (4) flow across
specified-flow, specified-head, or head-dependent boundaries; (5) decrease
of aquifer thickness to zero under extreme water-table decline and increase
of aquifer thickness from zero as the water table rises; and (6) head-
dependent fluxes from springs, drainage wells, leakage across riverbeds or
confining units combined with aquifer dewatering, and evapotranspiration.

The matrix equations produced by the finite-element method are solved
by the direct symmetric-Doolittle method or the iterative modified
incomplete-Cholesky conjugate-gradient method. The direct method can be
efficient for small- to medium-sized problems (less than about 500 nodes),
and the iterative method is generally more efficient for larger-sized
problems. Comparison of finite-element solutions with analytical solutions
for five example problems demonstrates that the finite-element model can
yield accurate solutions to ground-water flow problems.

INTRODUCTION

This report is the second part of a three-part series of reports (parts
1 and 3 are by Torak, 1992a and 1992b) that document the computer program
MODFE (modular finite-element model), which simulates steady- or unsteady-
state, areal or axisymmetric flow of ground water in a heterogeneous,
anisotropic aquifer. The model incorporates a variety of physical processes
necessary to simulate ground-water flow in the complicated settings that
often characterize actual field problems. Flow may be confined, unconfined
(using the Dupuit assumption), or a combination of both; known recharge and
discharge may be distributed areally, along lines such as specified-flow
boundaries, or at point sources and sinks such as pumping wells; and head-
dependent leakage may be distributed areally, such as through confining
units or wide riverbeds, or along lines such as narrow riverbeds. Confining
units may be rigid or may have elastic storage capacity. Special nonlinear,
head-dependent source and sink functions allow simulation of springs, drain-
age wells, rivers or confining units combined with aquifer dewatering, and
evapotranspiration.

1

The material in the three reports has evolved over the past 10 years from material presented by the authors in the courses entitled "Finite-Element Modeling of Ground-Water Flow" held at the U.S. Geological Survey National Training Center in Denver, Colorado. These reports formalize the course material and incorporate valuable suggestions and comments from attendees of the courses.

Features that appear to be new, at least to published finite-element programs for ground-water flow, include (1) the method of deriving the finite-element equations from a functional of the difference between the true and approximate solutions, (2) the method of approximating the variability of transmissivity over an element so that the coefficient matrix does not have to be reassembled element by element each time the saturated thickness changes, (3) the method of treating decreases of aquifer thickness to zero under conditions of extreme water-table decline and increases of aquifer thickness from zero as the water-table rises, (4) the finite-element in time method for deriving (a) the finite-element equations for unconfined flow and (b) the functions for nonlinear, head-dependent sources and sinks, and (5) the method for incorporating transient leakage from confining units.

PURPOSE AND SCOPE

The purpose of this second part of the three-part series of reports is to derive the finite-element equations for the physical processes contained in the finite-element model. A knowledge of the physics of ground-water flow, as explained by Bear (1979), for example, is assumed. The differential equations that describe the physics of the flow processes are stated and the situations under which they apply are briefly explained, but the equations are not derived here. Basic differential and integral calculus and the symbolic representation of systems of equations using matrix algebra are used extensively.

This report is organized as follows. First, the basic differential equation and boundary conditions for unsteady-state flow in a confined aquifer are stated and the finite-element equations for this system are derived in Cartesian coordinates. Next, the finite-element equations are extended to include unconfined or combined confined and unconfined flow; decreases of aquifer thickness to zero and increases from zero; the non-linear, head-dependent source and sink functions; and transient leakage from confining units. Following this, finite-element equations are derived in axisymmetric cylindrical coordinates and in steady-state form for either areal or axisymmetric problems. Finally, two matrix solution procedures are presented: a direct factorization method and an iterative, generalized conjugate-gradient procedure combined with approximate factorization.

Symbols used are defined where they first appear and in a special notation section at the end of the report. This should minimize confusion over use of similar symbols in different contexts.

ACKNOWLEDGMENT

FINITE-ELEMENT FORMULATION IN CARTESIAN COORDINATES

GOVERNING FLOW EQUATION AND BOUNDARY CONDITIONS

Ground-water flow in an aquifer where there are no discontinuities in transmissivity is assumed to be governed by the two-dimensional, unsteady-state flow equation (Bear, 1979, p. 103-116)

$$\frac{\partial}{\partial x}\left[T_{xx}\frac{\partial h}{\partial x} + T_{xy}\frac{\partial h}{\partial y}\right] + \frac{\partial}{\partial y}\left[T_{yx}\frac{\partial h}{\partial x} + T_{yy}\frac{\partial h}{\partial y}\right]$$

$$+ R(H-h) + W + P = S\frac{\partial h}{\partial t}, \tag{1}$$

where

(x,y) = Cartesian coordinate directions [length],

t = time [time],

$h(x,y,t)$ = hydraulic head in the aquifer [length],

$H(x,y,t)$ = hydraulic head at the distal side of a confining unit [length],

$\begin{bmatrix} T_{xx}(x,y,t) & T_{xy}(x,y,t) \\ T_{yx}(x,y,t) & T_{yy}(x,y,t) \end{bmatrix}$ = symmetric transmissivity tensor written in matrix form $\left[\text{length}^2/\text{time}\right]$,

$R(x,y,t)$ = hydraulic conductance (vertical hydraulic conductivity divided by thickness) of a confining unit [time^{-1}],

$S(x,y,t)$ = storage coefficient [0],

$W(x,y,t)$ = unit areal recharge or discharge rate [length/time] (positive for recharge), and

$P(x,y,t) = \sum_{j=1}^{p}\delta\left(x-a_j'\right)\delta\left(y-b_j'\right)Q_j(t)$ = designation using Dirac delta functions for p point sources or sinks, each of strength Q_j [length^3/time] (positive for injection) and located at $x = a_j'$ and $y = b_j'$.

Equation (1) is subject to the following boundary and initial conditions:

1. At a discontinuity in transmissivity within the aquifer, hydraulic head and the component of flow normal to the discontinuity are unchanged as the discontinuity is crossed (Bear, 1979, p. 100-102). Thus, at a discontinuity in transmissivity between transmissivity zones a and b (figure 1),

$$h\big|_a = h\big|_b \tag{2}$$

and

$$q_n\big|_a = q_n\big|_b, \tag{3}$$

where $\cdot\big|_a$ and $\cdot\big|_b$ indicate evaluation just within the a and b sides of the discontinuity, respectively, and $q_n(x,y,t)$ is the normal component of flow (specific discharge times aquifer thickness).
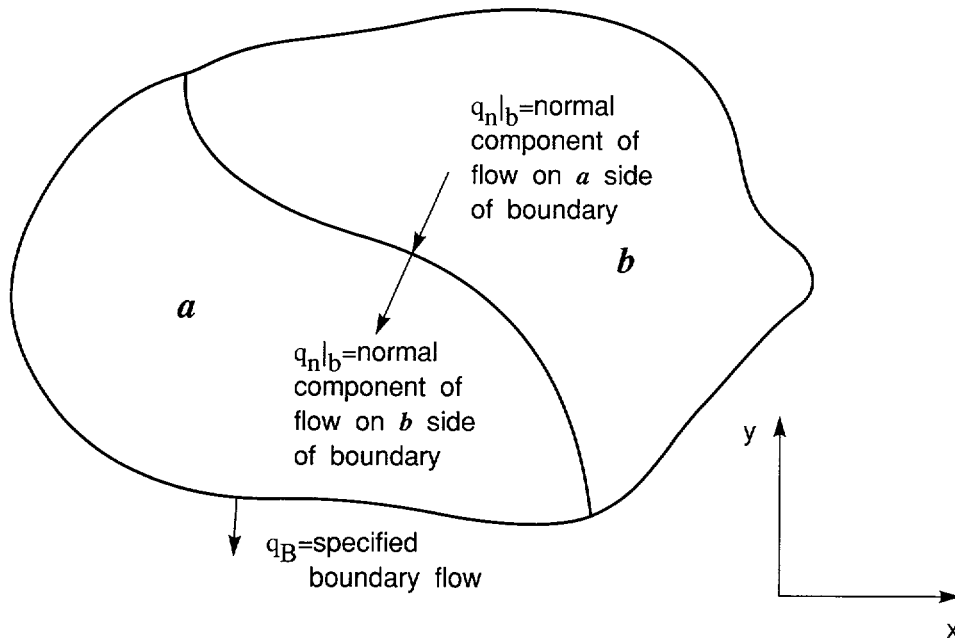
Figure 1. A hypothetical aquifer that has a discontinuity in transmissivity
between zones a and b.

2. The normal component of flow across a boundary of the aquifer is given
by the sum of specified and head-dependent flow components (Bear, 1979,
p. 117-120). Thus, on this type of boundary

$$q_n = q_B + \alpha\left(H_B - h\right),\qquad\qquad(4)$$

where

$q_B(x,y,t)$ = specified flow (specific discharge times aquifer thickness)
normal to the boundary [length^2/time] (positive for
inflow),

$\alpha(x,y,t)$ = a parameter that approaches infinity for a specified-head
(Dirichlet) condition, is zero for a specified flow
(Neumann) condition, and is finite and positive for a
general or mixed (Cauchy) condition [length/time], and

$H_B(x,y,t)$ = specified head at the boundary [length].

Note that although equation (4) is usually used to specify external boundary
conditions (see Bear, 1979, p. 116-123, for examples), it may also be used
to specify internal sources and sinks such as rivers (which are idealized as
lines) or springs (which are idealized as points).

3. The hydraulic head is known everywhere at the initial instant of time, or

$$h = H_o,\qquad\qquad(5)$$

where

$H_o(x,y)$ = the initial head [length].

For convenience in subsequent discussions, specified flow ($\alpha = 0$ in
equation (4)) and Cauchy ($0 < \alpha < \infty$ in equation 4)) boundary conditions are
referred to as Cauchy-type boundary conditions, because the former is simply
a special case of the latter. Specified-head boundary conditions are
treated separately from Cauchy-type boundary conditions.

FINITE-ELEMENT DISCRETIZATION

The finite-element method is used to solve equations (1) through (5). The basic concept underlying the finite-element method is that a complex flow region or domain may be subdivided into a network of subregions or elements, each having a simple shape (figure 2a). Each of these elements is then assumed to be small enough that at any instant of time the true solution, h, of equations (1) through (5) may be approximated within the

element by a simple function, $\hat{h}$. These local functions are continuous across element boundaries to ensure that the approximate solution is spatially continuous. Presumably, as each element is reduced in size and the number of elements is increased, the approximate solution approaches the true solution.

A)

B)
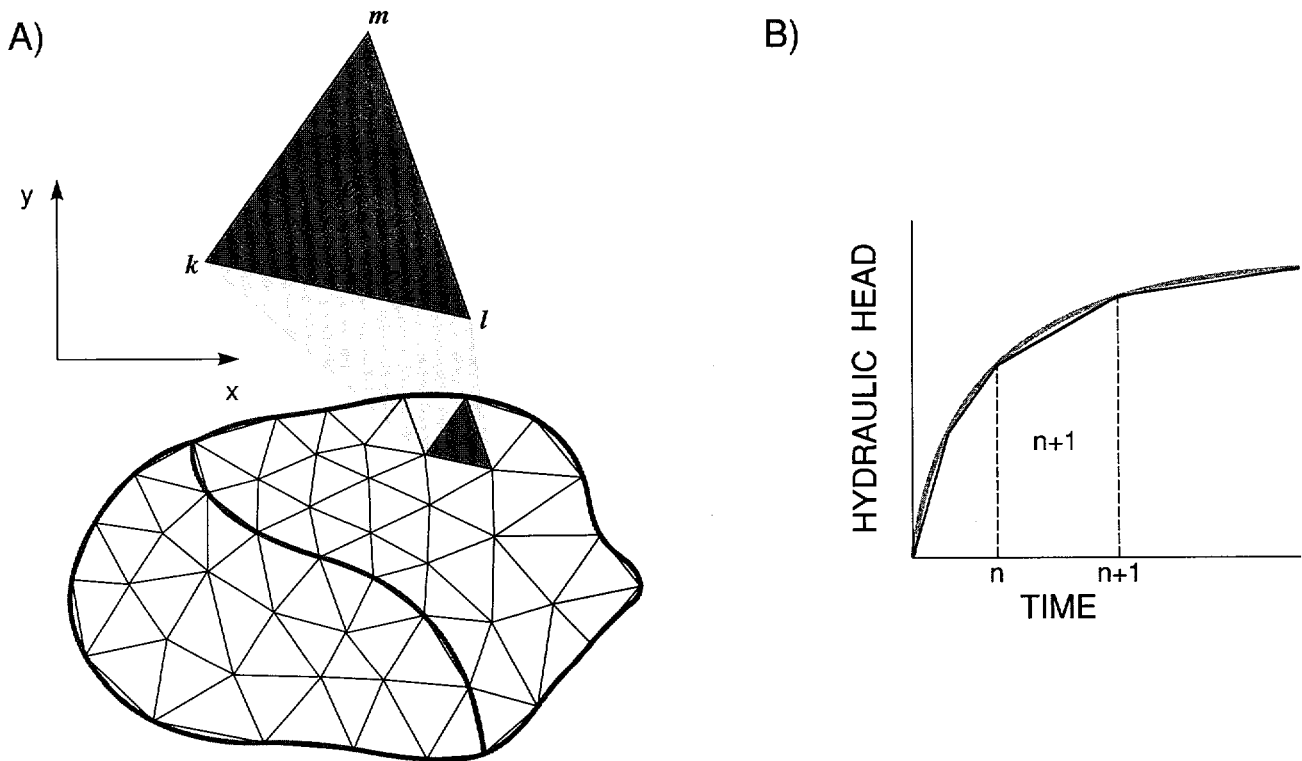
Figure 2. (a) Hypothetical aquifer of figure 1 subdivided into spatial finite elements, and (b) variation of hydraulic head with time subdivided into time elements.

The time domain of the true solution is similarly subdivided into elements (figure 2b), each bounded by two points in time at which local approximate functions are linked to form a piecewise continuous function of time. First the spatial functions are developed, then the time functions are superimposed.

In the present report, spatial element shapes are assumed to be triangles (figure 2a) and head, $\hat{h}$, is assumed to vary linearly within each element. Element corners are called nodes. Because three points define a plane, the three nodes of each triangular element are used to define the linear function.

5

At any point within typical element e (figure 2a) having nodes k, l, and m, the approximate solution may be written as

$$\hat{h} = A^e + B^e x + C^e y, \tag{6}$$

where constants A^e, B^e, and C^e can be found from the simultaneous equations that must be satisfied at the nodes:

$$\hat{h}_k = A^e + B^e x_k + C^e y_k,$$
$$\hat{h}_l = A^e + B^e x_l + C^e y_l, \tag{7}$$
$$\hat{h}_m = A^e + B^e x_m + C^e y_m,$$

Solution of equations (7) for A^e, B^e, and C^e, substitution of the results into equation (6), and rearrangement yields the final equation (Segerlind, 1976, p. 28-30)

$$\hat{h} = \hat{h}_k N_k^e + \hat{h}_l N_l^e + \hat{h}_m N_m^e, \tag{8}$$

where

$$\hat{h}_i = \hat{h}\left(x_i, y_i, t\right), \quad i = k, l, m,$$

$$\tag{9}$$

$$N_i^e = \left[a_i^e + b_i^e x + c_i^e y\right]/2\Delta^e, \quad i = k, l, m,$$

and the N_i^e are called basis (or coordinate) functions. In equations (9),

$$a_k^e = x_l y_m - x_m y_l,$$
$$b_k^e = y_l - y_m,$$
$$c_k^e = x_m - x_l,$$
$$a_l^e = x_m y_k - x_k y_m,$$
$$b_l^e = y_m - y_k, \tag{10}$$
$$c_l^e = x_k - x_m,$$
$$a_m^e = x_k y_l - x_l y_k,$$
$$b_m^e = y_k - y_l,$$
$$c_m^e = x_l - x_k,$$

and

$$2\Delta^e = \left(x_k - x_m\right)\left(y_l - y_m\right) - \left(x_m - x_l\right)\left(y_m - y_k\right). \tag{11}$$

If nodes k, l, and m are numbered counter-clockwise around element e,

then Δ^e is the area of element e. Otherwise, Δ^e is the negative of the area. Following the counter-clockwise numbering convention is critical to maintain the proper signs of quantities in the finite-element equations to be developed.
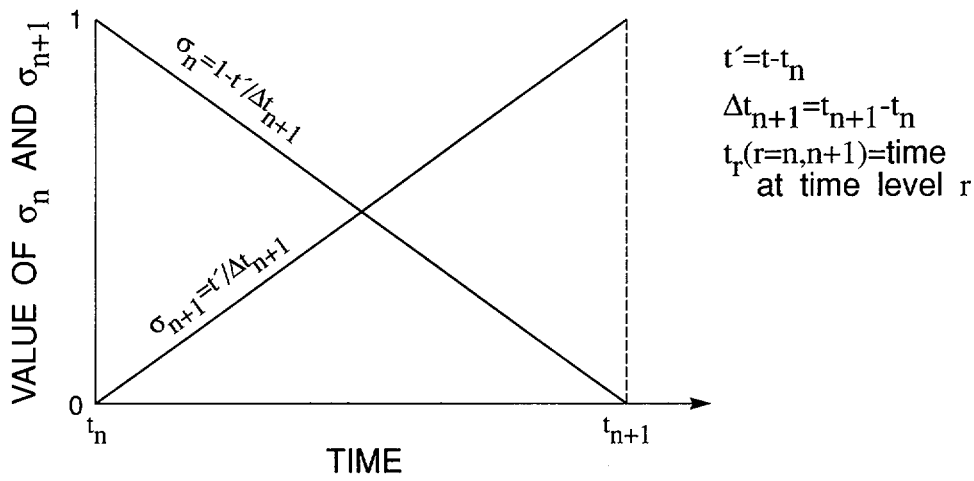
Figure 3. Finite-element discretization of time using basis functions σ_n and σ_{n+1} (after Zienkiewicz, 1971, p. 337).

Useful properties of the N_i^e are given by Wang and Anderson (1982, p. 120) as:

1. N_i^e is 1 at node i and 0 at the other two nodes.

2. N_i^e varies linearly with distance along any side.

3. N_i^e is 0 along the side opposite node i.

4. N_i^e is 1/3 at the centroid of the triangular element.

Another easily verified, useful property is that $N_k^e + N_l^e + N_m^e = 1$ at any point (x,y) in element e.

An approximate solution over time is developed by using the same finite-element concepts used to derive the approximate solution in space (Zienkiewicz, 1971, p. 335-337). Finite elements in time are chosen to be one-dimensional, and basis functions σ are chosen to be linear with a time node at each end of each element (figure 2b). If times at two time nodes are designated as t_n and t_{n+1}, and the length $t_{n+1} - t_n$ of a time element is Δt_{n+1} (figure 3), then hydraulic head $\hat{h}$ can be written for each space node i within each time element as

$$\hat{h}_i = \hat{h}_{i,n}\, \sigma_n + \hat{h}_{i,n+1}\, \sigma_{n+1}, \tag{12}$$

where the basis functions are given by

$$\sigma_n = 1 - \frac{t'}{\Delta t_{n+1}}, \tag{13}$$

$$\sigma_{n+1} = \frac{t'}{\Delta t_{n+1}},$$

$t' = t - t_n$, and $\hat{h}_{i,r} = \hat{h}_i\left(t_r\right)$, $r = n, n+1$. The basis functions σ_n and σ_{n+1} satisfy the first three properties listed for N_i^e previously, modified accordingly for the one-dimensional nature of the time element.

Combination of equations (8) and (12) yields the final approximate solution

$$\hat{h} = \sum_i \left[\hat{h}_{i,n} \, \sigma_n + \hat{h}_{i,n+1} \, \sigma_{n+1} \right] N_i^e, \quad i = k,1,m. \tag{14}$$

Nodal hydraulic heads in equation (14) are calculated so that $\hat{h}$ approximates the true solution, as described in the following section.

DERIVATION OF FINITE-ELEMENT EQUATIONS

Assume that there are N nodes in the flow domain, and that we wish to solve for values of hydraulic head at all N nodes. The necessary equations are generated by the approximate solution of equations (1) through (5), which is commonly derived using either weighted residual methods (Zienkiewicz, 1971, chap. 3; Norrie and deVries, 1973, chaps. 2 and 5; Pinder and Gray, 1977, chap. 3) or classical variational methods (Zienkiewicz, 1971, chaps. 3, 15, and 16; Remson and others, 1971, chap. 7; Norrie and deVries, 1973, chaps. 3-6, 9, 10). In weighted residual methods, solution over space is generally carried out separately from solution over time. To derive the necessary equations, the approximate solution given by equation (8) is substituted into equation (1) to form a residual, which is then multiplied by each member of a set of N weighting functions and integrated over the flow domain. The resulting set of N equations is then manipulated using the boundary conditions (equations (2) and (4)) to yield a set of N ordinary differential equations in time, which are usually solved with finite-difference methods. A commonly used weighted residual method is the Galerkin method, where the weighting functions are the basis functions

N_i, each of which is the union of all elemental basis functions N_i^e. A

Galerkin in time method was given by Zienkiewicz (1971, p. 335-336) as an alternative to the finite-difference solution over time.

The classical variational method involves use of a variational principle, which is an integral that, when minimized over the flow domain, yields equations (1) and (4). Because this variational principal is equivalent to the flow problem, the approximate solution may be substituted into it, and the integral may be minimized with respect to each nodal value of hydraulic head to yield the required finite-element equations. Variational and Galerkin finite-element methods applied to equations (1) through (5) yield the same set of finite-element equations when the same approximate solution (for example, equation (8)) is used.

Error-functional justification for the finite-element equations

Another method that is closely related to the classical variational method is to fit the approximate solution to the true solution using an

integral functional[1] of the error, $\hat{e} = h - \hat{h}$. In this author's opinion, derivation of the finite-element equations with this method is easier and provides more direct insight into the nature of the solution in terms of its error than the other methods.

[1]A functional is a function of a function. The integral is a function of the error $\hat{e} = h - \hat{h}$, and $\hat{e}$ is regarded as a function of the values of $\hat{h}_{i,n+1}$; hence, the integral is a functional.

To be useful, the functional, termed $\hat{I}(e)$, must be defined such that (1) $\hat{I}(e) \geq 0$, with equality occurring only if $\hat{e} = 0$, (2) the true solution, h, can be eliminated from the final finite-element equations, and (3) $\hat{I}(e)$ measures total (or integrated) error over the entire flow domain. The only error functional that satisfies these requirements and produces the same equations as produced by the Galerkin and classical variational methods is

$$\hat{I}(e) = \sum_e \int_0^{\Delta t_{n+1}} \left\{ \iint_{\Delta^e} \left[\frac{\partial \hat{e}}{\partial x}\left(T_{xx}\frac{\partial \hat{e}}{\partial x} + T_{xy}\frac{\partial \hat{e}}{\partial y}\right) + \frac{\partial \hat{e}}{\partial y}\left(T_{yx}\frac{\partial \hat{e}}{\partial x} + T_{yy}\frac{\partial \hat{e}}{\partial y}\right) \right. \right.$$

$$\left. \left. + R\hat{e}^2 + S\left(\frac{\partial \hat{e}}{\partial t}\right)^2 t' \right] dxdy + \int_{C_2^e} \alpha\hat{e}^2 \, dC \right\}dt', \tag{15}$$

where the sum over e indicates the sum over all elements, the double integral over Δ^e indicates integration over spatial element e, and the contour integral over C_2^e indicates integration over the side (if any) of element e that is part of a boundary where a Cauchy-type boundary condition applies. For equation (15) to be valid, the matrix of transmissivities must be symmetric and positive definite, and R, S, and α must be greater than or equal to zero. The requirement for the transmissivities guarantees that the sum of terms involving transmissivities is positive (or zero if $\hat{e} = 0$) because this sum is a positive-definite quadratic form (see Hohn, 1964, p. 336, 338). Note that for ground-water flow problems, all of these requirements are satisfied.

The approximate solution is fitted to the true solution by minimizing $\hat{I}(e)$ with respect to the approximate solution, which leads to an error distribution in which the error at any point (x,y,t) is as small as possible as measured by $\hat{I}(e)$. Because functional $\hat{I}(e)$ includes terms involving the error and its spatial and temporal derivatives, the minimization process minimizes the combination of the error and its derivatives. Magnitudes of T_{xx} (etc.), R, S, and α indicate which types of terms are more heavily weighted, and thus have more influence on the solution, for any given problem. For example, if terms involving the error directly were heavily weighted (that is, R and (or) α were large) compared to the other terms, then the average (integrated) error should be small, but if terms involving derivatives were heavily weighted, then the average error might be large if large errors were required to make the average derivatives of the error small. This latter situation could arise if space or time elements were too large or were poorly configured.

Minimization of equation (15) is accomplished by taking its derivative with respect to each value of $\hat{h}_{i,n+1}$, $i = 1,2,\cdots,N$, and setting each result to zero. Equation (15) does not also have to be minimized with respect to

9

$\hat{h}_{i,n}$ because an equation for time level n was created by minimizing equation (15) with respect to $\hat{h}_{i,n+1}$ for the previous time element. For the initial time element, $\hat{h}_{i,0}$ is the known initial condition so that equation (15) is not minimized with respect to it. It can be readily verified that the result of minimization is

$$\frac{\partial I}{\partial \hat{h}_{i,n+1}} = -2\sum_{e_i} \int_0^{\Delta t_{n+1}} \sigma_{n+1} \left\{ \int\int_{\Delta^e} \left[\frac{\partial N_i^e}{\partial x}\left(T_{xx}\frac{\partial \hat{e}}{\partial x} + T_{xy}\frac{\partial \hat{e}}{\partial y}\right) + \frac{\partial N_i^e}{\partial y}\left(T_{yx}\frac{\partial \hat{e}}{\partial x}\right.\right.\right.$$

$$\left.\left. + T_{yy}\frac{\partial \hat{e}}{\partial y}\right] + N_i^e R\hat{e} + N_i^e S\frac{\partial \hat{e}}{\partial t}\right] dxdy$$

$$\left. + \int_{C_2^e} N_i^e \alpha \hat{e} dC \right\} dt' = 0, \quad i = 1,2,\cdots,N, \tag{16}$$

where summation over e_i indicates summation over all elements sharing node i, termed a patch of elements by Wang and Anderson (1982, p. 12) (figure 4). Terms for all other elements over the flow domain drop out because $\hat{h}_{i,n+1}$ does not appear in the approximate solutions in these elements.

Element in patch sharing node i
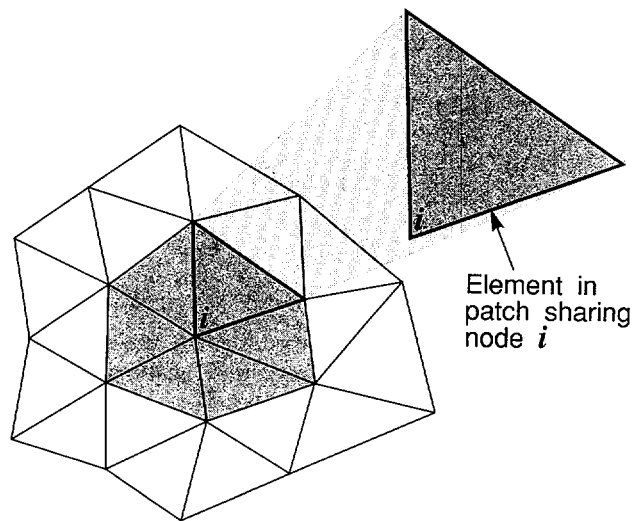
Figure 4. A typical patch of elements sharing node i.

Equation (16) can be separated into two parts, one written in terms of approximate solution $\hat{h}$ and the other written in terms of the true solution h. Thus,

$$\sum_e \int_0^{\Delta t_{n+1}} \sigma_{n+1} \left\{ \iint_{\Delta^e} \left[N_i^e \left(S \frac{\partial \hat{h}}{\partial t} - R\left(H-\hat{h}\right) - W - P \right) + \frac{\partial N_i^e}{\partial x} \left(T_{xx} \frac{\partial \hat{h}}{\partial x} + T_{xy} \frac{\partial \hat{h}}{\partial y} \right) \right. \right.$$

$$\left. \left. + \frac{\partial N_i^e}{\partial y} \left(T_{yx} \frac{\partial \hat{h}}{\partial x} + T_{yy} \frac{\partial \hat{h}}{\partial y} \right) \right] dxdy - \int_{C_2^e} N_i^e \left[q_B + \alpha\left(H_B - \hat{h}\right) \right] dC \right\} dt'$$

$$- \sum_e \int_0^{\Delta t_{n+1}} \sigma_{n+1} \left\{ \iint_{\Delta^e} \left[N_i^e \left(S \frac{\partial h}{\partial t} - R(H-h) - W - P \right) + \frac{\partial N_i^e}{\partial x} \left(T_{xx} \frac{\partial h}{\partial x} + T_{xy} \frac{\partial h}{\partial y} \right) \right. \right.$$

$$\left. \left. + \frac{\partial N_i^e}{\partial y} \left(T_{yx} \frac{\partial h}{\partial x} + T_{yy} \frac{\partial h}{\partial y} \right) \right] dxdy - \int_{C_2^e} N_i^e \left[q_B + \alpha\left(H_B - h\right) \right] dC \right\} dt' = 0. \quad (17)$$

Note that, to make each part of equation (17) complete, several terms were added to one part of the equation and subtracted from the other part. In appendix A the sum of the terms involving the true solution is shown to equal zero, so that equation (17) becomes

$$\sum_e \int_0^{\Delta t_{n+1}} \sigma_{n+1} \left\{ \iint_{\Delta^e} \left[N_i^e \left(S \frac{\partial \hat{h}}{\partial t} - R\left(H-\hat{h}\right) - W - P \right) + \frac{\partial N_i^e}{\partial x} \left(T_{xx} \frac{\partial \hat{h}}{\partial x} + T_{xy} \frac{\partial \hat{h}}{\partial y} \right) \right. \right.$$

$$\left. \left. + \frac{\partial N_i^e}{\partial y} \left(T_{yx} \frac{\partial \hat{h}}{\partial x} + T_{yy} \frac{\partial \hat{h}}{\partial y} \right) \right] dxdy - \int_{C_2^e} N_i^e \left[q_B + \alpha\left(H_B - \hat{h}\right) \right] dC \right\} dt' = 0, \quad i = 1,2,\cdots,N. \quad (18)$$

Equation (18) represents the required set of finite-element equations. Performing the indicated integrations yields the final set of operational equations. However, before the integrations can be accomplished, the specific space and time dependencies of the various terms in the integrals must be specified, and two desirable simplifications are made.

Integral approximations

The first simplification involves the integrals of $S\partial\hat{h}/\partial t$, $R(H-\hat{h})$, and $\alpha(H_B-\hat{h})$. These integrals do not involve spatial derivatives of $\hat{h}$ and can be shown to contribute positive terms to the diagonal and off-diagonal elements of the final coefficient matrix for the approximate solution (Segerlind, 1976, p. 216). In contrast, the integrals involving spatial derivatives contribute nonpositive off-diagonal terms and positive diagonal terms such that the sum of absolute values of the off-diagonal terms equals the diagonal term if all internal angles of the triangular elements are less than or equal to 90° (Narasimhan and others, 1978, p. 866). When specified-head boundary conditions are introduced, the coefficient matrix resulting

11

from the spatial derivative terms is a type of M-matrix known as a Stieltjes matrix (Varga, 1962, p. 85), which is ideal for the iterative matrix solution technique introduced further on. In addition, a Stieltjes final coefficient matrix can be shown to guarantee a nonoscillatory solution to equation (18) when combined with proper restrictions in time-element size (Briggs and Dixon, 1968). Addition of positive off-diagonal terms to the matrix can destroy the Stieltjes matrix property, so that it is desirable to replace the integrals of $S\partial\hat{h}/\partial t$, $R(\hat{H-h})$, and $\alpha(H_B-\hat{h})$ with integrals that contribute only positive diagonal terms. This replacement also simplifies the resulting finite-element equations so that their solution requires less computer time and storage than if the matrices resulting from the original integrals were used.

In structural dynamics problems, replacement of the so-called consistent mass matrix (the matrix resulting from an integral involving second derivatives of time that is analogous to the integral of $S\partial\hat{h}/\partial t$) with a diagonal approximation of the mass matrix has been reported to yield degraded results (Zienkiewicz, 1971, p. 326). Similar degraded solutions were reported when a diagonal approximation was used for advection-dominated advection-diffusion problems (Gresho and others, 1976). However, Narasimhan and others (1978, p. 863-864) argue that a diagonal approximation enhances the numerical performance when applied to the integral of $S\partial\hat{h}/\partial t$, and that retaining the nondiagonal form can lead to numerical difficulties. In addition, Wilson and others (1979) obtained good correspondence between analytical and finite-element solutions of equation (1) for several different test problems by using the same diagonal approximation, linear basis functions, and triangular spatial elements as used here. The author is aware of no study indicating degraded solutions when the diagonal approximation is applied to equations (1) through (5) using triangular spatial elements and linear basis functions, and the author's own numerical experiments have not revealed any significant degradation either. Finally, the author's analysis indicates that the method used here yields consistent mass balance over each patch of elements.

The method can be demonstrated for one integral, and results for the other two are similar. The diagonal approximation is

$$\iint_{\Delta^e} SN_i^e \frac{\partial \hat{h}}{\partial t} \, dxdy \simeq \iint_{\Delta^e} SN_i^e \frac{d\hat{h}_i}{dt} \, dxdy. \qquad (19)$$

The quadratic function $N_i^e \partial\hat{h}/\partial t$ is replaced by the linear function $N_i^e d\hat{h}_i/dt$, which, for constant S over the element (which is adopted for the present report), makes the approximation equivalent to the second-order correct trapezoidal rule (McCracken and Dorn, 1964, p. 161-166).

Rotation of coordinate axes

The second simplification, which is not an approximation, involves rotating the x and y coordinate axes locally, within each element, to axes $\bar{x}$ and $\bar{y}$ that coincide with the principal directions of the transmissivity tensor (figure 5) (Zienkiewicz and others, 1966). In the rotated coordinate

12
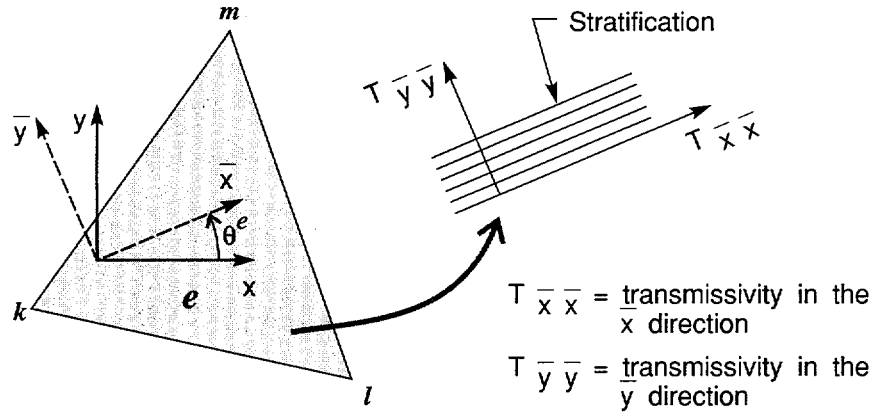
Figure 5. Rotation from global (x,y) to local $(\bar{x},\bar{y})$ coordinates in element e having node numbers k, l, and m.

system, the only nonzero components of the local transmissivity tensor are the diagonal (principal) components, $T_{\bar{x}\bar{x}}$ and $T_{\bar{y}\bar{y}}$. Coordinates $\bar{x}$ and $\bar{y}$ are obtained by using the rotation equations

$$\bar{x} = x \cos \theta^e + y \sin \theta^e,$$

$$\bar{y} = -x \sin \theta^e + y \cos \theta^e,$$

(20)

where θ^e is the angle of rotation of the axes, measured counter-clockwise, in element e (see figure 5). By replacing coordinates x and y and the original transmissivity tensor with rotated coordinates $\bar{x}$ and $\bar{y}$ and the diagonal transmissivity tensor, equation (18) can be transformed to become

$$\sum_{e_i} \int_0^{\Delta t_{n+1}} \sigma_{n+1} \left\{ \int\int_{\Delta^e} \left[\bar{N}_i^e \left(S \frac{d\hat{h}_i}{dt} - R\left(H_i - \hat{h}_i \right) - W - P \right) + \frac{\partial \bar{N}_i^e}{\partial \bar{x}} T_{\bar{x}\bar{x}} \frac{\partial \hat{h}}{\partial \bar{x}} \right. \right.$$

(21)

$$\left. \left. + \frac{\partial \bar{N}_i^e}{\partial \bar{y}} T_{\bar{y}\bar{y}} \frac{\partial \hat{h}}{\partial \bar{y}} \right] d\bar{x}d\bar{y} - \int_{C_2^e} \bar{N}_i^e \left[q_B + \alpha\left(H_{Bi} - \hat{h}_i \right) \right] d\bar{C} \right\} dt' = 0, \quad i = 1,2,\cdots,N,$$

where the bars over the variables indicate evaluation using $\bar{x}$ and $\bar{y}$ and equations like equation (19) were used to modify the appropriate integrals.

Evaluation of spatial integrals

To reduce notational complexity, the space and time integrations in equation (21) are performed in two separate steps. To perform the space integrations, it is assumed that S, R, and W are constant in each spatial element, and that $T_{\bar{x}\bar{x}}$ and $T_{\bar{y}\bar{y}}$ are linearly variable in each element as given

by relationships analogous to equation (8). That is,

$$T_{\bar{x}\bar{x}} = T_{\bar{x}\bar{x}k}^e \bar{N}_k^e + T_{\bar{x}\bar{x}l}^e \bar{N}_l^e + T_{\bar{x}\bar{x}m}^e \bar{N}_m^e,$$

(22)

and

$$T_{\bar{y}\bar{y}} = T_{\bar{y}\bar{y}k}^e \bar{N}_k^e + T_{\bar{y}\bar{y}l}^e \bar{N}_l^e + T_{\bar{y}\bar{y}m}^e \bar{N}_m^e,$$

(23)

13

where $T^e_{\bar{x}\bar{x}k}$, etc., are values of transmissivity at nodes k, etc., in element e. It is further assumed that q_B and α are constant along any Cauchy-type boundary side of each element. The integration is performed for typical element e bounded by nodes k, 1, and m using the general formulas (Segerlind, 1976, p. 45)

$$\iint_{\Delta^e} \left(\bar{N}^e_k\right)^p \left(\bar{N}^e_1\right)^q \left(\bar{N}^e_m\right)^r d\bar{x}d\bar{y} = \frac{p!q!r!}{(p+q+r+2)!} 2\Delta^e \qquad (24)$$

and

$$\int_{L_{k1}} \left(\bar{N}^e_k\right)^p \left(\bar{N}^e_1\right)^q d\bar{C} = \frac{p!q!}{(p+q+1)!} L_{k1}, \qquad (25)$$

where L_{k1} is the length of the element side between nodes k and 1. Thus, by writing $\hat{h}$ using equation (8) and substituting the appropriate expressions for $\bar{N}^e_i$, $\partial\bar{N}^e_i/\partial\bar{x}$, and $\partial\bar{N}^e_i/\partial\bar{y}$, i = k,1,m, the spatial integrals in equation (21) are evaluated for i = k (for example) as

$$\iint_{\Delta^e} \bar{N}^e_k \, S \, \frac{d\hat{h}_k}{dt} d\bar{x}d\bar{y} = \frac{1}{3}S^e\Delta^e \frac{d\hat{h}_k}{dt}, \qquad (26)$$

$$\iint_{\Delta^e} \bar{N}^e_k \, R\left(H_k-\hat{h}_k\right) d\bar{x}d\bar{y} = \frac{1}{3}R^e\Delta^e\left(H_k-\hat{h}_k\right), \qquad (27)$$

$$\iint_{\Delta^e} \bar{N}^e_k \, W d\bar{x}d\bar{y} = \frac{1}{3}W^e\Delta^e, \qquad (28)$$

$$\iint_{\Delta^e} \bar{N}^e_k \, P d\bar{x}d\bar{y} = \iint_{\Delta^e} \bar{N}^e_k \sum_{j=1}^{P_e} \delta\left(\bar{x}-\bar{a}'_j\right) \delta\left(\bar{y}-\bar{b}'_j\right)Q_j d\bar{x}d\bar{y}$$

$$= \sum_{j=1}^{P_e} \bar{N}^e_k\left(\bar{a}'_j,\bar{b}'_j\right)Q_j = P^e_k, \qquad (29)$$

$$\iint_{\Delta^e} \frac{\partial\bar{N}^e_k}{\partial\bar{x}} T_{\bar{x}\bar{x}} \frac{\partial\hat{h}}{\partial\bar{x}} d\bar{x}d\bar{y} = \iint_{\Delta^e} \frac{\partial\bar{N}^e_k}{\partial\bar{x}}\left(T^e_{\bar{x}\bar{x}k} \bar{N}^e_k + T^e_{\bar{x}\bar{x}1} \bar{N}^e_1 + T^e_{\bar{x}\bar{x}m} \bar{N}^e_m\right)\cdot$$

$$\cdot \left[\frac{\partial \bar{N}_k^e}{\partial \bar{x}} \hat{h}_k + \frac{\partial \bar{N}_1^e}{\partial \bar{x}} \hat{h}_1 + \frac{\partial \bar{N}_m^e}{\partial \bar{x}} \hat{h}_m \right] d\bar{x} d\bar{y}$$

$$= \frac{T_{\bar{x}\bar{x}}^e}{4\Delta^e} \left[\bar{b}_k^e \bar{b}_k^e \hat{h}_k + \bar{b}_k^e \bar{b}_1^e \hat{h}_1 + \bar{b}_k^e \bar{b}_m^e \hat{h}_m \right], \tag{30}$$

$$\int_{\Delta^e} \int \frac{\partial \bar{N}_k^e}{\partial \bar{y}} T_{\bar{y}\bar{y}} \frac{\partial \hat{h}}{\partial \bar{y}} d\bar{x} d\bar{y} = \int_{\Delta^e} \int \frac{\partial \bar{N}_k^e}{\partial \bar{y}} \left[T_{\bar{y}\bar{y}k}^e \bar{N}_k^e + T_{\bar{y}\bar{y}1}^e \bar{N}_1^e + T_{\bar{y}\bar{y}m}^e \bar{N}_m^e \right] \cdot$$

$$\cdot \left[\frac{\partial \bar{N}_k^e}{\partial \bar{y}} \hat{h}_k + \frac{\partial \bar{N}_1^e}{\partial \bar{y}} \hat{h}_1 + \frac{\partial \bar{N}_m^e}{\partial \bar{y}} \hat{h}_m \right] d\bar{x} d\bar{y}$$

$$= \frac{T_{\bar{y}\bar{y}}^e}{4\Delta^e} \left[\bar{c}_k^e \bar{c}_k^e \hat{h}_k + \bar{c}_k^e \bar{c}_1^e \hat{h}_1 + \bar{c}_k^e \bar{c}_m^e \hat{h}_m \right], \tag{31}$$

$$\int_{C_2^e} \bar{N}_k^e \left[q_B + \alpha \left(H_{Bk} - \hat{h}_k \right) \right] d\bar{c} = \frac{1}{2} \left[\left(q_B L \right)_{k1} + \left(\alpha L \right)_{k1} \left(H_{Bk} - \hat{h}_k \right) \right.$$

$$\left. + \left(q_B L \right)_{km} + \left(\alpha L \right)_{km} \left[H_{Bk} - \hat{h}_k \right] \right], \tag{32}$$

where S^e, R^e, and W^e are the constant values of S, R, and W in element e;

$$T_{\bar{x}\bar{x}}^e = \frac{1}{3} \left(T_{\bar{x}\bar{x}k}^e + T_{\bar{x}\bar{x}1}^e + T_{\bar{x}\bar{x}m}^e \right); \tag{33}$$

$$T_{\bar{y}\bar{y}}^e = \frac{1}{3} \left(T_{\bar{y}\bar{y}k}^e + T_{\bar{y}\bar{y}1}^e + T_{\bar{y}\bar{y}m}^e \right); \tag{34}$$

$\bar{b}_i^e$ and $\bar{c}_i^e$, $i = k, 1, m$, are defined by equations (10) and evaluated using $\bar{x}$ and $\bar{y}$; p_e is the number of point sources and sinks in element e; $\bar{N}_k^e \left(\bar{a}_j', \bar{b}_j' \right)$ is the basis function for node k evaluated at point $\left(\bar{a}_j', \bar{b}_j' \right)$; and L_{k1} and L_{km} are lengths of element sides between nodes k and 1 and between nodes k and m, respectively, on a Cauchy-type boundary. If a side is not on a Cauchy-type boundary, then L for that side is set to zero.

15

The term P_k^e represents the total amount of pumping that is allocated to node k in element e. If a well is located at node k (where $\bar{a}_j' = \bar{x}_k$ and $\bar{b}_j' = \bar{y}_k$ so that $\bar{N}_k^e\left(\bar{a}_j', \bar{b}_j'\right) = 1$), then the pumping rate Q_j can be allocated to one element so that when summed over all elements in the patch, the total rate is still Q_j. For other points in element e, the rate allocated to node k is less than the total rate Q_j because $\bar{N}_k^e\left(\bar{a}_j', \bar{b}_j'\right) < 1$ for $\bar{a}_j' \neq \bar{x}_k$ and (or) $\bar{b}_j' \neq \bar{y}_k$. However, parts of Q_j are also allocated to the other two nodes of the element so that, because $\bar{N}_k^e + \bar{N}_l^e + \bar{N}_m^e = 1$, the sum of the rates allocated to the three nodes is Q_j, as required.

By using equations (26) through (34), the spatial integrals for element e in equation (21) can be written as

$$\int\int_{\Delta^e} \left[\bar{N}_k^e\left(S\frac{d\hat{h}_k}{dt} - R\left(H_k - \hat{h}_k\right) - W - P \right) + \frac{\partial \bar{N}_k^e}{\partial\bar{x}} T_{\bar{x}\bar{x}} \frac{\partial\hat{h}}{\partial\bar{x}} + \frac{\partial \bar{N}_k^e}{\partial\bar{y}} T_{\bar{y}\bar{y}} \frac{\partial\hat{h}}{\partial\bar{y}} \right] d\bar{x}d\bar{y}$$

$$- \int_{C_2^e} \bar{N}_k^e\left[q_B + \alpha\left(H_{Bk} - \hat{h}_k\right) \right] d\bar{C}$$

$$= c_{kk}^e \frac{d\hat{h}_k}{dt} + \left(g_{kk}^e + v_{kk}^e \right)\hat{h}_k + g_{kl}^e \hat{h}_l + g_{km}^e \hat{h}_m - \frac{1}{3}R^e\Delta^e H_k - \frac{1}{3}W^e\Delta^e - P_k^e$$

$$- \frac{1}{2}\left[\left(q_B L\right)_{kl} + \left(q_B L\right)_{km} \right] - \frac{1}{2}\left[\left(\alpha L\right)_{kl} + \left(\alpha L\right)_{km} \right]H_{Bk}, \tag{35}$$

where

$$c_{kk}^e = \frac{1}{3}S^e\Delta^e, \tag{36}$$

$$v_{kk}^e = \frac{1}{3}R^e\Delta^e + \frac{1}{2}\left[\left(\alpha L\right)_{kl} + \left(\alpha L\right)_{km} \right], \tag{37}$$

$$g_{kk}^e = \frac{T_{\bar{x}\bar{x}}^e}{4\Delta^e} \bar{b}_k^e\bar{b}_k^e + \frac{T_{\bar{y}\bar{y}}^e}{4\Delta^e} \bar{c}_k^e\bar{c}_k^e, \tag{38}$$

16

$$g_{kl}^e = \frac{T_{xx}^e}{4\Delta^e} \bar{b}_k^e \bar{b}_l^e + \frac{T_{yy}^e}{4\Delta^e} \bar{c}_k^e \bar{c}_l^e, \tag{39}$$

$$g_{km}^e = \frac{T_{xx}^e}{4\Delta^e} \bar{b}_k^e \bar{b}_m^e + \frac{T_{yy}^e}{4\Delta^e} \bar{c}_k^e \bar{c}_m^e. \tag{40}$$

A small alteration in the integral formulations given by equations (30) and (31) is useful for computations and in developments further on. Because $\bar{N}_k^e + \bar{N}_l^e + \bar{N}_m^e = 1$, the terms

$$-\left(\frac{\partial \bar{N}_k^e}{\partial \bar{x}} + \frac{\partial \bar{N}_l^e}{\partial \bar{x}} + \frac{\partial \bar{N}_m^e}{\partial \bar{x}}\right)\hat{h}_k$$

and

$$-\left(\frac{\partial \bar{N}_k^e}{\partial \bar{y}} + \frac{\partial \bar{N}_l^e}{\partial \bar{y}} + \frac{\partial \bar{N}_m^e}{\partial \bar{y}}\right)\hat{h}_k$$

are both equal to zero and can be added into the terms

$$\frac{\partial \bar{N}_k^e}{\partial \bar{x}} \hat{h}_k + \frac{\partial \bar{N}_l^e}{\partial \bar{x}} \hat{h}_l + \frac{\partial \bar{N}_m^e}{\partial \bar{x}} \hat{h}_m$$

and

$$\frac{\partial \bar{N}_k^e}{\partial \bar{y}} \hat{h}_k + \frac{\partial \bar{N}_l^e}{\partial \bar{y}} \hat{h}_l + \frac{\partial \bar{N}_m^e}{\partial \bar{y}} \hat{h}_m$$

in equations (30) and (31), respectively. The resulting modifications of equations (30) and (31) are

$$\int\int_{\Delta^e} \frac{\partial \bar{N}_k^e}{\partial \bar{x}} T_{xx} \frac{\partial \hat{h}}{\partial \bar{x}} d\bar{x}d\bar{y} = \frac{T_{xx}^e}{4\Delta^e}\left[\bar{b}_k^e \bar{b}_l^e\left(\hat{h}_l - \hat{h}_k\right) + \bar{b}_k^e \bar{b}_m^e\left(\hat{h}_m - \hat{h}_k\right)\right] \tag{41}$$

and

$$\int\int_{\Delta^e} \frac{\partial \bar{N}_k^e}{\partial \bar{y}} T_{yy} \frac{\partial \hat{h}}{\partial \bar{y}} d\bar{x}d\bar{y} = \frac{T_{yy}^e}{4\Delta^e}\left[\bar{c}_k^e \bar{c}_l^e\left(\hat{h}_l - \hat{h}_k\right) + \bar{c}_k^e \bar{c}_m^e\left(\hat{h}_m - \hat{h}_k\right)\right], \tag{42}$$

which indicates that

$$g_{kk}^e = -g_{kl}^e - g_{km}^e. \tag{43}$$

The revised formulation, which was used by Narasimhan and others (1978, p. 875), saves both computer time and storage requirements because g_{kk}^e never need be explicitly computed using equation (38). An added advantage over the original formulation is that equations (41) and 42) generate less round-off error than equations (30) and (31) when solving the simultaneous systems of equations developed further on.

Substitution of equation (35) into equation (21) written for node k yields

$$\sum_{e_k} \int_0^{\Delta t_{n+1}} \sigma_{n+1} \left\{ c_{kk}^e \frac{d\hat{h}_k}{dt} + \left(g_{kk}^e + v_{kk}^e \right)\hat{h}_k + g_{kl}^e \hat{h}_l + g_{km}^e \hat{h}_m - \frac{1}{3}R^e\Delta^e H_k - \frac{1}{3}W^e\Delta^e - P_k^e \right.$$

$$\left. - \frac{1}{2}\left[\left(q_B L\right)_{kl} + \left(q_B L\right)_{km} \right] - \frac{1}{2}\left[\left(\alpha L\right)_{kl} + \left(\alpha L\right)_{km} \right]H_{Bk} \right\}dt' = 0. \qquad (44)$$

Equation (44) must apply to all N nodes of the finite-element mesh. These N equations can be written in matrix form as

$$\int_0^{\Delta t_{n+1}} \sigma_{n+1} \left[\underline{\underline{C}}\frac{d\hat{h}}{dt} + \left(\underline{\underline{G}} + \underline{\underline{V}} \right)\hat{\underline{h}} - \underline{B} \right]dt' = \int_0^{\Delta t_{n+1}} \sigma_{n+1}\left[\underline{\underline{C}}\frac{d\hat{h}}{dt} + \underline{\underline{A}}\hat{\underline{h}} - \underline{B} \right]dt' = \underline{0}, \qquad (45)$$

where

$$\underline{\underline{A}} = \underline{\underline{G}} + \underline{\underline{V}}, \qquad (46)$$

and doubly underscored letters indicate matrices and singly underscored letters indicate vectors. Entries of the matrices and vectors are defined as follows:

$$C_{ij} = \begin{cases} \sum_{e_i} c_{ij}^e, & i = j \\ 0, & i \neq j \end{cases}, \qquad (47)$$

$$V_{ij} = \begin{cases} \sum_{e_i} v_{ij}^e, & i = j \\ 0, & i \neq j \end{cases}, \qquad (48)$$

$$G_{ij} = \sum_{e_i} g_{ij}^e, \qquad (49)$$

$$B_i = \sum_{e_i} \left[\frac{1}{3}R^e\Delta^e H_i + \frac{1}{3}W^e\Delta^e + P_i^e + \frac{1}{2}\sum_{j'} \left(q_B L\right)_{ij'} + \frac{1}{2}\sum_{j'} (\alpha L)_{ij'} H_{Bi} \right], \qquad (50)$$

where the sum over j′ indicates the sum over the two nodes that are adjacent to node i in an element.

Specified-head boundaries were not considered in the preceding development. If node k was designated as a specified-head node, then equation (44) would be replaced by

$$\int_0^{\Delta t_{n+1}} \sigma_{n+1}\left(\hat{h}_k - H_{Bk} \right)dt' = 0, \qquad (51)$$

and this equation would replace equation k in matrix equation (45). Note that setting h_k equal to H_{Bk} is formally equivalent to letting $\alpha \to \infty$ at node k in equation (44).

18

Example of equation assembly

A simple finite-element mesh shown in figure 6 is used to demonstrate how the terms of equation (45) are assembled. Matrices $\underline{\underline{C}}$ and $\underline{\underline{A}}$, and vector $\underline{B}$, are assembled separately, then these are used to obtain the final system of equations.

Assembly is based on the patch of elements concept, where contributions to any equation i (that is, row i of $\underline{\underline{C}}$, $\underline{\underline{A}}$, or B) come from all elements sharing node i. By using this concept, $\underline{\underline{C}}$ can be assembled to yield: (Note: In the following equations all zero entries are left blank.)

$$
\underline{\underline{C}} =
\begin{array}{cccccc}
 & 1 & 2 & 3 & 4 & 5 \\
1 & \frac{1}{3}S^1 A^1 & & & & \\
2 & & \frac{1}{3}\left(S^1 A^1 + S^2 A^2\right) & & & \\
3 & & & \frac{1}{3}\left(S^1 A^1 + S^2 A^2 + S^3 A^3\right) & & \\
4 & & & & \frac{1}{3}\left(S^2 A^2 + S^3 A^3\right) & \\
5 & & & & & \frac{1}{3}S^3 A^3
\end{array},
$$
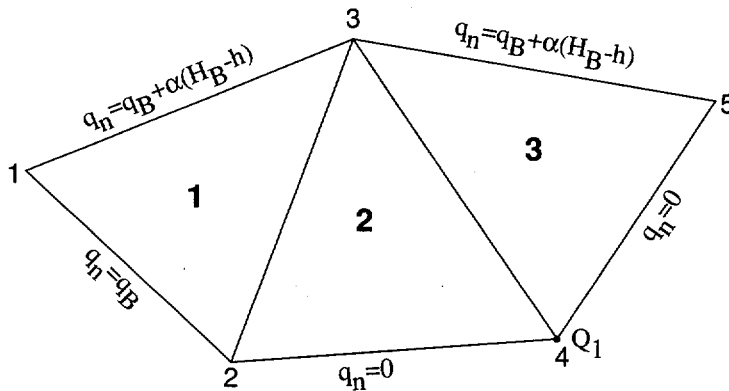
q_n = flow normal to an element boundary

q_B = specified flow normal to an element boundary

α = proportionality parameter for a Cauchy-type boundary condition

H_B = specified head at a boundary

h = hydraulic head

Q_1 = volumetric recharge from a well at node 4

Figure 6. Example of three elements and five nodes for demonstrating assembly of finite-element equations.

Matrix $\underline{\underline{A}}$ can be thought of as the sum of three matrices, a matrix composed of the g_{ij}^e ($i \neq j$) terms, a matrix composed of the $\frac{1}{3}R^e\Delta^e$ terms, and a matrix composed of the $\frac{1}{2}\left(\alpha L\right)_{ij}$, terms. These matrices are defined as $\underline{\underline{G}}$, $\underline{\underline{R}}$, and $\underline{\underline{\alpha}}$, respectively, where, from equations (37) and (48), $\underline{\underline{R}} + \underline{\underline{\alpha}} = \underline{\underline{V}}$. For the mesh shown in figure 6,

$$
\underline{\underline{G}} =
\begin{array}{c}
\quad \\ 1 \\ 2 \\ 3 \\ 4 \\ 5
\end{array}
\begin{bmatrix}
-\left(g_{12}^1+g_{13}^1\right) & g_{12}^1 & g_{13}^1 & & \\[2ex]
g_{21}^1 & -\left(g_{21}^1+g_{23}^1 + g_{23}^2+g_{24}^2\right) & g_{23}^1+g_{23}^2 & g_{24}^2 & \\[2ex]
g_{31}^1 & g_{32}^1+g_{32}^2 & -\left(g_{31}^1+g_{32}^1+g_{32}^2 + g_{34}^2+g_{34}^3+g_{35}^3\right) & g_{34}^2+g_{34}^3 & g_{35}^3 \\[2ex]
 & g_{42}^2 & g_{43}^2+g_{43}^3 & -\left(g_{42}^2+g_{43}^2 + g_{43}^3+g_{45}^3\right) & g_{45}^3 \\[2ex]
 & & g_{53}^3 & g_{54}^3 & -\left(g_{53}^3+g_{54}^3\right)
\end{bmatrix},
$$

with columns labeled $1, 2, 3, 4, 5$.

$$
\underline{\underline{R}} =
\begin{array}{c}
\quad \\ 1 \\ 2 \\ 3 \\ 4 \\ 5
\end{array}
\begin{bmatrix}
\frac{1}{3}R^1\Delta^1 & & & & \\[2ex]
 & \frac{1}{3}\left(R^1\Delta^1+R^2\Delta^2\right) & & & \\[2ex]
 & & \frac{1}{3}\left(R^1\Delta^1+R^2\Delta^2+R^3\Delta^3\right) & & \\[2ex]
 & & & \frac{1}{3}\left(R^2\Delta^2+R^3\Delta^3\right) & \\[2ex]
 & & & & \frac{1}{3}R^3\Delta^3
\end{bmatrix},
$$

with columns labeled $1, 2, 3, 4, 5$.

and

$$
\underline{\underline{\alpha}} =
\begin{array}{c}
\\
1 \\
2 \\
3 \\
4 \\
5
\end{array}
\begin{array}{ccccc}
1 & 2 & 3 & 4 & 5 \\
\end{array}
\begin{bmatrix}
\frac{1}{2}\big(\alpha L\big)_{13} & & & & \\
& & & & \\
& & \frac{1}{2}\Big[\big(\alpha L\big)_{31}+\big(\alpha L\big)_{35}\Big] & & \\
& & & & \\
& & & & \frac{1}{2}\big(\alpha L\big)_{53}
\end{bmatrix}
$$

Finally, the $\underline{B}$ vector, which contains all terms that do not multiply $\hat{\underline{h}}$ or $\hat{d\underline{h}}/dt$, is

$$
\underline{B} =
\begin{array}{c}
1 \\
\\
2 \\
\\
3 \\
\\
\\
4 \\
\\
5
\end{array}
\begin{bmatrix}
\frac{1}{3}R^1\Delta^1 H_1 + \frac{1}{3}W^1\Delta^1 + \frac{1}{2}\big(q_B L\big)_{12} + \frac{1}{2}\big(q_B L\big)_{13} + \frac{1}{2}\big(\alpha L\big)_{13}H_{B1} \\
\\
\frac{1}{3}\big(R^1\Delta^1 + R^2\Delta^2\big)H_2 + \frac{1}{3}\big(W^1\Delta^1 + W^2\Delta^2\big) + \frac{1}{2}\big(q_B L\big)_{21} \\
\\
\frac{1}{3}\big(R^1\Delta^1 + R^2\Delta^2 + R^3\Delta^3\big)H_3 + \frac{1}{3}\big(W^1\Delta^1 + W^2\Delta^2 + W^3\Delta^3\big) + \frac{1}{2}\Big[\big(q_B L\big)_{31} + \big(q_B L\big)_{35}\Big] \\
\quad + \frac{1}{2}\Big[\big(\alpha L\big)_{31} + \big(\alpha L_{35}\big)\Big]H_{B3} \\
\\
\frac{1}{3}\big(R^2\Delta^2 + R^3\Delta^3\big) + \frac{1}{3}\big(W^2\Delta^2 + W^3\Delta^3\big) + Q_1 \\
\\
\frac{1}{3}R^3\Delta^3 + \frac{1}{3}W^3\Delta^3 + \frac{1}{2}\big(q_B L\big)_{53} + \frac{1}{2}\big(\alpha L\big)_{53}H_{B5}
\end{bmatrix},
$$

and vectors $\hat{d\underline{h}}/dt$ and $\hat{\underline{h}}$ are

$$
\frac{\hat{d\underline{h}}}{dt} =
\begin{array}{c}
1 \\
2 \\
3 \\
4 \\
5
\end{array}
\begin{bmatrix}
\hat{dh}_1/dt \\
\hat{dh}_2/dt \\
\hat{dh}_3/dt \\
\hat{dh}_4/dt \\
\hat{dh}_5/dt
\end{bmatrix}
\quad \text{and} \quad
\hat{\underline{h}} =
\begin{array}{c}
1 \\
2 \\
3 \\
4 \\
5
\end{array}
\begin{bmatrix}
\hat{h}_1 \\
\hat{h}_2 \\
\hat{h}_3 \\
\hat{h}_4 \\
\hat{h}_5
\end{bmatrix}.
$$

The final set of equations corresponding to equation (45) can be written

$$\int_0^{\Delta t_{n+1}} \sigma_{n+1}\left[C_{11}\frac{d\hat{h}_1}{dt} + A_{11}\hat{h}_1 + A_{12}\hat{h}_2 + A_{13}\hat{h}_3 - B_1 \right] dt' = 0$$

$$\int_0^{\Delta t_{n+1}} \sigma_{n+1}\left[C_{22}\frac{d\hat{h}_2}{dt} + A_{21}\hat{h}_1 + A_{22}\hat{h}_2 + A_{23}\hat{h}_3 + A_{24}\hat{h}_4 - B_2 \right] dt' = 0$$

$$\int_0^{\Delta t_{n+1}} \sigma_{n+1}\left[C_{33}\frac{d\hat{h}_3}{dt} + A_{31}\hat{h}_1 + A_{32}\hat{h}_2 + A_{33}\hat{h}_3 + A_{34}\hat{h}_4 + A_{35}\hat{h}_5 - B_3 \right] dt' = 0$$

$$\int_0^{\Delta t_{n+1}} \sigma_{n+1}\left[C_{44}\frac{d\hat{h}_4}{dt} + A_{42}\hat{h}_2 + A_{43}\hat{h}_3 + A_{44}\hat{h}_4 + A_{45}\hat{h}_5 - B_4 \right] dt' = 0$$

$$\int_0^{\Delta t_{n+1}} \sigma_{n+1}\left[C_{55}\frac{d\hat{h}_5}{dt} + A_{53}\hat{h}_3 + A_{54}\hat{h}_4 + A_{55}\hat{h}_5 - B_5 \right] dt' = 0$$

where terms involving zero coefficients were omitted and an entry A_{ij} is $A_{ij} = G_{ij} + R_{ij} + \alpha_{ij}$.

There are no specified-head nodes in figure 6. If node 2 (for example) is designated as a specified-head node, then the second equation above is

replaced by $\int_0^{\Delta t_{n+1}} \sigma_{n+1}\left[\hat{h}_2 - H_{B2} \right] dt' = 0$ and $\hat{h}_2$ is replaced by H_{B2} in the

remaining equations, $i = 1, 3, 4$, and 5, so that the terms $A_{i2}H_{B2}$ are regarded

as knowns. To accomplish this, (1) all entries in row 2 and column 2 of matrices $\underline{\underline{G}}$, $\underline{\underline{C}}$, $\underline{\underline{R}}$, and $\underline{\underline{\alpha}}$ are set to zero except for entry (2,2) in matrix $\underline{\underline{\alpha}}$, which is set to unity, (2) row 2 in $\underline{B}$ is set to H_{B2}, and (3) all other rows

$i = 1, 3, 4$, and 5 in $\underline{B}$ have $A_{i2}H_{B2}$ subtracted from them.

Evaluation of time integral

Time integration of equation (45) is performed using a formula that is analogous to equation (25):

$$\int_0^{\Delta t_{n+1}} \left(\sigma_n\right)^p \left(\sigma_{n+1}\right)^q dt' = \frac{p!q!}{(p+q+1)!} \Delta t_{n+1}. \tag{52}$$

The simplest solution of equation (45) is obtained when coefficient matrices $\underline{\underline{C}}$ and $\underline{\underline{A}}$ and known vector $\underline{B}$ are constant in time. In this case, term by term integration of equation (45) yields

22

$$\int_0^{\Delta t_{n+1}} C_{ii} \frac{d\hat{h}_i}{dt} \sigma_{n+1} dt' = C_{ii} \int_0^{\Delta t_{n+1}} \left(\hat{h}_{i,n} \frac{d\sigma_n}{dt} + \hat{h}_{i,n+1} \frac{d\sigma_{n+1}}{dt} \right) \sigma_{n+1} dt'$$

$$= \frac{1}{2} C_{ii} \left(\hat{h}_{i,n+1} - \hat{h}_{i,n} \right), \tag{53}$$

$$\int_0^{\Delta t_{n+1}} A_{ij} \hat{h}_j \sigma_{n+1} dt' = A_{ij} \int_0^{\Delta t_{n+1}} \left(\hat{h}_{i,n} \sigma_n + \hat{h}_{i,n+1} \sigma_{n+1} \right) \sigma_{n+1} dt'$$

$$= \Delta t_{n+1} A_{ij} \left(\frac{1}{6} \hat{h}_{i,n} + \frac{1}{3} \hat{h}_{i,n+1} \right), \tag{54}$$

$$\int_0^{\Delta t_{n+1}} B_i \sigma_{n+1} dt' = B_i \int_0^{\Delta t_{n+1}} \sigma_{n+1} dt' = \frac{1}{2} \Delta t_{n+1} B_i. \tag{55}$$

Therefore, equation (45) is evaluated as

$$\underline{\underline{C}} \left(\hat{\underline{h}}_{n+1} - \hat{\underline{h}}_n \right) + \Delta t_{n+1} \underline{\underline{A}} \left(\frac{1}{3} \hat{\underline{h}}_n + \frac{2}{3} \hat{\underline{h}}_{n+1} \right) = \Delta t_{n+1} \underline{B}. \tag{56}$$

Solution of equation (56) produces round-off errors, which can be reduced by solving for a change in head between time levels rather than for the actual head values, $\hat{h}_{i,n+1}$. By defining $\underline{\delta}$ as 2/3 of the total head change between two time levels and substituting this into equation (56), a convenient equation for solution results. Thus, by defining

$$\underline{\delta} = \frac{2}{3} \left(\hat{\underline{h}}_{n+1} - \hat{\underline{h}}_n \right), \tag{57}$$

$\hat{\underline{h}}_{n+1} = \frac{3}{2} \underline{\delta} + \hat{\underline{h}}_n$ and equation (56) can be written in the form

$$\left[\frac{\underline{\underline{C}}}{(2/3)\Delta t_{n+1}} + \underline{\underline{A}} \right] \underline{\delta} = \underline{B} - \underline{\underline{A}} \hat{\underline{h}}_n. \tag{58}$$

Further reduction of round-off error is obtained by writing the diagonal terms of $\underline{\underline{G}}$ using equation (43) so that $\underline{\underline{G}} \hat{\underline{h}}_n$ can be written in terms of head differences of the form of equations (41) and (42).

Equations (57) and (58) are used to solve for head vectors $\hat{\underline{h}}_{n+1}$ at all time levels successively, starting with n = 0 at which $\underline{h}_0$ is the known initial condition. First, equation (58) is solved for $\underline{\delta}$ using one of the matrix solution routines discussed further on, and second, equation (57) is solved for $\hat{\underline{h}}_{n+1}$, which becomes $\hat{\underline{h}}_n$ for the next time level.

23

The finite-element in time method given by equation (56) is equivalent to the weighted finite-difference in time method,

$$\underline{\underline{C}}\left(\hat{\underline{h}}_{n+1} - \hat{\underline{h}}_n\right) + \Delta t_{n+1}\, \underline{\underline{A}}\left[(1-\theta)\hat{\underline{h}}_n + \theta\hat{\underline{h}}_{n+1}\right] = \Delta t_{n+1}\underline{B}, \qquad (59)$$

with weighting factor θ equal to 2/3. The weighted finite-difference in time method is unconditionally stable for $\theta \geq 1/2$ (Smith, 1965, p. 23-24), but Briggs and Dixon's (1968) criterion shows that use of $\theta < 1$ can cause oscillatory solutions if Δt_{n+1} is too large. Bettencourt and others (1981) reported very good accuracy and only slight oscillations in a solution obtained with the finite-element in time method ($\theta = 2/3$). In contrast, their solution to the same problem obtained with the well-known Crank-Nicolson method ($\theta = 1/2$) (Crank and Nicolson, 1947) exhibited large oscillations with little, if any, improvement in overall accuracy over the finite-element in time method. Numerical experiments conducted by the author also show that solutions are accurate and exhibit minimal oscillatory behavior if the sizes of time elements are not too large (which is problem dependent).

Time variability of B results if source-bed heads H, specified heads H_B, areal recharge W, or specified boundary flux q_B change with time. A simple method of approximating this time dependence in the finite-element equations is to assume linear time variability during each time element so that during time-element n+1

$$B_i = B_{i,n}\sigma_n + B_{i,n+1}\sigma_{n+1}. \qquad (60)$$

Thus, equation (55) is replaced with

$$\int_0^{\Delta t_{n+1}} B_i \sigma_{n+1}\, dt' = \int_0^{\Delta t_{n+1}} \left(B_{i,n}\sigma_n + B_{i,n+1}\sigma_{n+1}\right)\sigma_{n+1}\, dt'$$

$$= \frac{1}{6}\Delta t_{n+1}\left(B_{i,n} + 2B_{i,n+1}\right) = \frac{1}{2}\Delta t_{n+1}\bar{B}_i, \qquad (61)$$

where $\bar{B}_i$ is a weighted average value of B_i over timespan Δt_{n+1}, defined by

$$\bar{B}_i = \frac{1}{3}\left(B_{i,n} + 2B_{i,n+1}\right). \qquad (62)$$

Hence, time dependence of known heads and fluxes may be incorporated into equation (58) by replacing $\underline{B}$ with $\bar{\underline{B}}$.

Time variability of $\underline{\underline{C}}$, $\underline{\underline{A}}$, and B also results from processes such as unconfined flow, conversions from confined to unconfined flow (and vice versa), nonlinearity of stream-aquifer interactions, and discharges from springs, drains, or evapotranspiration. These types of time variabilities are treated in the sections covering these topics.

Mass-balance calculation

A mass balance based on equation (56) is needed to allow hydrologic budget analysis of the model and to assess the accuracy of the matrix solution methods discussed further on. Total quantities of water moved

during the timespan Δt_{n+1} are computed according to equation (61) as the product of weighted average discharges and Δt_{n+1}. To compute these totals, the mass-balance equations are formulated in terms of weighted average discharges and weighted average head, defined as

$$\bar{h} = \frac{1}{3}\hat{h}_n + \frac{2}{3}\hat{h}_{n+1}. \tag{63}$$

By employing equations (56), (62), and (63), along with the definitions of the quantities in these equations, the system of nodal mass-balance equations is written as

$$\frac{1}{2}\sum_{e_i} S^e \Delta^e \left(\bar{h}_i - \hat{h}_{i,n}\right) - \frac{1}{3}\sum_{e_i} R^e \Delta^e \left(\bar{H}_i - \bar{h}_i\right)\Delta t_{n+1} - \frac{1}{3}\sum_{e_i} \bar{W}^e \Delta^e \Delta t_{n+1} - \sum_{e_i} \bar{P}_i^e \Delta t_{n+1}$$

$$- \sum_{\substack{j=1 \\ i \neq j}}^{N} \sum_{e_i} g_{ij}^e \left(\bar{h}_j - \bar{h}_i\right)\Delta t_{n+1} - \bar{Q}_{Bi}\Delta t_{n+1} - \frac{1}{2}\sum_{j'} \left[\left(\bar{q}_B L\right)_{ij'}\right.$$

$$\left. + \left(\alpha L\right)_{ij'}\left(\bar{H}_{Bi} - \bar{h}_i\right)\right]\Delta t_{n+1} \simeq 0, \quad i = 1,2,\cdots,N, \tag{64}$$

where $\bar{Q}_{Bi} = 0$ unless node i is a specified-head node, in which case $\bar{Q}_{Bi}$ is the volumetric discharge across the node (positive for inflow) obtained by direct solution of equation (64) for $\bar{Q}_{Bi}$. Bars over quantities in equation (64) indicate weighted averages over time.

To obtain the total mass balance over the flow domain, equation (64) is summed over i. When this is done, it can be seen that

$$\sum_{i=1}^{N} \sum_{\substack{j=1 \\ i \neq j}}^{N} \sum_{e_i} g_{ij}^e \left(\bar{h}_j - \bar{h}_i\right) = 0,$$

because $g_{ij}^e = g_{ji}^e$ so that $g_{ij}^e \left(\bar{h}_j - \bar{h}_i\right) + g_{ji}^e \left(\bar{h}_i - \bar{h}_j\right) = 0$. Thus, the

components that should sum to give nearly zero are:

Total depletion or accretion of water in storage $= \frac{1}{2} \sum_{i=1}^{N} \sum_{e_i} S^e \Delta^e \left(\bar{h}_i - \hat{h}_{i,n}\right).$

Total leakage across confining units $= \frac{1}{3} \sum_{i=1}^{N} \sum_{e_i} R^e \Delta^e \left(\bar{H}_i - \bar{h}_i\right)\Delta t_{n+1}.$

Total areal recharge or discharge $= \frac{1}{3} \sum_{i=1}^{N} \sum_{e_i} \bar{W}^e \Delta^e \Delta t_{n+1}.$

25

Total water pumped into or out of wells $= \sum\limits_{i=1}^{N} \sum\limits_{e_i} \bar{P}_i^e \Delta t_{n+1} = \sum\limits_{j=1}^{P} \bar{Q}_j \Delta t_{n+1}$.

Total water crossing specified-head boundaries $= \sum\limits_{i=1}^{N} \bar{Q}_{Bi} \Delta t_{n+1}$.

Total water crossing Cauchy-type boundaries

$$= \frac{1}{2} \sum\limits_{i=1}^{N} \sum\limits_{j}' \left[\left(\bar{q}_B L \right)_{ij'} + \left(\alpha L \right)_{ij'} \left(\bar{H}_{Bi} - \bar{h}_i \right) \right] \Delta t_{n+1}.$$

Average volumetric flow rates in time element n+1 can be obtained by dividing the components by Δt_{n+1}, and running totals over time can be obtained by summing the components over all preceding time elements. The mass imbalance in time element n+1 is obtained by summing the components, and a running mass imbalance is obtained by summing mass imbalances over all preceding time elements.

EXTENSIONS OF THE BASIC EQUATIONS

Unconfined flow

When equation (1) is applied to areal flow in an unconfined aquifer by using the Dupuit approximation (Bear, 1979, p. 111-114), transmissivities are functions of the current saturated thickness of the aquifer, as follows:

$$\begin{aligned} T &= Kb \\ &= K\left(h - z_b \right), \end{aligned} \tag{65}$$

where b is the saturated thickness $h - z_b$ of the aquifer, h is the elevation of the water table above some datum, z_b is the elevation of the aquifer bottom referred to the same datum, and subscripts $\bar{x}$ and $\bar{y}$ were omitted from T and K for simplicity. Because b is head dependent and varies in time, equation (1) is nonlinear, with transmissivities that are head dependent and vary in time.

Time variance of the transmissivities can be handled in the same manner as time variance of B_i. That is, the G_{ij} coefficients, which contain the transmissivities, can be written for time element n+1 as

$$G_{ij} = G_{ij,n} \sigma_n + G_{ij,n+1} \sigma_{n+1}, \tag{66}$$

so that, by using the relationship $A_{ij} = G_{ij} + V_{ij}$, equation (54) is replaced with

$$\int_0^{\Delta t_{n+1}} \hat{A}_{ij} h_j \sigma_{n+1} dt'$$

26

$$= \int_0^{\Delta t_{n+1}} \left(G_{ij,n}\sigma_n + G_{ij,n+1}\sigma_{n+1} + V_{ij} \right) \left(\hat{h}_{j,n}\sigma_n + \hat{h}_{j,n+1}\sigma_{n+1} \right) \sigma_{n+1} dt'$$

$$= \frac{1}{12}\Delta t_{n+1}\left(G_{ij,n} + G_{ij,n+1} + 2V_{ij} \right)\hat{h}_{j,n}$$

$$+ \frac{1}{12}\Delta t_{n+1}\left(G_{ij,n} + 3G_{ij,n+1} + 4V_{ij} \right)\hat{h}_{j,n+1}. \tag{67}$$

For an aquifer that remains unconfined (that is, h never exceeds the elevation of the base of an overlying confining bed) throughout the simulation period, matrix $\underline{\underline{C}}$ is modified by replacing the storage coefficient, S^e, in each element by specific yield, S_y^e. Therefore, $\underline{\underline{C}}$ is constant in time. Conversions from confined to unconfined flow (and vice versa) and their effect on $\underline{\underline{C}}$ is discussed in a later section.

Use of equation (67) in place of equation (54) modifies equation (56) to

$$\underline{\underline{C}}\left(\hat{\underline{h}}_{n+1} - \hat{\underline{h}}_n \right) + \frac{1}{6}\Delta t_{n+1}\left(\underline{\underline{G}}_n + 3\underline{\underline{G}}_{n+1} + 4\underline{\underline{V}} \right)\hat{\underline{h}}_{n+1}$$

$$+ \frac{1}{6}\Delta t_{n+1}\left(\underline{\underline{G}}_n + \underline{\underline{G}}_{n+1} + 2\underline{\underline{V}} \right)\hat{\underline{h}}_n = \Delta t_{n+1}\bar{\underline{B}}, \tag{68}$$

where equation (61) was used for B. Equation (68) can be written in a form analogous to equation (58) by using equation (57) and weighted average values of G_{ij}, defined as

$$\tilde{G}_{ij} = \frac{1}{4}\left(G_{ij,n} + 3G_{ij,n+1} \right) \tag{69}$$

and

$$\bar{G}_{ij} = \frac{1}{3}\left(G_{ij,n} + 2G_{ij,n+1} \right). \tag{70}$$

Thus,

$$\left[\frac{\underline{\underline{C}}}{(2/3)\Delta t_{n+1}} + \tilde{\underline{\underline{G}}} + \underline{\underline{V}} \right]\underline{\delta} = \bar{\underline{B}} - \left(\bar{\underline{\underline{G}}} + \underline{\underline{V}} \right)\hat{\underline{h}}_n. \tag{71}$$

Define the terms

$$d_{ij}^e = \frac{K_{xx}^e}{4\Delta^e}b_i^e b_j^e + \frac{K_{yy}^e}{4\Delta^e}c_i^e c_j^e, \quad i \neq j. \tag{72}$$

Then an off-diagonal element of $\underline{\underline{G}}$ is given by

$$G_{ij} = \sum_{e_i}\left(\frac{1}{3}\sum_s b_s \right)d_{ij}^e, \quad i \neq j, \tag{73}$$

27

where b_s is the aquifer thickness at node $s = k,l,m$, and $\frac{1}{3}\sum_s b_s$ is the average aquifer thickness in element e, assuming that thickness varies linearly over the element. Computation of G_{ij} using equation (73) requires reassembling G_{ij} element by element each time thickness b_s is changed. An approximation that Cooley (1971) found to be good for a subdomain finite-element solution of axisymmetric, variably saturated flow problems avoids this reassembly. The approximation is to evaluate the head-dependent coefficient in G_{ij} using the head half-way between nodes i and j. For the present problem, this approximation is equivalent to replacing $\frac{1}{3}\sum_s b_s$ in equation (73) with the average thickness between nodes i and j. Thus, the approximation is

$$G_{ij} \simeq \frac{1}{2}\left(b_i + b_j\right)\sum_{\substack{e \\ i}} d^e_{ij} = \frac{1}{2}\left(b_i + b_j\right)D_{ij}, \qquad (74)$$

where

$$D_{ij} = \sum_{\substack{e \\ i}} d^e_{ij}. \qquad (75)$$

Because aquifer thickness is dependent upon head, a means of predicting this thickness at an advanced time level, n+1, is needed prior to solving equation (71). A simple and effective method is the predictor-corrector technique described by Douglas and Jones (1963). In the predictor step of this two-step process, the previously calculated thicknesses are used in $\underline{\underline{G}}$ to form an equation of the same form as equation (58). This equation is then solved for the head changes over the time element, and heads at the advanced time level are predicted based on equation (57). Aquifer thicknesses are then updated using the predicted head changes, and these updated thicknesses are used to form $\underline{\underline{\widetilde{G}}}$ and $\underline{\underline{\bar{G}}}$. These matrices are used in equation (71) to solve for the head changes over the time element, which is the corrector step.

The predictor step is expressed by the following equations. Based on equation (58),

$$\left(\frac{\underline{\underline{C}}}{(2/3)\Delta t_{n+1}} + \underline{\underline{G}}_n + \underline{\underline{V}}\right)\underline{\delta}^* = \underline{\bar{B}} - \left(\underline{\underline{G}}_n + \underline{\underline{V}}\right)\underline{\hat{h}}_n \qquad (76)$$

where $\underline{\delta}^*$ is the predicted head-change vector and

$$G_{ij,n} = \frac{1}{2}\left(b_{i,n} + b_{j,n}\right)D_{ij}, \quad i \neq j. \qquad (77)$$

The thickness $b_{i,n}$ is

$$b_{i,n} = \hat{h}_{i,n} - z_{bi}, \qquad (78)$$

and $b_{j,n}$ is defined similarly. From equation (57), the predicted head vector, $\underline{\hat{h}}^*_n$, at time level t_{n+1} is

$$\hat{\underline{h}}^* = \frac{3}{2}\underline{\delta}^* + \hat{\underline{h}}_n. \tag{79}$$

The corrector step uses heads $\hat{\underline{h}}^*$ to form the following corrector equations. Based on equation (71)

$$\left[\frac{\underline{\underline{C}}}{(2/3)\Delta t_{n+1}} + \underline{\underline{\tilde{G}}}^* + \underline{\underline{V}}\right]\underline{\delta} = \underline{\bar{B}} - \left(\underline{\underline{\bar{G}}}^* + \underline{\underline{V}}\right)\hat{\underline{h}}_n, \tag{80}$$

where $\underline{\underline{\tilde{G}}}^*$ and $\underline{\underline{\bar{G}}}^*$ are the approximations of $\underline{\underline{\tilde{G}}}$ and $\underline{\underline{\bar{G}}}$, defined, using equations (69), (70), (74), and (79), as

$$\tilde{G}_{ij}^* = \frac{1}{4}\left\{\frac{1}{2}\left(b_{i,n} + b_{j,n}\right) + 3\left[\frac{1}{2}\left(b_{i,n} + \frac{3}{2}\delta_i^* + b_{j,n} + \frac{3}{2}\delta_j^*\right)\right]\right\}D_{ij}$$

$$= \frac{9}{16}\left[\delta_i^* + \delta_j^* + \frac{8}{9}\left(b_{i,n} + b_{j,n}\right)\right]D_{ij}, \quad i \neq j, \tag{81}$$

and

$$\bar{G}_{ij}^* = \frac{1}{3}\left\{\frac{1}{2}\left(b_{i,n} + b_{j,n}\right) + 2\left[\frac{1}{2}\left(b_{i,n} + \frac{3}{2}\delta_i^* + b_{j,n} + \frac{3}{2}\delta_j^*\right)\right]\right\}D_{ij}$$

$$= \frac{1}{2}\left(\delta_i^* + \delta_j^* + b_{i,n} + b_{j,n}\right)D_{ij}, \quad i \neq j. \tag{82}$$

In practice, to reduce round-off error $\left\{\underline{\underline{C}}/\left[(2/3)\Delta t_{n+1}\right] + \underline{\underline{\tilde{G}}}^* + \underline{\underline{V}}\right\}\underline{\delta}^*$ is subtracted from both sides of equation (80) to create a residual form of the equation so that $\underline{\delta} - \underline{\delta}^*$ is actually solved for. Head change $\underline{\delta}$ is then directly computed as $\underline{\delta} = \left(\underline{\delta} - \underline{\delta}^*\right) + \underline{\delta}^*$. The head at the end of the time element is calculated using equation (57).

Mass-balance calculations could be based on equation (80). However, more information about the accuracy of the predictor-corrector scheme can be obtained by computing mass-balance components from an equation derived from equation (71) in which $\underline{\underline{\tilde{G}}}$ and $\underline{\underline{\bar{G}}}$ are computed using $\hat{\underline{h}}_{n+1}$, which is

$$\frac{\underline{\underline{C}}}{(2/3)\Delta t_{n+1}}\underline{\delta} + \left(\underline{\underline{\tilde{G}}} + \underline{\underline{V}}\right)\underline{\delta} + \left(\underline{\underline{\bar{G}}} + \underline{\underline{V}}\right)\hat{\underline{h}}_n - \underline{\bar{B}}$$

$$= \frac{\underline{\underline{C}}}{(2/3)\Delta t_{n+1}}\underline{\delta} + \left(\underline{\underline{\bar{G}}} + \underline{\underline{V}}\right)\underline{\bar{h}} + \left(\underline{\underline{\tilde{G}}} - \underline{\underline{\bar{G}}}\right)\underline{\delta} - \underline{\bar{B}}$$

$$\simeq \underline{0}, \tag{83}$$

where, by employing equations (57), (69), (70), and (74), it can be verified that

$$\widetilde{G}_{ij} - \bar{G}_{ij} = \frac{1}{16}\left(\delta_i + \delta_j\right)D_{ij}. \tag{84}$$

Equation (83) becomes equation (64) when $\underline{\underline{G}}$ is not time variant.

Because $\widetilde{\underline{\underline{G}}}$ and $\bar{\underline{\underline{G}}}$ are computed using $\hat{\underline{h}}_{n+1}$ instead of $\hat{\underline{h}}^*$, $\underline{\delta}$ is only the approximate solution of equation (83). If the time element Δt_{n+1} is too large, then the approximate solution will be poor, and this will result in a large mass imbalance as computed using equation (83). In this case, the time-element size should be reduced.

The algorithm used to implement the predictor-corrector method is summarized by the following steps.

1. Predictor: Solve equation (76) for $\underline{\delta}^*$, and solve equation (79) for $\hat{\underline{h}}^*$. Then compute a predicted average head vector $\bar{\underline{h}}^*$ using $\bar{\underline{h}}^* = \underline{\delta}^* + \hat{\underline{h}}_n$.

2. Compute elements of $\widetilde{\underline{\underline{G}}}^*$ and $\bar{\underline{\underline{G}}}^*$ using equations (81) and (82).

3. Corrector: Solve the residual form of equation (80) for $\underline{\delta} - \underline{\delta}^*$, and compute the average head $\bar{\underline{h}}$ using $\bar{\underline{h}} = \underline{\delta} - \underline{\delta}^* + \bar{\underline{h}}^*$, which is obtained by combining $\bar{\underline{h}}^* = \underline{\delta}^* + \hat{\underline{h}}_n$ and $\bar{\underline{h}} = \underline{\delta} + \hat{\underline{h}}_n$. Compute $\underline{\delta} = \left(\underline{\delta} - \underline{\delta}^*\right) + \underline{\delta}^*$.

4. Compute the weighted average mass-balance components using equation (83).

5. Update the aquifer thickness for the next time element using

$$\underline{b}_{n+1} = \frac{3}{2}\underline{\delta} + \underline{b}_n, \tag{85}$$

which is obtained by using the definition of b and equation (57).

6. Compute $\hat{\underline{h}}_{n+1}$ using

$$\hat{\underline{h}}_{n+1} = \frac{1}{2}\underline{\delta} + \bar{\underline{h}}, \tag{86}$$

which is derived by combining equations (57) and (63).

7. Advance the time-element index, n, define a new Δt_{n+1}, and return to 1, unless the simulation time limit has been reached.

Drying and resaturation of nodes

If the water table declines to the base of the aquifer at a node during a simulation, then the node is said to "go dry" (figure 7). Although the aquifer thickness at the node is zero, horizontal flow to or from adjacent saturated nodes can still exist by virtue of equation (74). Thus, the node should remain active and hydraulic head at the node should still be calculated. An approximate method of simulating this process is to solve the finite-element equation (equation (71)) using zero aquifer thickness at dry nodes. Because the storage term C_{ii} is not altered when a node goes

dry, this use of equation (71) assumes (1) that water is released from or taken into storage in the aquifer where the saturated thickness is greater than zero and in the material underlying the aquifer where the saturated thickness is zero, (2) that the specific yield of the aquifer and underlying materials is the same, and (3) that both of these sources of water contribute to horizontal flow in the aquifer. If the material underlying the aquifer is explicitly incorporated into the simulation as a confining unit (R > 0), then this unit serves to vertically convey water to or from the aquifer in addition to releasing or taking on stored water at dry nodes. Because dry nodes are active nodes in the flow system, solution of equation (71) can produce heads that decline below the aquifer base so that the water table can move laterally away from the dry nodes (figure 7). In this case, the computed heads at the dry nodes can be thought of as effective heads that allow approximation of horizontal flow in the aquifer near the dry nodes. If all nodes j adjacent to a dry node i also are dry, then, from equation (74), all $G_{ij} = 0$, and horizontal flow in the aquifer near the dry node ceases. Water table decline at the dry node will also cease unless the underlying unit is a confining unit (R > 0) and $\hat{h}_i > H_i$, or the node is on a Cauchy-type boundary (α>0) and $\hat{h}_i > H_{Bi}$, or known sources and sinks in B_i are negative, which is treated below.

If a pumping well (or other specified sink) is located at a node that goes dry, then the net discharge at the node is too large for the aquifer to sustain. This incompatibility must be rectified by the investigator.

$\hat{h}_{i,r}$ ($i=k,l,m$; $r=n,n+1$)=computed hydraulic head at node i and time level r

$b_{i,r}$ ($i=k,l,m$; $r=n,n+1$)=saturated thickness at node i and time level r
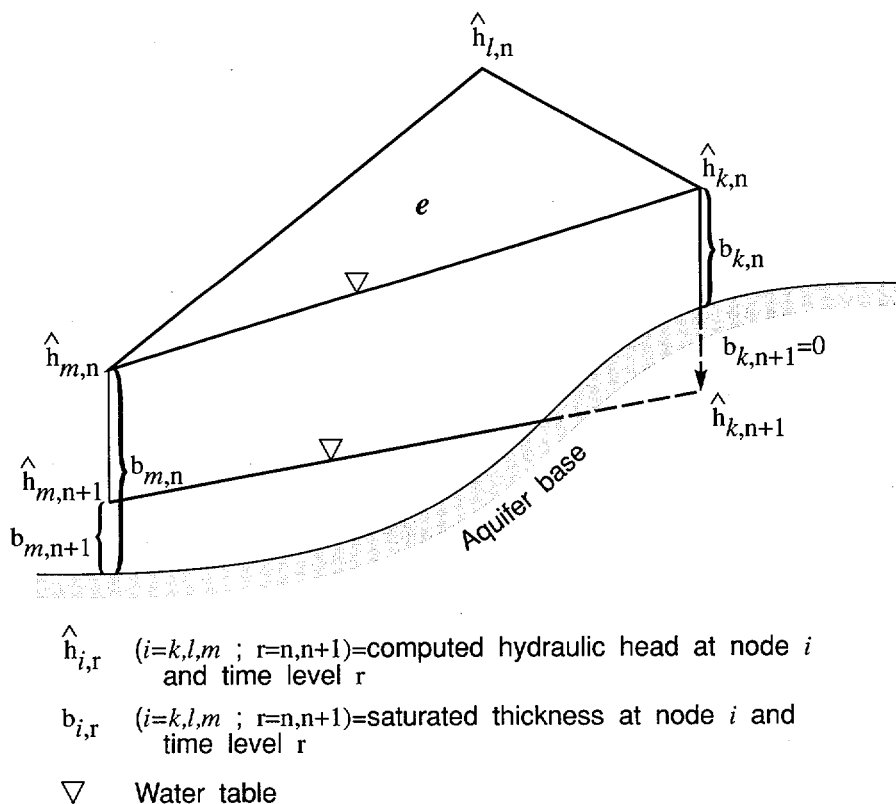
∇ Water table

Figure 7. Node k in element e drying up as the water table declines during simulation.

However, to allow the simulation to continue, and to suggest how much discharge the model can supply, the following automatic procedure is followed. If the net specified flux is negative at a node that is predicted to go dry at the end of a predictor step, then the net flux at the node is permanently cut in half beginning with the corrector step. Discharge can continue to be reduced by a factor of two on subsequent time elements if the node continues to go dry. If this procedure is insufficient to maintain a positive saturated thickness, the head at the node may drop below the aquifer base, and the dry node may at least temporarily supply water to the sink, which is physically unrealistic. However, the results of this procedure should indicate to the investigator how the model input could be changed to yield a physically compatible situation.

Combined confined and unconfined flow

Equation (1) can be applied to a problem where there is confined flow in some areas of the aquifer and unconfined flow in other areas. In this case, conversion can take place from one type of flow to another at any time or place in the aquifer (figure 8). Where flow is confined, the storage coefficient in equation (1) is the artesian storage coefficient, S, and transmissivity is constant in time. Where flow is unconfined, the storage coefficient is the specific yield, S_y, and transmissivity is time variant, as given by equation (65).

If flow at node i converts from confined to unconfined, or vice versa, during time-element n+1, the time interval Δt_{n+1} is divided into two

subintervals, $\theta_i \Delta t_{n+1}$ and $(1-\theta_i)\Delta t_{n+1}$, where θ_i is the unknown proportionate

point in the time interval when node i converts. The storage-change term analogous to equation (53) is then approximated as the sum of the two storage-change terms resulting from treating the two subintervals as subelements, each having its own basis functions and approximate function for hydraulic head. Thus, for the subinterval $\theta_i \Delta t_{n+1}$

$$\hat{h}_i = \hat{h}_{i,n}\sigma_n^{(1)} + z_{ti}\sigma_\theta^{(1)}, \tag{87}$$

where

$$\sigma_n^{(1)} = 1 - \frac{\sigma_{n+1}}{\theta_i}, \tag{88}$$

$$\sigma_\theta^{(1)} = \frac{\sigma_{n+1}}{\theta_i}, \tag{89}$$

and z_{ti} is the elevation of the top of the aquifer and equals the head at

node i at time $t_n + \theta_i \Delta t_{n+1}$. For the subinterval $\left(1-\theta_i\right)\Delta t_{n+1}$,

$$\hat{h}_i = z_{ti}\sigma_\theta^{(2)} + \hat{h}_{i,n+1}\sigma_{n+1}^{(2)}, \tag{90}$$

where

$$\sigma_\theta^{(2)} = \frac{\sigma_n}{1-\theta_i}, \tag{91}$$

$$\sigma_{n+1}^{(2)} = 1 - \frac{\sigma_n}{1-\theta_i}. \tag{92}$$
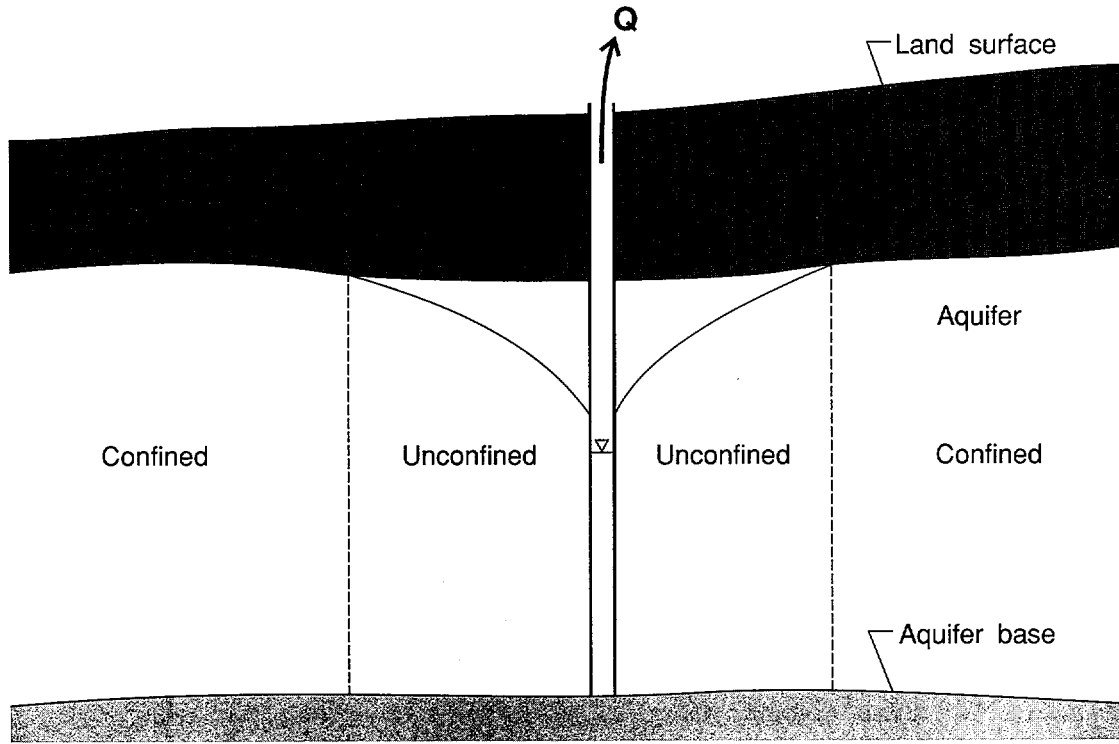
▽ Locates potentiometric surface in the confined part of the aquifer and water-table position in the unconfined part

Figure 8. Cross section showing conversion from confined to unconfined flow at time t near a well pumped at volumetric rate Q.

By using equations (87) through (92), the integrals involving the storage term C_{ii} (see equation (53)) for both subintervals are formed and evaluated as

$$\int_0^{\theta_i \Delta t_{n+1}} C_{ii} \frac{d\hat{h}_i}{dt} \sigma_\theta^{(1)} dt' = C_{ii}^{(1)} \int_0^{\theta_i \Delta t_{n+1}} \left[\hat{h}_{i,n} \frac{d\sigma_n^{(1)}}{dt} + z_{ti} \frac{d\sigma_\theta^{(1)}}{dt} \right] \sigma_\theta^{(1)} dt'$$

$$= \frac{1}{2} C_{ii}^{(1)} \left[z_{ti} - \hat{h}_{i,n} \right] \tag{93}$$

and

$$\int_{\theta_i \Delta t_{n+1}}^{\Delta t_{n+1}} C_{ii} \frac{d\hat{h}_i}{dt} \sigma_{n+1}^{(2)} dt' = C_{ii}^{(2)} \int_{\theta_i \Delta t_{n+1}}^{\Delta t_{n+1}} \left[z_{ti} \frac{d\sigma_\theta^{(2)}}{dt} + \hat{h}_{i,n+1} \frac{d\sigma_{n+1}^{(2)}}{dt} \right] \sigma_{n+1}^{(2)} dt'$$

$$= \frac{1}{2} C_{ii}^{(2)} \left[\hat{h}_{i,n+1} - z_{ti} \right], \tag{94}$$

where $C_{ii}^{(1)}$ is the storage term before conversion, and $C_{ii}^{(2)}$ is the storage term after conversion.

33

The required finite-element equation for node i and time-element n+1 is obtained by summing the equations for the two subintervals, so that the storage-change term is the sum of equations (93) and (94). This sum is one-half the total change in water stored during time element n+1 (Prickett and Lonnquist, 1971, p. 40-41; Trescott and others, 1976, p. 10-11; and Wilson and others, 1979, p. 52). It is instructive to note that this sum can be written as

$$\frac{1}{2}c_{ii}^{(1)}\left[z_{ti} - \hat{h}_{i,n}\right] + \frac{1}{2}c_{ii}^{(2)}\left[\hat{h}_{i,n+1} - z_{ti}\right]$$

$$= \frac{1}{2}c_{ii}^{(1)}\left[z_{ti} - \hat{h}_{i,n}\right] + \frac{1}{2}c_{ii}^{(2)}\left[\hat{h}_{i,n+1} - \hat{h}_{i,n}\right] - \frac{1}{2}c_{ii}^{(2)}\left[z_{ti} - \hat{h}_{i,n}\right]$$

$$= \frac{1}{2}\left[c_{ii}^{(2)} + \left[c_{ii}^{(1)} - c_{ii}^{(2)}\right]\hat{\theta}_i\right]\left[\hat{h}_{i,n+1} - \hat{h}_{i,n}\right], \qquad (95)$$

where

$$\hat{\theta}_i = \frac{z_{ti} - \hat{h}_{i,n}}{\hat{h}_{i,n+1} - \hat{h}_{i,n}}. \qquad (96)$$

Therefore, if head is assumed to vary linearly within time-element n+1, $\hat{\theta}_i$ is an estimate of θ_i and the term in brackets defines an effective storage coefficient for time-element n+1.

By making use of equations (95) and (96) and approximating the sum of terms of the form of equation (67) for the two subintervals by the analogous term (equation 67) for the entire time element, an equation of the form of equation (68) may be written to include the possibility of one or more conversions within time-element n+1 as

$$\underline{\underline{c}}^{(2)}\left[\hat{\underline{h}}_{n+1} - \hat{\underline{h}}_n\right] + \left[\underline{\underline{c}}^{(1)} - \underline{\underline{c}}^{(2)}\right]\left[\underline{z}_t - \hat{\underline{h}}_n\right] + \frac{1}{6}\Delta t_{n+1}\left[\underline{\underline{G}}_n + 3\underline{\underline{G}}_{n+1} + 4\underline{\underline{V}}\right]\hat{\underline{h}}_{n+1}$$

$$+ \frac{1}{6}\Delta t_{n+1}\left[\underline{\underline{G}}_n + \underline{\underline{G}}_{n+1} + 2\underline{\underline{V}}\right]\hat{\underline{h}}_n = \Delta t_{n+1}\underline{\underline{B}}, \qquad (97)$$

where $c_{jj}^{(1)} = c_{jj}^{(2)}$ for all nodes that do not convert in the time interval Δt_{n+1}, and $\underline{z}_t$ is the vector of nodal aquifer-top elevations. Hence, the equation to replace equation (71) is

$$\left[\frac{\underline{\underline{c}}^{(2)}}{(2/3)\Delta t_{n+1}} + \underline{\underline{\tilde{G}}} + \underline{\underline{V}}\right]\underline{\delta} = \underline{\underline{B}} - \left[\underline{\underline{\tilde{G}}} + \underline{\underline{V}}\right]\hat{\underline{h}}_n + \frac{\underline{\underline{c}}^{(2)} - \underline{\underline{c}}^{(1)}}{\Delta t_{n+1}}\left[\underline{z}_t - \hat{\underline{h}}_n\right]. \qquad (98)$$

The predictor-corrector method is used to solve equation (98). For the predictor step, the predicted head vector $\hat{\underline{h}}^*$ is obtained using equations (76) and (79) with $\underline{\underline{c}} = \underline{\underline{c}}^{(1)}$. These predicted heads are used to determine which nodes, if any, convert during the time interval and to estimate the saturated thickness for all nodes that are either unconfined during the entire time interval or convert from confined to unconfined conditions during the time interval.

Predicted heads $\hat{h}_i^*$ were found to be poor estimates of $\hat{h}_{i,n+1}$ for nodes that convert from confined to unconfined conditions; thus, they cannot be directly used to calculate saturated thickness for the corrector step. This problem occurs because the artesian storage coefficient is usually several orders of magnitude smaller than the specific yield, so that unless $\hat{\theta}_i \approx 1$,

34

the change from $c_{ii}^{(1)}$ to $c_{ii}^{(2)} + \left[c_{ii}^{(1)} - c_{ii}^{(2)}\right]\hat{\theta}_i$ is large. It was also found

that total storage changes during confined to unconfined conversions are usually determined much more accurately in the predictor step than the head values. Therefore, an expression that can be used to revise a predicted

head can be developed by equating the predicted change in storage, $c_{ii}^{(1)}\left(\hat{h}_i^* - \right.$

$\left.\hat{h}_{i,n}\right)$, to the change in storage calculated with the conversion, to obtain

$$c_{ii}^{(1)}\left[z_{ti} - \hat{h}_{i,n}\right] + c_{ii}^{(2)}\left[\hat{h}_i' - z_{ti}\right] = c_{ii}^{(1)}\left[\hat{h}_i^* - \hat{h}_{i,n}\right], \qquad (99)$$

where $\hat{h}_i'$ is the revised predicted head. Solution of equation (99) for $\hat{h}_i'$
yields

$$\hat{h}_i' = \frac{c_{ii}^{(1)}}{c_{ii}^{(2)}}\left[\hat{h}_i^* - z_{ti}\right] + z_{ti}. \qquad (100)$$

For confined to unconfined conversions, $c_{ii}^{(1)} \ll c_{ii}^{(2)}$ so that $\hat{h}_{i,n} - \hat{h}_i'$ is

much smaller than $\hat{h}_{i,n} - \hat{h}_i^*$. However, if $\hat{h}_i^*$ predicts a conversion, so
will $\hat{h}_i'$.

The corrector step incorporates any possible conversions. If $\hat{h}_i^* < z_{ti}$

$< \hat{h}_{i,n}$ or $\hat{h}_{i,n} < z_{ti} < \hat{h}_i^*$, then a conversion is assumed to have taken place.
The corrector equation is

$$\left[\frac{\underline{\underline{C}}^{(2)}}{(2/3)\Delta t_{n+1}} + \underline{\underline{\tilde{G}}}^* + \underline{\underline{V}}\right]\underline{\delta} = \underline{\bar{B}} - \left[\underline{\underline{\tilde{G}}}^* + \underline{\underline{V}}\right]\hat{\underline{h}}_n + \frac{\underline{\underline{C}}^{(2)} - \underline{\underline{C}}^{(1)}}{\Delta t_{n+1}}\left[\underline{z}_t - \hat{\underline{h}}_n\right], \qquad (101)$$

where entries $\tilde{G}_{ij}^*$ and $\bar{G}_{ij}^*$ are computed using $\hat{h}_i^*$ or $\hat{h}_i'$ as appropriate.

As in the case of purely unconfined flow, the magnitudes of the errors generated by the predictor-corrector method are indicated by the mass-balance errors. If the errors are large, then the time element sizes should be reduced.

Point head-dependent discharge (springs and drainage wells)

The discharge rate from springs or drainage wells varies with the head in the aquifer and declines to zero as the head declines to some elevation z_p (figure 9). Discharge is zero as long as the head remains below z_p. Discharges from springs or drainage wells may be simulated by adding a point head-dependent sink function to equation (1) to give an equation of the form

$$\frac{\partial}{\partial x}\left[T_{xx}\frac{\partial h}{\partial x} + T_{xy}\frac{\partial h}{\partial y}\right] + \frac{\partial}{\partial y}\left[T_{yx}\frac{\partial h}{\partial x} + T_{yy}\frac{\partial h}{\partial y}\right]$$
$$+ R(H-h) + S_p + W + P = S\frac{\partial h}{\partial t}. \qquad (102)$$

The term S_p is the sink function, given by

$$S_p = \sum_{j=1}^{P_p}\delta\left(x-a_{pj}'\right)\delta\left(y-b_{pj}'\right)Q_{pj}, \qquad (103)$$
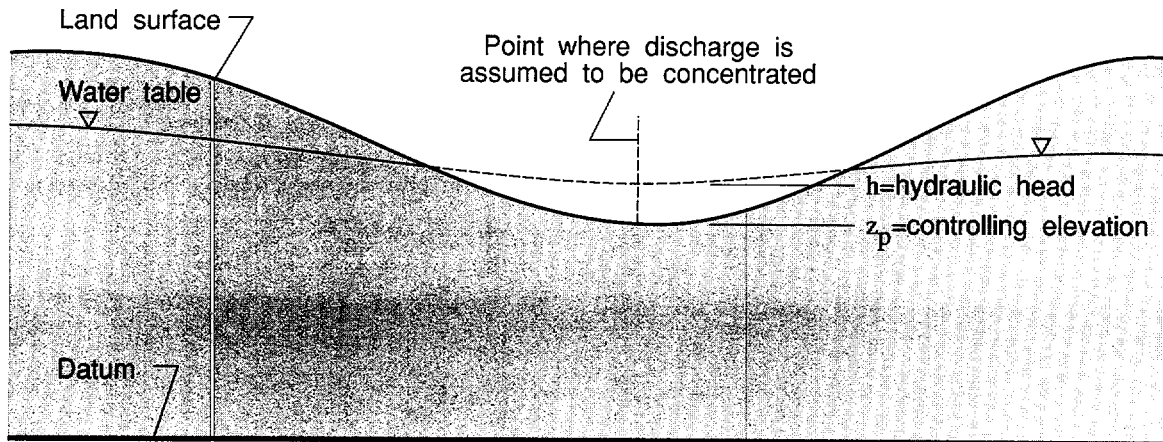
Figure 9. Cross section showing configuration of water-table position, ▽, and controlling elevation for point head-dependent discharge functions.

where Q_{pj} is the volumetric rate of head-dependent discharge [length^3/time] (negative for a sink) at point $\left(a'_{pj}, b'_{pj}\right)$, and there are p_p such points. Discharge Q_{pj} is assumed to vary linearly with head as long as head is greater than z_p. Thus, it is calculated from

$$Q_{pj} = \begin{cases} C_{pj}\left(z_p - h\right), & h > z_p \\ 0 & , h \leq z_p \end{cases}, \tag{104}$$

where C_{pj} is a function of hydraulic conductance of aquifer materials in the vicinity of the spring or drainage well [length^2/time]. Nonlinearity of the sink function results from the fact that the form of the function is dependent on the head in the aquifer.

To incorporate equations (103) and (104) into the finite-element equations, it is assumed that the point sinks are located only at node points. Hence, spatial finite-element discretization is applied to equations (103) and (104) to yield

$$\sum_e \int_{\Delta_e} \int \tilde{N}_i^e \sum_{j=1}^{p_p} \delta\left(\bar{x} - \bar{a}'_{pj}\right) \delta\left(\bar{y} - \bar{b}'_{pj}\right) Q_{pj} d\bar{x}d\bar{y}$$

$$= Q_{pi} \tag{105}$$

where

$$Q_{pi} = \begin{cases} C_{pi}\left(z_{pi} - \hat{h}_i\right), & \hat{h}_i > z_{pi} \\ 0 & , \hat{h}_i \leq z_{pi} \end{cases}, \tag{106}$$

and subscript i indicates that the quantity is at node i.

36

Case 1. Head above z_{pi} through-out the time element
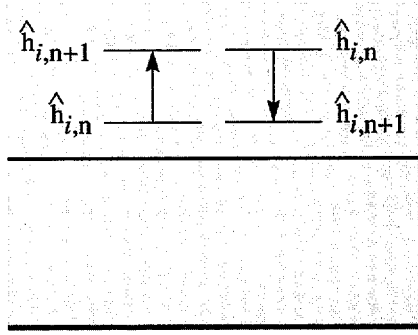
Case 3. Head rises above z_{pi} within the time element
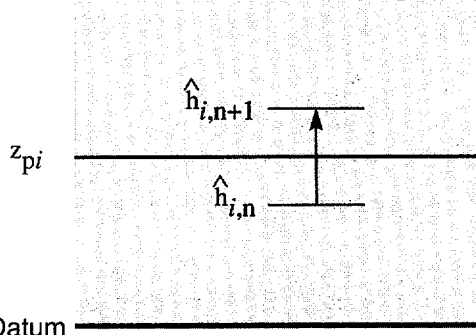
Case 2. Head drops below z_{pi} within the time element
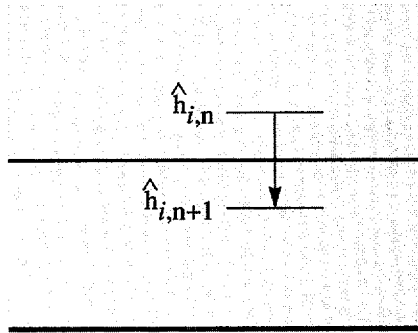
Case 4. Head below z_{pi} through-out the time element
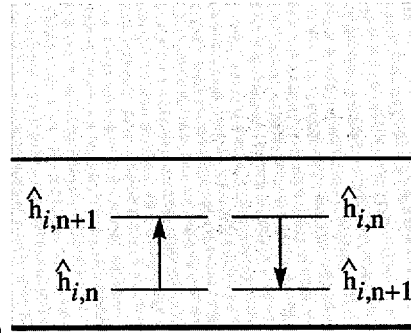
z_{pi} = controlling elevation at node i

$\hat{h}_{i,r}$ (r=n,n+1) = hydraulic head at node i and at time level r

Figure 10. Four possible cases involving change in head over time element n+1 during which there is point head-dependent discharge.

To time integrate the sink function, four cases involving the change in head $\hat{h}_i$ over time element n+1 must be distinguished (figure 10). If the sink function changes form within the time element, then the element is divided into two subintervals, $\phi_i \Delta t_{n+1}$ and $\left(1 - \phi_i\right)\Delta t_{n+1}$, defined by proportional change-over point ϕ_i. Because head is assumed to vary linearly within the time element, ϕ_i is defined by

$$\phi_i = \frac{z_{pi} - \hat{h}_{i,n}}{\hat{h}_{i,n+1} - \hat{h}_{i,n}}. \tag{107}$$

Formulations for each of the four cases in figure 10 are as follows:

1. Head above z_{pi} throughout the time element.

$$\int_0^{\Delta t_{n+1}} C_{pi}\Big(z_{pi} - \hat{h}_i\Big)\sigma_{n+1}dt' = \int_0^{\Delta t_{n+1}} C_{pi}\Big(z_{pi} - \hat{h}_{i,n}\sigma_n - \hat{h}_{i,n+1}\sigma_{n+1}\Big)\sigma_{n+1}dt'$$

$$= \tfrac{1}{6}\Delta t_{n+1}C_{pi}\Big[z_{pi} - \hat{h}_{i,n} + 2\Big(z_{pi} - \hat{h}_{i,n+1}\Big)\Big]. \tag{108}$$

2. Head drops below z_{pi} within the time element. This sink function must allow for linear variation of head from $\hat{h}_{i,n}$ to z_{pi} during the time interval t_n to $t_n + \phi_i\Delta t_{n+1}$, after which the sink function vanishes.

$$\int_0^{\Delta t_{n+1}} C_{pi}\Big(z_{pi} - \hat{h}_i\Big)\sigma_{n+1}dt' = \int_0^{\Delta t_{n+1}} C_{pi}\Big(z_{pi} - \hat{h}_{i,n}\sigma_n - \hat{h}_{i,n+1}\sigma_{n+1}\Big)\sigma_{n+1}dt'$$

$$= C_{pi}\int_0^{\phi_i\Delta t_{n+1}}\Big[z_{pi} - \hat{h}_{i,n}\sigma_n - \Big(\Big(z_{pi} - \hat{h}_{i,n}\Big)\tfrac{1}{\phi_i} + \hat{h}_{i,n}\Big)\sigma_{n+1}\Big]\sigma_{n+1}dt'$$

$$= C_{pi}\Big(z_{pi} - \hat{h}_{i,n}\Big)\int_0^{\phi_i\Delta t_{n+1}}\Big(1 - \frac{\sigma_{n+1}}{\phi_i}\Big)\sigma_{n+1}dt'$$

$$= \tfrac{1}{6}\phi_i^2\Delta t_{n+1}C_{pi}\Big(z_{pi} - \hat{h}_{i,n}\Big), \tag{109}$$

where equation (107) was used to eliminate $\hat{h}_{i,n+1}$.

3. Head rises above z_{pi} within the time element. This sink function must allow for linear variation of head from z_{pi} to $\hat{h}_{i,n+1}$ during the time interval $t_n + \phi_i\Delta_{n+1}$ to t_{n+1}, before which the sink function vanishes.

$$\int_0^{\Delta t_{n+1}} C_{pi}\Big(z_{pi} - \hat{h}_i\Big)\sigma_{n+1}dt' = \int_0^{\Delta t_{n+1}} C_{pi}\Big(z_{pi} - \hat{h}_{i,n}\sigma_n - \hat{h}_{i,n+1}\sigma_{n+1}\Big)\sigma_{n+1}dt'$$

$$= C_{pi}\int_{\phi_i\Delta t_{n+1}}^{\Delta t_{n+1}}\Big[z_{pi} - \Big(z_{pi} - \phi_i\hat{h}_{i,n+1}\Big)\frac{\sigma_n}{1 - \phi_i} - \hat{h}_{i,n+1}\sigma_{n+1}\Big]\sigma_{n+1}dt'$$

$$= C_{pi}\left(z_{pi} - \hat{h}_{i,n+1}\right)\int_{\phi_i \Delta t_{n+1}}^{\Delta t_{n+1}}\left(1 - \frac{\sigma_n}{1 - \phi_i}\right)\sigma_{n+1} dt'$$

$$= \frac{1}{3}\left(1 - \phi_i'\right)\Delta t_{n+1} C_{pi}\left(z_{pi} - \hat{h}_{i,n+1}\right), \tag{110}$$

where equation (107) was used to eliminate $\hat{h}_{i,n}$, and

$$\phi_i' = \frac{\phi_i\left(\phi_i + 1\right)}{2}, \tag{111}$$

4. Head below z_{pi} throughout the time element. The sink function vanishes during the entire time element.

To obtain the terms that add into equation (71), the results of cases 1 through 4 must be multiplied by $-2/\Delta t_{n+1}$ and converted to residual form using equation (57) to give

1. $-\frac{1}{3}C_{pi}\left[z_{pi} - \hat{h}_{i,n} + 2\left(z_{pi} - \hat{h}_{i,n+1}\right)\right]$

$$= C_{pi}\delta_i - C_{pi}\left(z_{pi} - \hat{h}_{i,n}\right). \tag{112}$$

2. $-\frac{1}{3}\phi_i^2 C_{pi}\left(z_{pi} - \hat{h}_{i,n}\right).$ \hfill (113)

3. $-\frac{2}{3}\left(1 - \phi_i'\right)C_{pi}\left(z_{pi} - \hat{h}_{i,n+1}\right)$

$$= \left(1 - \phi_i'\right)C_{pi}\delta_i - \frac{2}{3}\left(1 - \phi_i'\right)C_{pi}\left(z_{pi} - \hat{h}_{i,n}\right). \tag{114}$$

4. No formulation.

Addition of these terms into equation (71) consists of adding the coefficient of δ_i (that is, C_{pi} or $(1 - \phi_i')C_{pi}$) into matrix $\underline{\underline{V}}$ and adding the term containing $z_{pi} - \hat{h}_{i,n}$ onto the right-hand side. Because ϕ_i is unknown at the beginning of the time element, the predictor-corrector method is used to solve the modified equation (71).

The predictor step is initiated by checking whether $\hat{h}_{i,n} \geq z_{pi}$ or $\hat{h}_{i,n} < z_{pi}$. If the former is true, then case 1 is assumed and if the latter is true, then case 4 is assumed. Prediction equation (76) is then solved with the appropriate terms added in, and predicted heads $\hat{\underline{h}}^*$ are obtained using equation (79).

39

H = hydraulic head at top of confining unit or riverbed

z_t = elevation of top of aquifer

h = hydraulic head
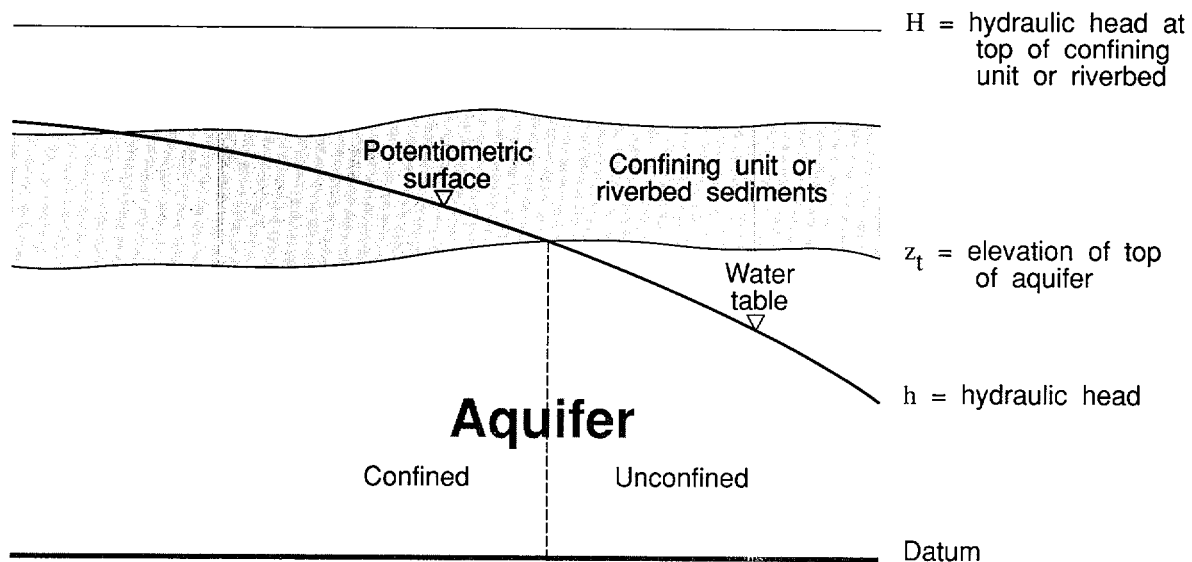
Figure 11. Cross section showing aquifer dewatering beneath a confining unit or riverbed sediments having low permeability.

To initiate the corrector step, heads $\hat{h}_i^*$ and $\hat{h}_{i,n}$ are checked to determine which of cases 1 through 4 apply. If case 2 or case 3 applies, then ϕ_i is estimated from

$$\phi_i = \frac{z_{pi} - \hat{h}_{i,n}}{\hat{h}_i^* - \hat{h}_{i,n}}. \tag{115}$$

Predicted head $\hat{h}_i^*$ was found to be a good prediction of $\hat{h}_{i,n+1}$ unless the time-element size was too large. The corrector equation is formed by adding the appropriate terms into equation (80), in which $\tilde{G}^*$ and $\underset{=}{G}^*$ may or may not be time variant depending on whether flow is unconfined or confined.

Areal head-dependent leakage combined with aquifer dewatering

Vertical leakage through a confining unit overlying an aquifer being dewatered, or leakage through the bed materials (assumed to have low permeability) of a river that is wide enough that it cannot be considered to be a line source or sink, may be simulated using a function in equation (1) similar to R(H - h) (figure 11) (Prickett and Lonnquist, 1971, p. 33-35). The difference is that in the present case the maximum rate of leakage to the aquifer is attained when the head in the aquifer declines below the base of the overlying confining unit or riverbed sediments. Any further decline in head results in a constant rate of leakage. There is no maximum rate of leakage from the aquifer when the head rises above the base of the confining unit or riverbed sediments. With this leakage function included, equation (1) may be written as

$$\frac{\partial}{\partial x}\left(T_{xx} \frac{\partial h}{\partial x} + T_{xy} \frac{\partial h}{\partial y}\right) + \frac{\partial}{\partial y}\left(T_{yx} \frac{\partial h}{\partial x} + T_{yy} \frac{\partial h}{\partial y}\right)$$
$$+ R(H - h) + S_a + W + P = S\frac{\partial h}{\partial t}, \tag{116}$$

where

$$S_a = \begin{cases} R_a\left(H_a - h\right), & h > z_t \\ R_a\left(H_a - z_t\right), & h \le z_t \end{cases}$$ (117)

In equation (117), R_a is the hydraulic conductance [time^{-1}] of the confining unit or riverbed sediments overlying the aquifer, H_a is the head at the distal side of the confining unit or riverbed sediments, and z_t is the elevation of the base of the overlying confining unit or riverbed sediments. The term $R(H - h)$ is retained in equation (116) to allow for a confining unit underlying the aquifer.

Spatial finite-element discretization applied to equation (117) results in an equation analogous to equation (27). Therefore, the leakage term resulting from the patch of elements for node i can be written

$$Q_{ai} = \begin{cases} C_{ai}\left(H_{ai} - \hat{h}_i\right), & \hat{h}_i > z_{ti} \\ C_{ai}\left(H_{ai} - z_{ti}\right), & \hat{h}_i \le z_{ti} \end{cases}$$ (118)

where

$$C_{ai} = \frac{1}{3}\sum_{e_i} R_a^e \Delta^e,$$ (119)

and Q_{ai} is the volumetric flow rate at node i [length^3/time] from leakage through the overlying confining unit or riverbed sediments.

Four cases similar to those developed for the point head-dependent sink functions (figure 10) are used to integrate equation (118) over time. The time element is divided in the same manner into two subintervals if the head in the aquifer crosses the base of the overlying confining unit or riverbed sediments within the element. For convenience, the same designation ϕ_i is used for the changeover point in time. The four cases can be expressed as follows:

1. Head above z_{ti} throughout the time element.

$$\int_0^{\Delta t_{n+1}} C_{ai}\left(H_{ai} - \hat{h}_i\right)\sigma_{n+1} dt'$$

$$= C_{ai}\int_0^{\Delta t_{n+1}} \left[\left(H_{ai,n} - \hat{h}_{i,n}\right)\sigma_n + \left(H_{ai,n+1} - \hat{h}_{i,n+1}\right)\sigma_{n+1}\right]\sigma_{n+1} dt'$$

$$= \frac{1}{6}\Delta t_{n+1} C_{ai}\left[H_{ai,n} - \hat{h}_{i,n} + 2\left(H_{ai,n+1} - \hat{h}_{i,n+1}\right)\right]$$ (120)

41

2. Head drops below z_{ti} within the time element. This function must allow for linear variation of head from $\hat{h}_{i,n}$ to z_{ti} during time interval t_n to $t_n + \phi_i \Delta t_{n+1}$, after which the head-dependent function is replaced by a known function.

$$\int_0^{\phi_i \Delta t_{n+1}} C_{ai}\left(H_{ai} - \hat{h}_i\right)\sigma_{n+1} dt' + \int_{\phi_i \Delta t_{n+1}}^{\Delta t_{n+1}} C_{ai}\left(H_{ai} - z_{ti}\right)\sigma_{n+1} dt'$$

$$= \int_0^{\Delta t_{n+1}} C_{ai} H_{ai}\sigma_{n+1} dt' - \int_0^{\phi_i \Delta t_{n+1}} C_{ai}\hat{h}_i\sigma_{n+1} dt' - \int_{\phi_i \Delta t_{n+1}}^{\Delta t_{n+1}} C_{ai} z_{ti}\sigma_{n+1} dt'$$

$$= C_{ai}\int_0^{\Delta t_{n+1}} \left(H_{ai,n}\sigma_n + H_{ai,n+1}\sigma_{n+1}\right)\sigma_{n+1} dt'$$

$$- C_{ai}\int_0^{\phi_i \Delta t_{n+1}} \left[\hat{h}_{i,n}\sigma_n + \left(\left(z_{ti} - \hat{h}_{i,n}\right)\frac{1}{\phi_i} + \hat{h}_{i,n}\right)\sigma_{n+1}\right]\sigma_{n+1} dt'$$

$$- C_{ai} z_{ti}\int_{\phi_i \Delta t_{n+1}}^{\Delta t_{n+1}} \sigma_{n+1} dt'$$

$$= \frac{1}{6}\phi_i^2 \Delta t_{n+1} C_{ai}\left[H_{ai,n} - \hat{h}_{i,n} + 2\left(H_{ai,n+1} - z_{ti}\right)\right]$$

$$+ \frac{1}{6}\left(1 - \phi_i^2\right)\Delta t_{n+1} C_{ai}\left[H_{ai,n} - z_{ti} + 2\left(H_{ai,n+1} - z_{ti}\right)\right], \qquad (121)$$

where equation (107) was used to eliminate $\hat{h}_{i,n+1}$.

3. Head rises above z_{ti} within the time element. This function must allow for linear variation of head from z_{ti} to $\hat{h}_{i,n+1}$ during the time interval $t_n + \phi_i \Delta t_{n+1}$ to t_{n+1}, before which the head dependent function is replaced by a known function.

$$\int_0^{\phi_i \Delta t_{n+1}} C_{ai}\left(H_{ai} - z_{ti}\right)\sigma_{n+1} dt' + \int_{\phi_i \Delta t_{n+1}}^{\Delta t_{n+1}} C_{ai}\left(H_{ai} - \hat{h}_i\right)\sigma_{n+1} dt'$$

$$= \int_0^{\Delta t_{n+1}} C_{ai} H_{ai}\sigma_{n+1} dt' - \int_0^{\phi_i \Delta t_{n+1}} C_{ai} z_{ti}\sigma_{n+1} dt' - \int_{\phi_i \Delta t_{n+1}}^{\Delta t_{n+1}} C_{ai} \hat{h}_i \sigma_{n+1} dt'$$

$$= C_{ai}\int_0^{\Delta t_{n+1}}\left(H_{ai,n}\sigma_n + H_{ai,n+1}\sigma_{n+1}\right)\sigma_{n+1} dt' - C_{ai} z_{ti}\int_0^{\phi_i \Delta t_{n+1}}\sigma_{n+1} dt'$$

$$- C_{ai}\int_{\phi_i \Delta t_{n+1}}^{\Delta t_{n+1}}\left[\left(z_{ti} - \phi_i \hat{h}_{i,n+1}\right)\frac{\sigma_n}{1-\phi_i} + \hat{h}_{i,n+1}\sigma_{n+1}\right]\sigma_{n+1} dt'$$

$$= \frac{1}{6}\phi_i' \Delta t_{n+1} C_{ai}\left[H_{ai,n} - z_{ti} + 2\left(H_{ai,n+1} - z_{ti}\right)\right]$$

$$+ \frac{1}{6}\left(1 - \phi_i'\right)\Delta t_{n+1} C_{ai}\left[H_{ai,n} - z_{ti} + 2\left(H_{ai,n+1} - \hat{h}_{i,n+1}\right)\right], \tag{122}$$

where equation (107) was used to eliminate $\hat{h}_{i,n}$, and ϕ_i' is defined by equation (111).

4. Head below z_{ti} throughout the time element.

$$\int_0^{\Delta t_{n+1}} C_{ai}\left(H_{ai} - z_{ti}\right)\sigma_{n+1} dt'$$

$$= C_{ai}\int_0^{\Delta t_{n+1}}\left[\left(H_{ai,n} - z_{ti}\right)\sigma_n + \left(H_{ai,n+1} - z_{ti}\right)\sigma_{n+1}\right]\sigma_{n+1} dt'$$

$$= \frac{1}{6}\Delta t_{n+1} C_{ai}\left[H_{ai,n} - z_{ti} + 2\left(H_{ai,n+1} - z_{ti}\right)\right]. \tag{123}$$

As for the point head-dependent sink functions, the terms that add into equation (71) are obtained by multiplying the above results by $-2/\Delta t_{n+1}$ and converting to residual form using equation (57). The results are:

1. $-\frac{1}{3}C_{ai}\left[H_{ai,n} - \hat{h}_{i,n} + 2\left(H_{ai,n+1} - \hat{h}_{i,n+1}\right)\right]$

$$= C_{ai}\delta_i - \frac{1}{3}C_{ai}\left[H_{ai,n} - \hat{h}_{i,n} + 2\left(H_{ai,n+1} - \hat{h}_{i,n}\right)\right]. \tag{124}$$

2. $-\frac{1}{3}\phi_i^2 C_{ai}\left[H_{ai,n} - \hat{h}_{i,n} + 2\left(H_{ai,n+1} - z_{ti}\right)\right]$

$$-\frac{1}{3}\left(1 - \phi_i^2\right)C_{ai}\left[H_{ai,n} - z_{ti} + 2\left(H_{ai,n+1} - z_{ti}\right)\right]. \tag{125}$$

3. $-\frac{1}{3}\phi_i' C_{ai}\left[H_{ai,n} - z_{ti} + 2\left(H_{ai,n+1} - z_{ti}\right)\right]$

$-\frac{1}{3}\left(1 - \phi_i'\right)C_{ai}\left[H_{ai,n} - z_{ti} + 2\left(H_{ai,n+1} - \hat{h}_{i,n+1}\right)\right]$

$= \left(1 - \phi_i'\right)C_{ai}\delta_i - \frac{1}{3}\phi_i' C_{ai}\left[H_{ai,n} - z_{ti} + 2\left(H_{ai,n+1} - z_{ti}\right)\right]$

$$-\frac{1}{3}\left(1 - \phi_i'\right)C_{ai}\left[H_{ai,n} - z_{ti} + 2\left(H_{ai,n+1} - \hat{h}_{i,n}\right)\right]. \tag{126}$$

4. $$-\frac{1}{3}C_{ai}\left[H_{ai,n} - z_{ti} + 2\left(H_{ai,n+1} - z_{ti}\right)\right]. \tag{127}$$

The terms in the above four cases are incorporated into equation (71) and the predictor-corrector method is employed in exactly the same manner as for the point head-dependent sink functions.

Areal head-dependent discharge (evapotranspiration)

Another areally distributed function allows for discharge-only processes such as evapotranspiration (figure 12). The rate of discharge from the aquifer is assumed to reach a maximum when the water table (or head in the aquifer) reaches the top of the aquifer, which is land surface. The minimum rate of zero is reached when the head declines to some lower threshold elevation (Prickett and Lonnquist, 1971, p. 37-38). This discharge function occupies the same position in equation (1) as S_a (see equation (116)) and is stated in the form

$$S_e = \begin{cases} R_e\left(z_e - z_t\right), & h \geq z_t \\ R_e\left(z_e - h\right), & z_e < h < z_t, \\ 0 & , & h \leq z_e \end{cases} \tag{128}$$

where R_e linearly relates the discharge rate to the head difference, z_e is the elevation below which the function vanishes, z_t is the elevation of the top of the aquifer, and $z_t > z_e$. An expression for R_e is (Prickett and Lonnquist, 1971, p. 37)

44

$$R_e = \frac{v_e}{d_e}, \tag{129}$$

where

$$d_e = z_t - z_e, \tag{130}$$

and v_e is the absolute value of the maximum unit discharge rate

[length/time] due to evapotranspiration from the aquifer.

Spatial integration of equation (128) is performed by using the same process used for equations (118) and (119). Therefore,

$$Q_{ei} = \begin{cases} C_{ei}\left(z_{ei} - z_{ti}\right), & \hat{h}_i \geq z_{ti} \\ C_{ei}\left(z_{ei} - \hat{h}_i\right), & z_{ei} < \hat{h}_i < z_{ti}, \\ 0 & , \hat{h}_i \leq z_{ei} \end{cases} \tag{131}$$

where

$$C_{ei} = \frac{1}{3} \sum_{e_i} R_e \Delta^e, \tag{132}$$

and Q_{ei} is volumetric discharge at node i [length^3/time].

Time integration yields nine separate cases (figure 13) involving the positions of $\hat{h}_{i,n}$ and $\hat{h}_{i,n+1}$ relative to z_{ei} and z_{ti}, and the time element is divided into either one, two, or three subintervals depending on the case applying. Changeover points in time are designated ϕ_{ti} and ϕ_{ei} to conform with the changeover elevations z_{ti} and z_{ei}, so that
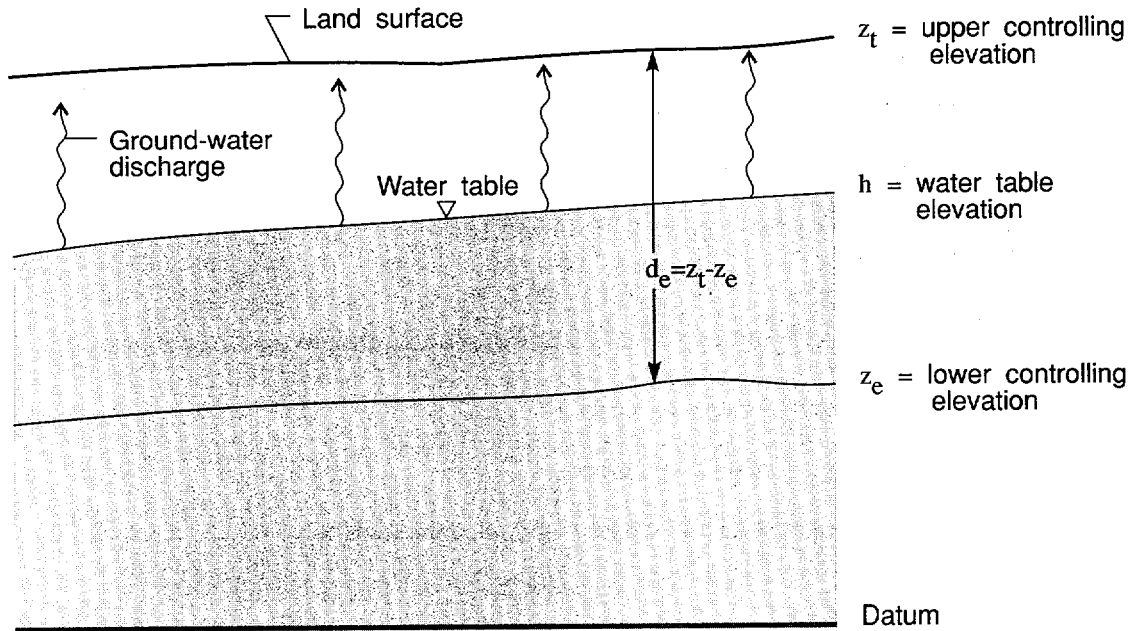
Figure 12. Cross section showing configuration of the water table and controlling elevations for evapotranspiration type of head-dependent discharge.

45

Case 1.

$\hat{h}_{i,n}$　　　$\hat{h}_{i,n+1}$

$\hat{h}_{i,n+1}$　　　$\hat{h}_{i,n}$

z_{ti}

z_{ei}

Datum

Case 4.

$\hat{h}_{i,n}$　　　$\hat{h}_{i,n+1}$

$\hat{h}_{i,n+1}$　　　$\hat{h}_{i,n}$

z_{ti}

z_{ei}

Datum

Case 7.

z_{ti}

z_{ei}

$\hat{h}_{i,n}$　　　$\hat{h}_{i,n+1}$

$\hat{h}_{i,n+1}$　　　$\hat{h}_{i,n}$

Datum

Case 2.

$\hat{h}_{i,n}$

z_{ti}

$\hat{h}_{i,n+1}$

z_{ei}

Datum

Case 5.

z_{ti}

$\hat{h}_{i,n}$

z_{ei}

$\hat{h}_{i,n+1}$

Datum

Case 8.

$\hat{h}_{i,n}$

z_{ti}

z_{ei}

$\hat{h}_{i,n+1}$

Datum

Case 3.

$\hat{h}_{i,n+1}$

z_{ti}

$\hat{h}_{i,n}$

z_{ei}

Datum

Case 6.

z_{ti}

$\hat{h}_{i,n+1}$

z_{ei}

$\hat{h}_{i,n}$

Datum

Case 9.

$\hat{h}_{i,n+1}$

z_{ti}

z_{ei}

$\hat{h}_{i,n}$

Datum
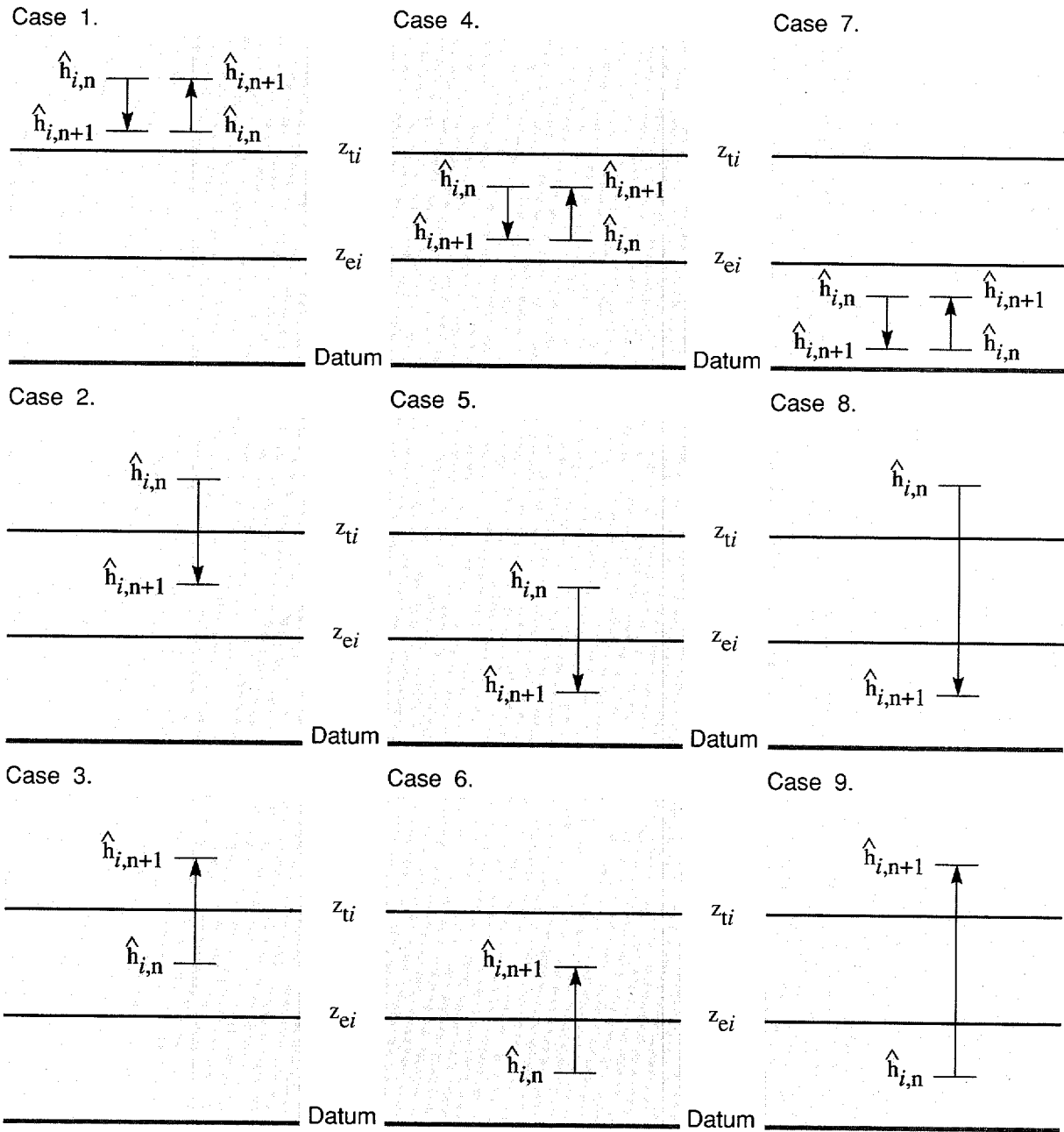
z_{ti} = upper controlling elevation at node i

z_{ei} = lower controlling elevation at node i

$\hat{h}_{i,r}$ (r=n,n+1) = hydraulic head at node i and at time level r

Figure 13. Nine possible cases involving change in head over time-element n+1 during which there is areal head-dependent discharge.

$$\phi_{ti} = \frac{z_{ti} - \hat{h}_{i,n}}{\hat{h}_{i,n+1} - \hat{h}_{i,n}}, \tag{133}$$

and

$$\phi_{ei} = \frac{z_{ei} - \hat{h}_{i,n}}{\hat{h}_{i,n+1} - \hat{h}_{i,n}}. \tag{134}$$

The cases are as follows.

1. Head above z_{ti} throughout the time element.

$$\int_0^{\Delta t_{n+1}} C_{ei}\left(z_{ei} - z_{ti}\right)\sigma_{n+1} dt' = \frac{1}{2}\Delta t_{n+1} C_{ei}\left(z_{ei} - z_{ti}\right). \tag{135}$$

2. Head drops below z_{ti} but stays above z_{ei} within the time element.

$$\int_0^{\phi_{ti}\Delta t_{n+1}} C_{ei}\left(z_{ei} - z_{ti}\right)\sigma_{n+1} dt' + \int_{\phi_{ti}\Delta t_{n+1}}^{\Delta t_{n+1}} C_{ei}\left(z_{ei} - \hat{h}_i\right)\sigma_{n+1} dt'$$

$$= C_{ei}\left(z_{ei} - z_{ti}\right)\int_0^{\phi_{ti}\Delta t_{n+1}} \sigma_{n+1} dt'$$

$$+ C_{ei}\int_{\phi_{ti}\Delta t_{n+1}}^{\Delta t_{n+1}} \left[z_{ei} - \left(z_{ti} - \phi_{ti}\hat{h}_{i,n+1}\right)\frac{\sigma_n}{1 - \phi_{ti}} - \hat{h}_{i,n+1}\sigma_{n+1}\right]\sigma_{n+1} dt'$$

$$= \frac{1}{2}\phi'_{ti}\Delta t_{n+1}C_{ei}\left(z_{ei} - z_{ti}\right) + \frac{1}{6}\left(1 - \phi'_{ti}\right)\Delta t_{n+1}C_{ei}\left[z_{ei} - z_{ti} + 2\left(z_{ei} - \hat{h}_{i,n+1}\right)\right], \tag{136}$$

where ϕ'_{ti} is given by equation (111) with ϕ_{ti} replacing ϕ_i.

3. Head rises from between z_{ei} and z_{ti} to above z_{ti} within the time element.

$$\int_0^{\phi_{ti}\Delta t_{n+1}} C_{ei}\left(z_{ei} - \hat{h}_i\right)\sigma_{n+1} dt' + \int_{\phi_{ti}\Delta t_{n+1}}^{\Delta t_{n+1}} C_{ei}\left(z_{ei} - z_{ti}\right)\sigma_{n+1} dt'$$

$$= C_{ei} \int_0^{\phi_{ti}\Delta t_{n+1}} \left[z_{ei} - \hat{h}_{i,n}\sigma_n - \left(\left(z_{ti} - \hat{h}_{i,n} \right) \frac{1}{\phi_{ti}} + \hat{h}_{i,n} \right) \sigma_{n+1} \right] \sigma_{n+1} dt'$$

$$+ C_{ei} \left(z_{ei} - z_{ti} \right) \int_{\phi_{ti}\Delta t_{n+1}}^{\Delta t_{n+1}} \sigma_{n+1} dt'$$

$$= \frac{1}{6}\phi_{ti}^2 \Delta t_{n+1} C_{ei} \left[z_{ei} - \hat{h}_{i,n} + 2\left(z_{ei} - z_{ti} \right) \right] + \frac{1}{2}\left(1 - \phi_{ti}^2 \right) \Delta t_{n+1} C_{ei} \left(z_{ei} - z_{ti} \right). \quad (137)$$

4. Head stays between z_{ei} and z_{ti} throughout the time element.

$$\int_0^{\Delta t_{n+1}} C_{ei}\left(z_{ei} - \hat{h}_i \right)\sigma_{n+1} dt' = C_{ei} \int_0^{\Delta t_{n+1}} \left[\left(z_{ei} - \hat{h}_{i,n} \right)\sigma_n + \left(z_{ei} - \hat{h}_{i,n+1} \right)\sigma_{n+1} \right]\sigma_{n+1} dt'$$

$$= \frac{1}{6}\Delta t_{n+1} C_{ei}\left[z_{ei} - \hat{h}_{i,n} + 2\left(z_{ei} - \hat{h}_{i,n+1} \right) \right]. \quad (138)$$

5. Head drops from between z_{ei} and z_{ti} to below z_{ei} within the time element.

$$\int_0^{\phi_{ei}\Delta t_{n+1}} C_{ei}\left(z_{ei} - \hat{h}_i \right)\sigma_{n+1} dt'$$

$$= C_{ei} \int_0^{\phi_{ei}\Delta t_{n+1}} \left[z_{ei} - \hat{h}_{i,n}\sigma_n - \left(\left(z_{ei} - \hat{h}_{i,n} \right)\frac{1}{\phi_{ei}} + \hat{h}_{i,n} \right)\sigma_{n+1} \right]\sigma_{n+1} dt'$$

$$= \frac{1}{6}\phi_{ei}^2 \Delta t_{n+1} C_{ei}\left(z_{ei} - \hat{h}_{i,n} \right). \quad (139)$$

6. Head rises above z_{ei} but stays below z_t within the time element.

$$\int_{\phi_{ei}\Delta t_{n+1}}^{\Delta t_{n+1}} C_{ei}\left(z_{ei} - \hat{h}_i\right)\sigma_{n+1}dt' = C_{ei}\int_{\phi_{ei}\Delta t_{n+1}}^{\Delta t_{n+1}}\left[z_{ei} - \left(z_{ei} - \phi_{ei}\hat{h}_{i,n+1}\right)\frac{\sigma_n}{1-\phi_{ei}}\right.$$

$$\left. - \hat{h}_{i,n+1}\sigma_{n+1}\right]\sigma_{n+1}dt' = \frac{1}{3}\left(1 - \phi'_{ei}\right)\Delta t_{n+1}C_{ei}\left(z_{ei} - \hat{h}_{i,n+1}\right), \tag{140}$$

where ϕ'_{ei} is given by equation (111) with ϕ_{ei} replacing ϕ_i.

7. Head stays below z_{ei} throughout the time element. The discharge function vanishes within the entire time element.

8. Head drops from above z_{ti} to below z_{ei} within the time element. This function must allow for constant discharge during timespan t_n to $t_n + \phi_{ti}\Delta t_{n+1}$ and head-dependent discharge during timespan $t_n + \phi_{ti}\Delta t_{n+1}$ to $t_n + \phi_{ei}\Delta t_{n+1}$, and must vanish during timespan $t_n + \phi_{ei}\Delta t_{n+1}$ to t_{n+1}. The head-dependent discharge can be expressed as a function of ϕ_{ei}, ϕ_{ti}, z_{ei}, and z_{ti} by eliminating $\hat{h}_{i,n}$ and $\hat{h}_{i,n+1}$ using equations (133) and (134). First $\hat{h}_{i,n+1}$ is eliminated by solving equation (134) for $\hat{h}_{i,n+1}$, then $\hat{h}_{i,n}$ is eliminated by combining equations (133) and (134) to get

$$\hat{h}_{i,n} = \frac{\phi_{ei}z_{ti} - \phi_{ti}z_{ei}}{\phi_{ei} - \phi_{ti}} \tag{141}$$

Thus,

$$\int_0^{\phi_{ti}\Delta t_{n+1}} C_{ei}\left(z_{ei} - z_{ti}\right)\sigma_{n+1}dt' + \int_{\phi_{ti}\Delta t_{n+1}}^{\phi_{ei}\Delta t_{n+1}} C_{ei}\left(z_{ei} - \hat{h}_i\right)\sigma_{n+1}dt'$$

$$= C_{ei}\left(z_{ei} - z_{ti}\right)\int_0^{\phi_{ti}\Delta t_{n+1}}\sigma_{n+1}dt' + C_{ei}\int_{\phi_{ti}\Delta t_{n+1}}^{\phi_{ei}\Delta t_{n+1}}\left[z_{ei} - \hat{h}_{i,n}\sigma_n - \left(\left(z_{ei} - \hat{h}_{i,n}\right)\cdot\right.\right.$$

$$\left.\left.\cdot\frac{1}{\phi_{ei}} + \hat{h}_{i,n}\right)\sigma_{n+1}\right]\sigma_{n+1}dt'$$

$$= C_{ei}\left(z_{ei} - z_{ti}\right)\int_0^{\phi_{ti}\Delta t_{n+1}} \sigma_{n+1}dt' + C_{ei}\left(z_{ei} - \hat{h}_{i,n}\right)\int_{\phi_{ti}\Delta t_{n+1}}^{\phi_{ei}\Delta t_{n+1}}\left(1 - \frac{\sigma_{n+1}}{\phi_{ei}}\right)\sigma_{n+1}dt'$$

$$= C_{ei}\left(z_{ei} - z_{ti}\right)\int_0^{\phi_{ti}\Delta t_{n+1}} \sigma_{n+1}dt' + C_{ei}\left(z_{ei} - z_{ti}\right)\int_{\phi_{ti}\Delta t_{n+1}}^{\phi_{ei}\Delta t_{n+1}}\left[\frac{\phi_{ei} - \sigma_{n+1}}{\phi_{ei} - \phi_{ti}}\right]\sigma_{n+1}dt'$$

$$= \frac{1}{6}\left[\phi_{ti}\left(\phi_{ei} + \phi_{ti}\right) + \phi_{ei}^2\right]\Delta t_{n+1}C_{ei}\left(z_{ei} - z_{ti}\right). \tag{142}$$

9. Head rises from below z_{ei} to above z_{ti} within the time element. This case is analogous to case 8, except that the head changes in the opposite direction. Hence, the resulting equations are

$$\int_{\phi_{ei}\Delta t_{n+1}}^{\phi_{ti}\Delta t_{n+1}} C_{ei}\left(z_{ei} - \hat{h}_i\right)\sigma_{n+1}dt' + \int_{\phi_{ti}\Delta t_{n+1}}^{\Delta t_{n+1}} C_{ei}\left(z_{ei} - z_{ti}\right)\sigma_{n+1}dt'$$

$$= C_{ei}\left(z_{ei} - z_{ti}\right)\int_{\phi_{ei}\Delta t_{n+1}}^{\phi_{ti}\Delta t_{n+1}}\left[\frac{\phi_{ei} - \sigma_{n+1}}{\phi_{ei} - \phi_{ti}}\right]\sigma_{n+1}dt' + C_{ei}\left(z_{ei} - z_{ti}\right)\int_{\phi_{ti}\Delta t_{n+1}}^{\Delta t_{n+1}} \sigma_{n+1}dt'$$

$$= \frac{1}{6}\left[\left(\phi_{ei} + 2\phi_{ti}\right)\left(\phi_{ti} - \phi_{ei}\right) + 3\left(1 + \phi_{ti}\right)\left(1 - \phi_{ti}\right)\right]\Delta t_{n+1}C_{ei}\left(z_{ei} - z_{ti}\right). \tag{143}$$

Multiplication of the above discharge functions by $-2/\Delta t_{n+1}$ and conversion to residual form using equation (57) yields the terms that add into equation (71). The results are:

1. $$- C_{ei}\left(z_{ei} - z_{ti}\right). \tag{144}$$

2. $$- \phi'_{ti}C_{ei}\left(z_{ei} - z_{ti}\right) - \frac{1}{3}\left(1 - \phi'_{ti}\right)C_{ei}\left[z_{ei} - z_{ti} + 2\left(z_{ei} - \hat{h}_{i,n+1}\right)\right]$$

$$= \left(1 - \phi'_{ti}\right)C_{ei}\delta_i - \phi'_{ti}C_{ei}\left(z_{ei} - z_{ti}\right) - \frac{1}{3}\left(1 - \phi'_{ti}\right)C_{ei}\left[z_{ei} - z_{ti}\right.$$

$$\left. + 2\left(z_{ei} - \hat{h}_{i,n}\right)\right]. \tag{145}$$

3. $\quad - \frac{1}{3}\phi_{ti}^2 C_{ei}\left[z_{ei} - \hat{h}_{i,n} + 2\left(z_{ei} - z_{ti}\right)\right] - \left(1 - \phi_{ti}^2\right)C_{ei}\left(z_{ei} - z_{ti}\right).$ (146)

4. $\quad - \frac{1}{3}C_{ei}\left[z_{ei} - \hat{h}_{i,n} + 2\left(z_{ei} - \hat{h}_{i,n+1}\right)\right]$

$$= C_{ei}\delta_i - C_{ei}\left(z_{ei} - \hat{h}_{i,n}\right). \tag{147}$$

5. $\quad - \frac{1}{3}\phi_{ei}^2 C_{ei}\left(z_{ei} - \hat{h}_{i,n}\right).$ (148)

6. $\quad - \frac{2}{3}\left(1 - \phi'_{ei}\right)C_{ei}\left(z_{ei} - \hat{h}_{i,n+1}\right)$

$$= \left(1 - \phi'_{ei}\right)C_{ei}\delta_i - \frac{2}{3}\left(1 - \phi'_{ei}\right)C_{ei}\left(z_{ei} - \hat{h}_{i,n}\right). \tag{149}$$

7. No formulation.

8. $\quad - \frac{1}{3}\left[\phi_{ti}\left(\phi_{ei} + \phi_{ti}\right) + \phi_{ei}^2\right]C_{ei}\left(z_{ei} - z_{ti}\right).$ (150)

9. $\quad - \frac{1}{3}\left[\left(\phi_{ei} + 2\phi_{ti}\right)\left(\phi_{ti} - \phi_{ei}\right) + 3\left(1 + \phi_{ti}\right)\left(1 - \phi_{ti}\right)\right]C_{ei}\left(z_{ei} - z_{ti}\right).$ (151)

Procedures for use of the above nine cases in equation (71) and solution using the predictor-corrector method are analogous to those used for the previous two types of head-dependent functions. For the predictor, if $\hat{h}_{i,n} \geq z_{ti}$, then equation (144) is used in equation (76) (case 1); if $z_{ei} < \hat{h}_{i,n} < z_{ti}$, then equation (147) is used (case 4); and if $\hat{h}_{i,n} \leq z_{ti}$, then no terms are used (case 7). Estimates of ϕ_{ei} and ϕ_{ti} to use in equations (144) through (151) are obtained using equations (133) and (134), with $\hat{h}_i^*$ substituted for $\hat{h}_{i,n+1}$. The corrector is employed by adding one of equations (144) through (151) into corrector equation (80), as appropriate based on checking $\hat{h}_i^*$ and $\hat{h}_{i,n}$ against z_{ei} and z_{ti}, and solving for δ.

Line head-dependent leakage combined with aquifer dewatering

The final type of head-dependent function is a line source or sink of the general form of the boundary condition given by equation (4), except in the present case the function yields a maximum flux when the head in the
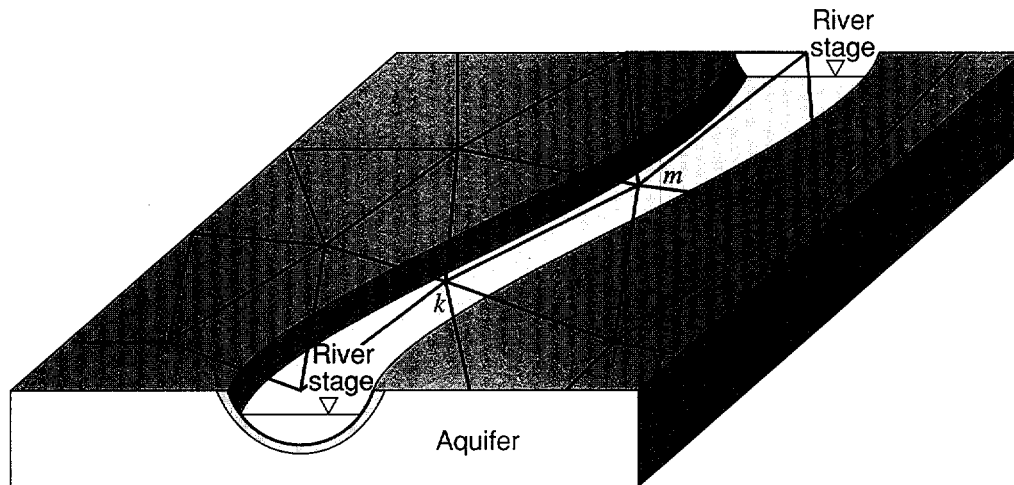
Figure 14a. Block diagram of a river idealized as a line source or sink along spatial element sides.

aquifer declines to a specified elevation. This function is most often used to simulate a river that is narrow enough to be replaced by a line (figure 14a). As for the case of a wide river, the maximum leakage rate from the river to the aquifer is attained when the head in the aquifer declines to the bottom of the riverbed sediments (assumed to have low permeability) (figure 14b).

The line source or sink function is written in the form of a flow across an internal or external boundary, or

$$
q_n = \begin{cases} \alpha_r \left(H_r - h \right) , & h > z_r \\ \alpha_r \left(H_r - z_r \right) , & h \leq z_r \end{cases} ,
\tag{152}
$$

where α_r [length/time] is a parameter that is a function of the hydraulic conductivity of sediments through which leakage occurs, H_r is the controlling head (for rivers, the river-stage elevation), and z_r is the elevation at which the discharge to the aquifer is a maximum (for rivers, the elevation of the bottom of the riverbed sediments). For riverbed sediments, α_r is given by

$$
\alpha_r = \frac{K_r W_r}{b_r} ,
\tag{153}
$$

where K_r is the hydraulic conductivity of the riverbed sediments, W_r is the width of the river, and b_r is the thickness of the riverbed sediments.

Equation (152) is incorporated into the spatial finite-element equations in the same manner as equation (4) (see equation (32)). That is, the total discharge across the line source or sink in the patch of elements for node i is

$$
Q_{ri} = \begin{cases} C_{ri} \left(H_{ri} - \hat{h}_i \right) , & \hat{h}_i > z_{ri} \\ C_{ri} \left(H_{ri} - z_{ri} \right) , & \hat{h}_i \leq z_{ri} \end{cases} ,
\tag{154}
$$

52
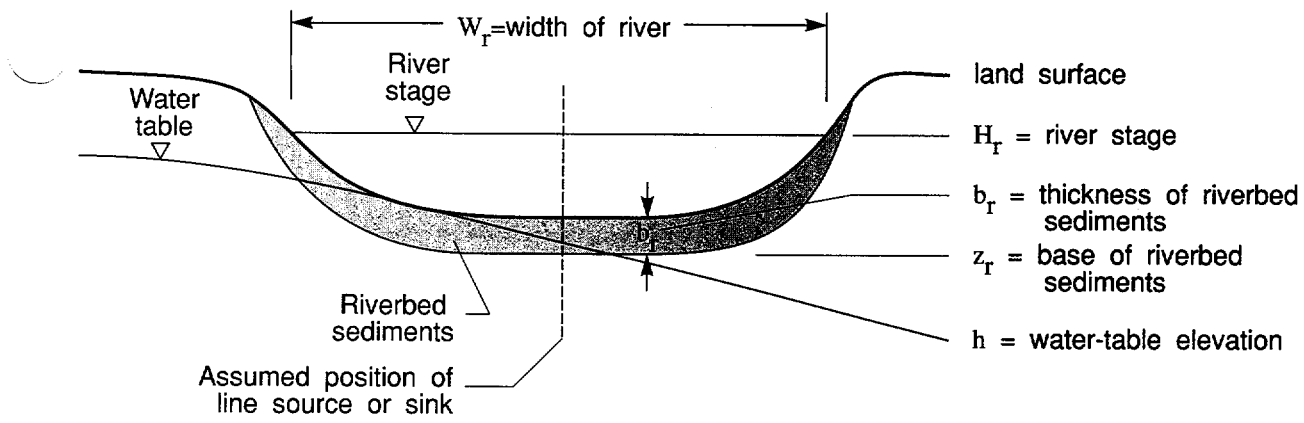
Figure 14b. Cross section showing a configuration of the water-table elevation under a river that is idealized as a line or source sink.

where

$$C_{ri} = \frac{1}{2}\sum_j \left(\alpha_r L\right)_{ij}, ,$$ (155)

Q_{ri} is the volumetric discharge at node i [length^3/time] from leakage involving the line sources or sinks, and L_{ij}, is defined the same as for equation (32) (figure 14c).

H_{ri} (i=k,m) = river-stage elevation at node i

z_{ri} (i=k,m) = base of riverbed sediments at node i

W_r = width of river in element e

L_{km} = length of side k-m of element e
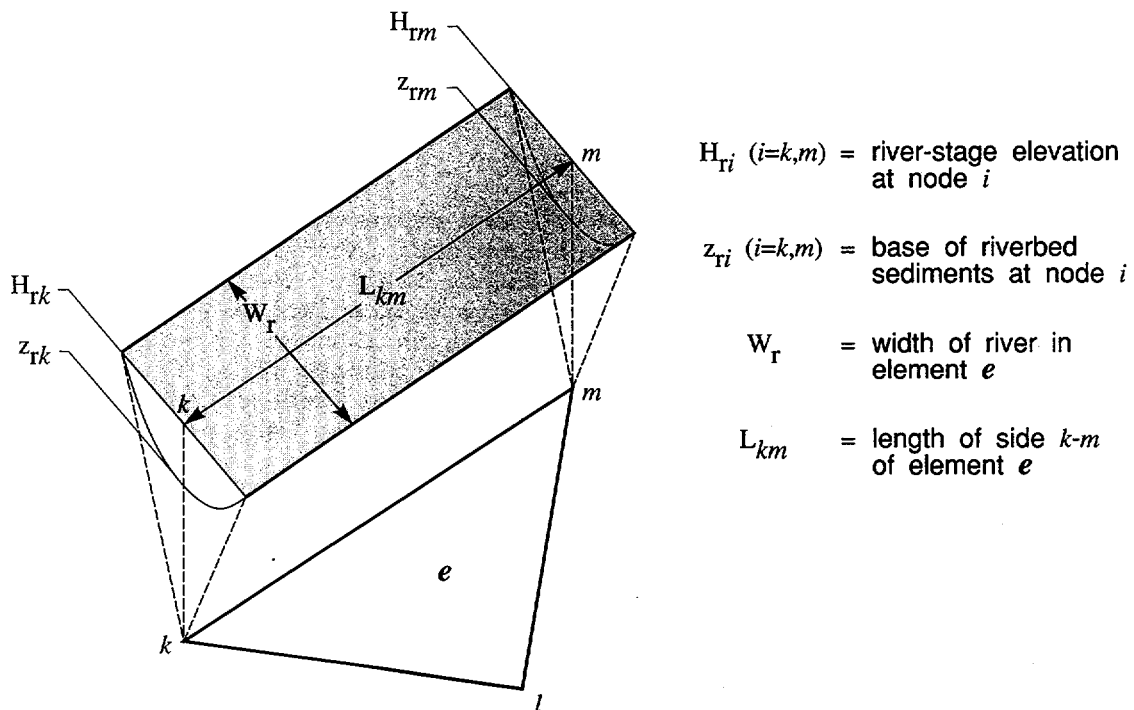
Figure 14c. Nomenclature for side k-m of element e that forms a line head-dependent source or sink.

53

Equation (154) is exactly the same as equation (118) expressing areal head-dependent leakage, except that H_{ri} replaces H_{ai}, C_{ri} replaces C_{ai}, and z_{ri} replaces z_{ti}. Therefore, the expressions and the predictor-corrector solution procedure derived for areal, head-dependent leakage apply to the present case as well.

Leakage of water stored elastically in a confining unit (transient leakage)

Equation (1) can be rewritten in a general form that includes leakage to or from a confining unit as follows (Cooley, 1974, p. 3-9):

$$\frac{\partial}{\partial x}\left[T_{xx}\frac{\partial h}{\partial x} + T_{xy}\frac{\partial h}{\partial y}\right] + \frac{\partial}{\partial y}\left[T_{yx}\frac{\partial h}{\partial x} + T_{yy}\frac{\partial h}{\partial y}\right]$$

$$\pm\ K'_{zz}\frac{\partial h'}{\partial z}\bigg|_{z=z_c} + W + P = S\frac{\partial h}{\partial t}\ , \tag{156}$$

where K'_{zz} is the vertical hydraulic conductivity of the confining unit, h' is the head in the confining unit, and z_c is the elevation of the base of the confining unit (if the confining unit overlies the aquifer) or the elevation of the top of the confining unit (if the confining unit underlies the aquifer).

If the confining unit has no elastic storage capacity and flow in the confining unit is almost vertical, then the leakage rate is

$$\pm\ K'_{zz}\frac{\partial h'}{\partial z}\bigg|_{z=z_c} = R(H-h)\ , \tag{157}$$

and equation (1) results. However, if the confining unit has elastic storage capacity, then the leakage rate must be computed using an unsteady-state equation for flow in the confining unit. The formulation of this problem was developed by Hantush (1960), and expanded by Neuman and Witherspoon (1969) and Herrera and his coworkers (Herrera, 1970, 1974; Herrera and Rodarte, 1973; Herrera and Yates, 1977). The following approach is an expansion of their approaches to apply to the finite-element method. For notational simplicity, the confining unit is assumed to overlie the aquifer, because the final function describing leakage to or from the aquifer is the same whether the confining unit overlies or underlies the aquifer. In this case, $z_c = z_t$, the elevation of the base of the confining unit.

Flow in the confining unit is assumed to be almost vertical, which is a reasonable approximation when $K'_{zz}/K_{xx} < 10^{-2}$ (Neuman and Witherspoon, 1969, p. 804). With this assumption, flow in the confining unit at some location (x,y) can be described by the following initial value problem:

$$K'_{zz}\frac{\partial^2 h'}{\partial z^2} = S'_s\frac{\partial h'}{\partial t} \tag{158}$$

54

subject to

$$h' = h(t), \quad z = z_t, \quad t \geq 0,$$

$$h' = H(t), \quad z = z_t + b', \quad t \geq 0, \tag{159}$$

$$h' = H_0'(z), \quad z_t \leq z \leq z_t + b', \quad t = 0,$$

where S_s' is specific storage of the confining unit [length^{-1}], b' is the

thickness of the confining unit, h is head in the aquifer, H is head at the distal side of the confining unit (assumed to be a known function of time), $H_0'(z)$ is the initial steady-state distribution of head in the confining

unit, $H_0'\left(z_t\right) = h(0)$, and $H_0'\left(z_t + b'\right) = H(0)$.

To initiate development of the leakage function to be inserted into the finite-element equations, spatial finite-element discretization is applied to the general leakage rate in the same manner as it was to R(H-h) to yield

$$\sum_e \int\!\!\int_{\Delta^e} \bar{N}_i^e K'_{zz} \frac{\partial h'}{\partial z}\bigg|_{z=z_t} d\bar{x}d\bar{y} \simeq \frac{1}{3}\sum_e K'_{zz}{}^e \Delta^e \frac{\partial h_i'}{\partial z}\bigg|_{z=z_t}, \tag{160}$$

where, to conform with R, K'_{zz} is assumed to be constant in element e.

Equation (160) expresses the total leakage rate across the patch of elements for node i in terms of the hydraulic gradient at node i. Thus, the equation for flow in the confining unit at node i must yield equation (160) when used to obtain the leakage rate. This flow equation is derived as follows. Integration of equation (158) from z_t to $z \leq z_t + b'$ and solution for the

leakage rate across the base of the confining unit yields

$$K'_{zz} \frac{\partial h'}{\partial z}\bigg|_{z=z_t} = K'_{zz} \frac{\partial h'}{\partial z} - S_s' \int_{z_t}^{z} \frac{\partial h'}{\partial t} dz, \tag{161}$$

which, when substituted into the integral in equation (160), yields an expression for the approximate leakage across the patch of elements for node i in the form of

$$\frac{1}{3}\sum_e K'_{zz}{}^e \Delta^e \frac{\partial h_i'}{\partial z}\bigg|_{z=z_t} = \frac{1}{3}\sum_e K'_{zz}{}^e \Delta^e \frac{\partial h'}{\partial z} - \frac{1}{3}\sum_e S_s'{}^e \Delta^e \int_{z_t}^{z} \frac{\partial h'}{\partial t} dz, \tag{162}$$

where $S_s'{}^e$ is the specific storage of the confining unit in element e.

Differentiation of equation (162) with respect to z yields the equation for flow in the confining unit at node i as

$$\sum_e K'_{zz}{}^e \Delta^e \frac{\partial^2 h_i'}{\partial z^2} = \sum_e S_s'{}^e \Delta^e \frac{\partial h_i'}{\partial t}. \tag{163}$$

The leakage rate given by equation (160) is calculated using the head in the confining unit, h_i', which is obtained by analytically solving

equation (163). Appropriate boundary and initial conditions are given by $_\wedge$ equations (159) written for node i using approximate head in the aquifer h_i in place of exact head h_i. That is,

55

$$h_i' = \hat{h}_i(t), \quad z = z_{ti}, \quad t \geq 0,$$

$$h_i' = H_i(t), \quad z = z_{ti} + b_i', \quad t \geq 0, \tag{164}$$

$$h_i' = H_{0i}'(z), \quad z_{ti} \leq z \leq z_{ti} + b_i', \quad t = 0,$$

where b_i' is a weighted average thickness of the confining unit for all

elements in the patch for node i that is derived further on. The solution to this initial value problem was obtained by Carslaw and Jaeger (1959, p. 102-104) and can be written in the form

$$h_i'(z,t) = 2 \sum_{m=1}^{\infty} \frac{1}{m\pi} \sin \frac{m\pi \left(z - z_t\right)}{b_i'} \int_0^t \frac{d\eta_{mi}(\tau)}{d\tau} \, e^{-(m\pi)^2 \gamma_i (t-\tau)} d\tau$$

$$+ \frac{H_i(t) - \hat{h}_i(t)}{b_i'} \left(z - z_{ti}\right) + \hat{h}_i(t), \tag{165}$$

where

$$\eta_{mi}(t) = (-1)^m H_i(t) - \hat{h}_i(t) \tag{166}$$

and

$$\gamma_i = \frac{\sum_{e_i} K_{zz}'^e \Delta^e}{b_i'^2 \sum_{e_i} S_s'^e \Delta^e}. \tag{167}$$

The leakage rate at time t, $K_{zz}' \partial h_i'/\partial z \big|_{z=z_t}$, is computed from equations

(165) and (166). When this is substituted into equation (160), the leakage rate across the patch of elements for node i is obtained as

$$\frac{1}{3} \sum_{e_i} K_{zz}'^e \Delta^e \frac{\partial h_i'}{\partial z}\bigg|_{z=z_t} = \frac{2}{3b_i'} \sum_{e_i} K_{zz}'^e \Delta^e \sum_{m=1}^{\infty} (-1)^m \int_0^t \frac{dH_i(\tau)}{d\tau} e^{-(m\pi)^2 \gamma_i (t-\tau)} d\tau$$

$$- \frac{2}{3b_i'} \sum_{e_i} K_{zz}'^e \Delta^e \sum_{m=1}^{\infty} \int_0^t \frac{d\hat{h}_i(\tau)}{d\tau} e^{-(m\pi)^2 \gamma_i (t-\tau)} d\tau$$

$$+ \frac{1}{3b_i'} \sum_{e_i} K_{zz}'^e \Delta^e \left[H_i(t) - \hat{h}_i(t)\right]. \tag{168}$$

56

As $S'^e_s \to 0$, equation (168) should approach the leakage term based on equation (27). That is,

$$\frac{1}{3b'_i} \sum_{e_i} K'^e_{zz} \Delta^e \left(H_i - \hat{h}_i \right) = \frac{1}{3} \sum_{e_i} R^e \Delta^e \left(H_i - \hat{h}_i \right).$$
(169)

Therefore, the weighted average thickness b'_i should be defined by

$$b'_i = \frac{\sum\limits_{e_i} K'^e_{zz} \Delta^e}{\sum\limits_{e_i} R^e \Delta^e}$$
(170)

to make equations (168) and (27) consistent. Because R^e is K'^e_{zz} divided by the confining-unit thickness for element e, b'_i is a weighted harmonic mean of confining-unit thicknesses.

The leakage functions to be inserted into the finite-element equations are developed from equation (168). To yield a useful form for the leakage functions, the integrals must be evaluated and the infinite series must be approximated. The integrals can be approximated by using the procedure of Cooley (1972), in which the integrals for time level $t = t_{n+1}$ are evaluated in terms of integrals for time level $t = t_n$, which have already been evaluated. Thus, the necessity for storing the heads for all previous time levels is avoided. The procedure is applied to the second integral in equation (168), for example, as follows:

$$\int_0^{t_{n+1}} \frac{d\hat{h}_i(\tau)}{d\tau} e^{-(m\pi)^2 \gamma_i \left(t_{n+1} - \tau \right)} d\tau$$

$$= \int_0^{t_n} \frac{d\hat{h}_i(\tau)}{d\tau} e^{-(m\pi)^2 \gamma_i \left(t_n + \Delta t_{n+1} - \tau \right)} d\tau + \int_{t_n}^{t_{n+1}} \frac{d\hat{h}_i(\tau)}{d\tau} e^{-(m\pi)^2 \gamma_i \left(t_{n+1} - \tau \right)} d\tau$$

$$= e^{-(m\pi)^2 \gamma_i \Delta t_{n+1}} \int_0^{t_n} \frac{d\hat{h}_i(\tau)}{d\tau} e^{-(m\pi)^2 \gamma_i \left(t_n - \tau \right)} d\tau$$

$$+ \frac{\hat{h}_{i,n+1} - \hat{h}_{i,n}}{\Delta t_{n+1}} \int_{t_n}^{t_{n+1}} e^{-(m\pi)^2 \gamma_i \left(t_{n+1} - \tau\right)} d\tau$$

$$= e^{-(m\pi)^2 \gamma_i \Delta t_{n+1}} \int_0^{t_n} \frac{d\hat{h}_i(\tau)}{d\tau} e^{-(m\pi)^2 \gamma_i \left(t_n - \tau\right)} d\tau$$

$$+ \frac{\hat{h}_{i,n+1} - \hat{h}_{i,n}}{\Delta t_{n+1}} \frac{1}{(m\pi^2)\gamma_i} \left[1 - e^{-(m\pi)^2 \gamma_i \Delta t_{n+1}} \right]. \tag{171}$$

By multiplying equation (171) by 2 (for convenience in later manipulations) and defining

$$I_{mi,n} = 2 \int_0^{t_n} \frac{d\hat{h}_i(\tau)}{d\tau} e^{-(m\pi)^2 \gamma_i \left(t_n - \tau\right)} d\tau, \tag{172}$$

a recursive relation for evaluating the integral is obtained as

$$I_{mi,n+1} = e^{-(m\pi)^2 \gamma_i \Delta t_{n+1}} I_{mi,n} + \frac{\hat{h}_{i,n+1} - \hat{h}_{i,n}}{\Delta t_{n+1}} \frac{2}{(m\pi)^2 \gamma_i} \left[1 - e^{-(m\pi)^2 \gamma_i \Delta t_{n+1}} \right], \tag{173}$$

where $I_{mi,0} = 0$. Equation (173) permits evaluation of $I_{mi,n+1}$ from $I_{mi,n}$ and the heads at only the current (n+1) and previous (n) time levels. In an analogous manner, $2(-1)^m$ times the first integral in equation (168) is evaluated as

$$J_{mi,n+1} = e^{-(m\pi)^2 \gamma_i \Delta t_{n+1}} J_{mi,n}$$

$$+ \frac{H_{i,n+1} - H_{i,n}}{\Delta t_{n+1}} \frac{2(-1)^m}{(m\pi)^2 \gamma_i} \left[1 - e^{-(m\pi)^2 \gamma_i \Delta t_{n+1}} \right], \tag{174}$$

where

$$J_{mi,n} = 2(-1)^m \int_0^{t_n} \frac{dH_i(\tau)}{d\tau} e^{-(m\pi)^2 \gamma_i \left(t_n - \tau\right)} d\tau. \tag{175}$$

Next, the infinite series in equation (168) are approximated by finite series so that a large number of terms given by equations (172) through (175) do not have to be computed and saved at each time level. The

58

coefficients of the finite series are determined so that a small number of terms of the finite series can be used to give a good approximation of results obtained with the infinite series. The approximations are

$$2\sum_{m=1}^{\infty}\int_0^{t_{n+1}}\frac{d\hat{h}_i(\tau)}{d\tau}\ e^{-(m\pi)^2\gamma_i\left(t_{n+1}-\tau\right)}d\tau \simeq \sum_{m=1}^{N_1}A'_m\int_0^{t_{n+1}}\frac{d\hat{h}_i(\tau)}{d\tau}\ e^{-\alpha_m\gamma_i\left(t_{n+1}-\tau\right)}d\tau \quad (176)$$

and

$$2\sum_{m=1}^{\infty}\left(-1\right)^m\int_0^{t_{n+1}}\frac{dH_i(\tau)}{d\tau}\ e^{-(m\pi)^2\gamma_i\left(t_{n+1}-\tau\right)}dt \simeq \sum_{m=1}^{N_2}B'_m\int_0^{t_{n+1}}\frac{dH_i(\tau)}{d\tau}\ e^{-\beta_m\gamma_i\left(t_{n+1}-\tau\right)}d\tau, \quad (177)$$

where A'_m, α_m, B'_m, and β_m are coefficients to be determined, and N_1 and N_2 are the numbers of terms in the two finite series.

By repeating the derivation leading to equation (171) using the approximations, it can be seen that equations (173) and (175) are approximated by

$$\hat{I}_{mi,n+1} = e^{-\alpha_m\gamma_i\Delta t_{n+1}}\hat{I}_{mi,n} + \frac{\hat{h}_{i,n+1} - \hat{h}_{i,n}}{\Delta t_{n+1}}\ \frac{A'_m}{\alpha_m\gamma_i}\left(1-e^{-\alpha_m\gamma_i\Delta t_{n+1}}\right), \quad (178)$$

where

$$\hat{I}_{mi,n} = A'_m\int_0^{t_n}\frac{d\hat{h}_i(\tau)}{d\tau}\ e^{-\alpha_m\gamma_i\left(t_n-\tau\right)}d\tau, \quad (179)$$

and

$$\hat{J}_{mi,n+1} = e^{-\beta_m\gamma_i\Delta t_{n+1}}\hat{J}_{mi,n} + \frac{H_{i,n+1} - H_{i,n}}{\Delta t_{n+1}}\ \frac{B'_m}{\beta_m\gamma_i}\left(1 - e^{-\beta_m\gamma_i\Delta t_{n+1}}\right), \quad (180)$$

where

$$\hat{J}_{mi,n} = B'_m\int_0^{t_n}\frac{dH_i(\tau)}{d\tau}\ e^{-\beta_m\gamma_i\left(t_n-\tau\right)}d\tau. \quad (181)$$

The leakage rate computed using equations (178) and (180) is approximately equal to the exact leakage rate computed using equations (173) and (174) if

$$2\sum_{m=1}^{\infty}\frac{1}{(m\pi)^2}\left(1 - e^{-(m\pi)^2\gamma_i\Delta t_{n+1}}\right) \simeq \sum_{m=1}^{N_1}\frac{A'_m}{\alpha_m}\left(1 - e^{-\alpha_m\gamma_i\Delta t_{n+1}}\right), \quad (182)$$

and

$$2\sum_{m=1}^{\infty}\frac{(-1)^m}{(m\pi)^2}\left(1 - e^{-(m\pi)^2\gamma_i\Delta t_{n+1}}\right) \simeq \sum_{m=1}^{N_2}\frac{B'_m}{\beta_m}\left(1 - e^{-\beta_m\gamma_i\Delta t_{n+1}}\right). \quad (183)$$

To obtain the best approximations, equations (182) and (183) should hold exactly when $\Delta t_{n+1} = 0$ and when $\Delta t_{n+1} \to \infty$. The first requirement helps

59

yield an accurate solution for small time elements and is automatically fulfilled because both pairs of series equal zero when $\Delta t_{n+1} = 0$. The second requirement ensures that the total yield from storage in the confining unit under a unit head change that is fixed indefinitely is preserved by the approximation (Herrera and Yates, 1977, p. 726-727). Herrera and Yates (1977, p. 727) found the accuracy of their approximate solutions to be highly sensitive to fulfillment of this requirement. By letting $\Delta t_{n+1} \rightarrow \infty$ in equations (182) and (183) and using the sums of the resulting infinite series (Herrera and Yates, 1977, p. 727), it can be seen that this requirement is fulfilled if

$$2 \sum_{m=1}^{\infty} \frac{1}{(m\pi)^2} = \frac{1}{3} = \sum_{m=1}^{N_1} \frac{A'_m}{\alpha_m} \qquad (184)$$

and

$$2 \sum_{m=1}^{\infty} \frac{(-1)^m}{(m\pi)^2} = -\frac{1}{6} = \sum_{m=1}^{N_2} \frac{B'_m}{\beta_m}. \qquad (185)$$

It remains to find coefficients A'_m, α_m, B'_m, and β_m so that approximations given by equations (182) and (183) are good with a small number of terms, N_1 and N_2. For notational convenience, dimensionless time element Δt_D is defined by $\Delta t_D = \gamma_i \Delta t_n$, and the series are denoted as

$$S_1\left(\Delta t_D\right) = 2 \sum_{m=1}^{\infty} \frac{1}{(m\pi)^2}\left[1 - e^{-(m\pi)^2 \Delta t_D}\right], \qquad (186)$$

$$S_2\left(\Delta t_D\right) = 2 \sum_{m=1}^{\infty} \frac{(-1)^m}{(m\pi)^2}\left[1 - e^{-(m\pi)^2 \Delta t_D}\right], \qquad (187)$$

$$M_1\left(\Delta t_D\right) = \sum_{m=1}^{N_1} A_m\left[1 - e^{-\alpha_m \Delta t_D}\right], \qquad (188)$$

and

$$M_2\left(\Delta t_D\right) = \sum_{m=1}^{N_2} B_m\left[1 - e^{-\beta_m \Delta t_D}\right], \qquad (189)$$

where

$$A_m = \frac{A'_m}{\alpha_m} \qquad (190)$$

and

$$B_m = \frac{B'_m}{\beta_m}. \qquad (191)$$

Because both sets of series are functions of only Δt_D, the coefficients A_m, B_m, α_m, and β_m can be uniquely obtained by fitting approximate, finite series $M_1\left(\Delta t_D\right)$ and $M_2\left(\Delta t_D\right)$ to infinite series $S_1\left(\Delta t_D\right)$ and $S_2\left(\Delta t_D\right)$, respectively, using a range of values for Δt_D sufficient to include most time-element sizes and values of γ_i.

The coefficients were determined using nonlinear least squares (see Cooley and Naff, 1990, p. 61-64) applied to the weighted sum of squared error functions

$$SS_1 = \sum_{n=1}^{P_t} \frac{\left[S_1\left(\Delta t_D\right) - M_1\left(\Delta t_D\right)\right]^2}{\left|S_1(\Delta t_D)\right|} \tag{192}$$

and

$$SS_2 = \sum_{n=1}^{P_t} \frac{\left[S_2\left(\Delta t_D\right) - M_2\left(\Delta t_D\right)\right]^2}{\left|S_2\left(\Delta t_D\right)\right|} \tag{193}$$

subject to the constraints given by equations (184) and (185). The number of dimensionless times used in the fitting process, P_t, was set equal to 25, and $\Delta t_D = 1 \times 10^{-6}$, 2.5×10^{-6}, 5×10^{-6}, 1×10^{-5}, 2.5×10^{-5}, $5 \times 10^{-5}, \ldots, 1 \times 10^{2}$. The weights $1/\left|S_1\left(\Delta t_D\right)\right|$ and $1/\left|S_2\left(\Delta t_D\right)\right|$ are somewhat arbitrary, but were found to give good approximations for both small and large dimensionless time elements. Constraints were applied by specifying

$$A_{N_1} = \frac{1}{3} - \sum_{m=1}^{N_1-1} A_m \tag{194}$$

and

$$B_{N_2} = -\frac{1}{6} - \sum_{m=1}^{N_2-1} B_m. \tag{195}$$

Thus, the coefficients determined by nonlinear least squares are N_1 values of α_m and N_1-1 values of A_m for equation (192) and N_2 values of β_m and N_2-1 values of B_m for equation (193).

Good fits for both least squares problems were obtained with $N_1 = 3$ and $N_2 = 2$, and the resulting approximations are illustrated in figures 15 and 16. Values of the coefficients determined are

$A_1 = 0.26484,$ $\alpha_3 = 49538,$

$A_2 = 0.060019,$ $B_1 = -0.25754,$

$A_3 = 0.0084740,$ $B_2 = 0.090873,$

$\alpha_1 = 13.656,$ $\beta_1 = 10.764,$

$\alpha_2 = 436.53,$ $\beta_2 = 19.805.$
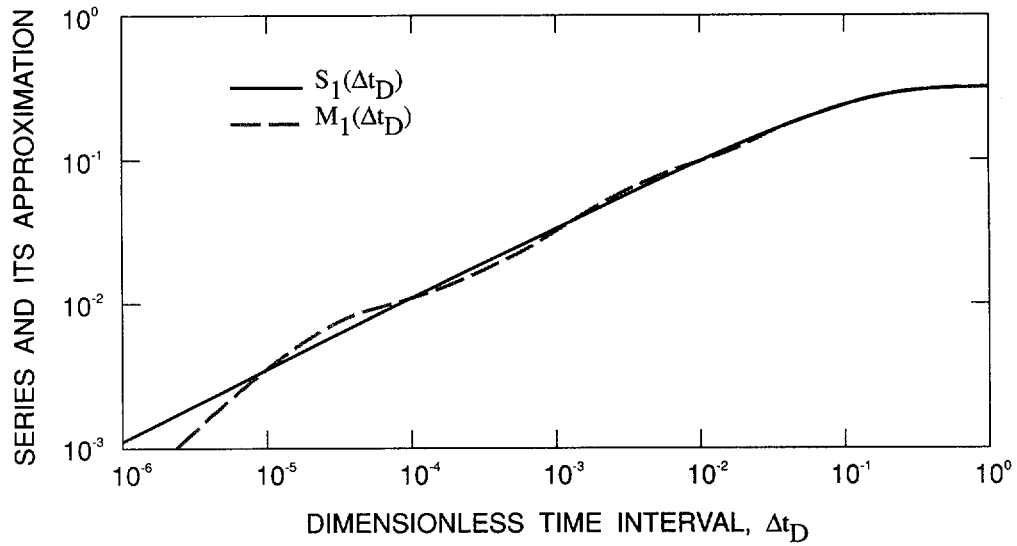
Figure 15. Relationship between series $S_1(\Delta t_D)$ defined by equation (186) and its approximation $M_1(\Delta t_D)$ defined by equation (188).
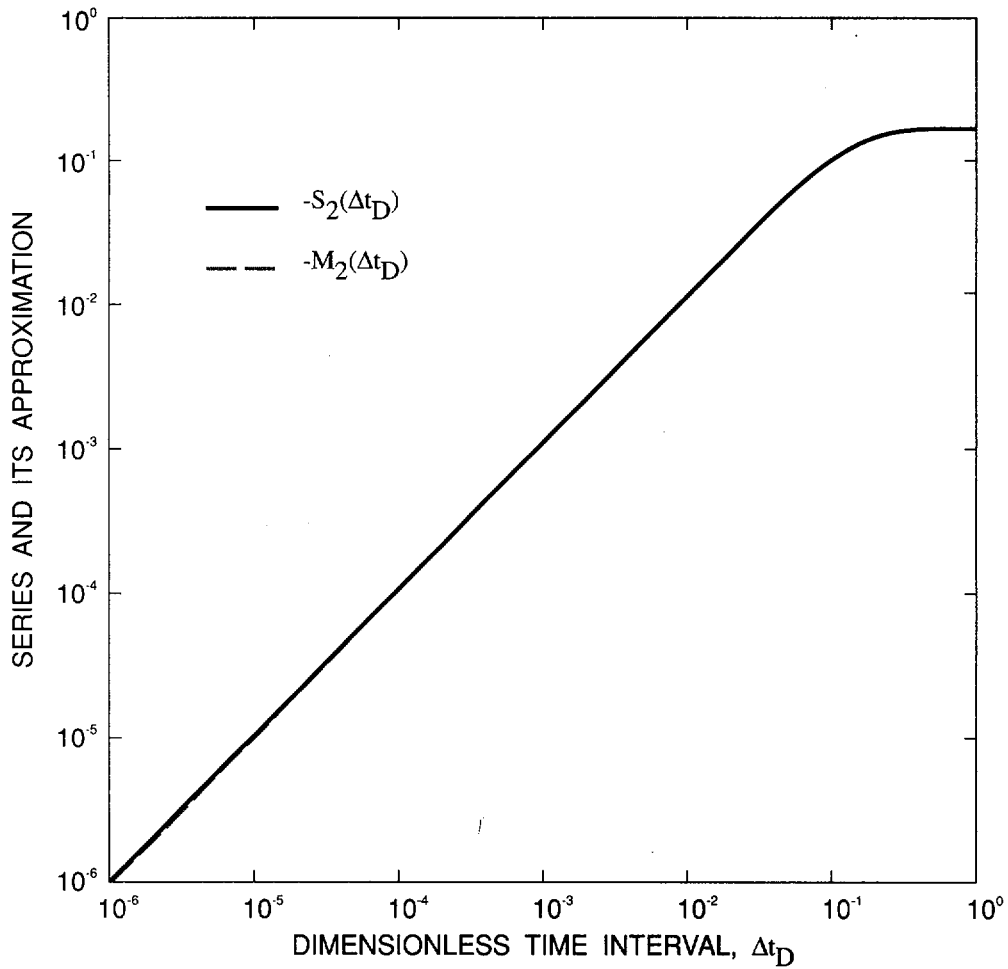
Figure 16. Relationship between series $-S_2(\Delta t_D)$ defined by equation (187) and its approximation $-M_2(\Delta t_D)$ defined by equation (189).

The final step leading to the leakage function to incorporate into the finite-element equations is to integrate the product of equation (168) and σ_{n+1} over time element n+1 by assuming that $\partial h_i'/\partial z\big|_{z=z_t}$ varies linearly like $\hat{h}_i$ (which is consistent with the treatment of R(H - h)). Therefore (see equation (54)),

$$\left[\frac{1}{3}\sum_{\bar{e}_i} K'^e_{zz}\Delta^e\right]\int_0^{\Delta t_{n+1}} \frac{\partial h_i'}{\partial z}\bigg|_{z=z_t} \sigma_{n+1}\,dt'$$

$$= \frac{1}{6}\Delta t_{n+1}\left[\frac{1}{3}\sum_{\bar{e}_i} K'^e_{zz}\Delta^e\right]\left[\frac{\partial h_{i,n}'}{\partial z} + 2\frac{\partial h_{i,n+1}'}{\partial z}\right]_{z=z_t} \tag{196}$$

By using equations (168), (170), and (176) through (181), the leakage rates at time levels n and n+1 in equation (196) are evaluated as, respectively,

$$\left[\frac{1}{3}\sum_{\bar{e}_i} K'^e_{zz}\Delta^e\right]\frac{\partial h_{i,n}'}{\partial z}\bigg|_{z=z_t} \simeq \left[\frac{1}{3}\sum_{\bar{e}_i} R^e\Delta^e\right]\left[\sum_{m=1}^{N_2}\hat{J}_{mi,n} - \sum_{m=1}^{N_1}\hat{I}_{mi,n} + H_{i,n} - \hat{h}_{i,n}\right] \tag{197}$$

and

$$\left[\frac{1}{3}\sum_{\bar{e}_i} K'^e_{zz}\Delta^e\right]\frac{\partial h_{i,n+1}'}{\partial z}\bigg|_{z=z_t} \simeq \left[\frac{1}{3}\sum_{\bar{e}_i} R^e\Delta^e\right]\left[\sum_{m=1}^{N_2} e^{-\beta_m\gamma_i\Delta t_{n+1}}\hat{J}_{mi,n}\right.$$

$$+ \frac{H_{i,n+1} - H_{i,n}}{\Delta t_{n+1}\gamma_i} M_2\left(\gamma_i\Delta t_{n+1}\right)$$

$$- \sum_{m=1}^{N_1} e^{-\alpha_m\gamma_i\Delta t_{n+1}}\hat{I}_{mi,n} - \frac{\hat{h}_{i,n+1} - \hat{h}_{i,n}}{\Delta t_{n+1}\gamma_i} M_1\left(\gamma_i\Delta t_{n+1}\right) + H_{i,n+1} - \hat{h}_{i,n+1}\right] \tag{198}$$

To obtain the leakage function to incorporate into the finite-element equations, equation (196) is multiplied by $-2/\Delta t_{n+1}$ and equations (197) and (198) are substituted into it. For notational compactness of the resulting expression, the following quantities are defined.

$$P_{hi,n} = \left[\frac{1}{3}\sum_{\bar{e}_i} R^e\Delta^e\right]\sum_{m=1}^{N_1}\hat{I}_{mi,n}, \tag{199}$$

$$P_{Hi,n} = \left[\frac{1}{3}\sum_{\bar{e}_i} R^e\Delta^e\right]\sum_{m=1}^{N_2}\hat{J}_{mi,n}, \tag{200}$$

$$Q_{hi,n} = \left[\frac{1}{3}\sum_{\bar{e}_i} R^e\Delta^e\right]\sum_{m=1}^{N_1} e^{-\alpha_m\gamma_i\Delta t_{n+1}}\hat{I}_{mi,n}, \tag{201}$$

$$Q_{Hi,n} = \left[\frac{1}{3}\sum_{e_i} R^e \Delta^e\right] \sum_{m=1}^{N_2} e^{-\beta_m \gamma_i \Delta t_{n+1}} \hat{J}_{mi,n},$$
(202)

$$C_{hi,n+1} = \left[\frac{1}{3}\sum_{e_i} R^e \Delta^e\right] \frac{M_1\left(\gamma_i \Delta t_{n+1}\right)}{\gamma_i},$$
(203)

$$C_{Hi,n+1} = \left[\frac{1}{3}\sum_{e_i} R^e \Delta^e\right] \frac{M_2\left(\gamma_i \Delta t_{n+1}\right)}{\gamma_i},$$
(204)

$$C_{Ri} = \frac{1}{3}\sum_{e_i} R^e \Delta^e.$$
(205)

Therefore, the final leakage function is

$$-\frac{1}{3}\left(P_{Hi,n} + 2Q_{Hi,n}\right) - \frac{2}{3}C_{Hi,n+1}\frac{H_{i,n+1} - H_{i,n}}{\Delta t_{n+1}} + \frac{1}{3}\left(P_{hi,n} + 2Q_{hi,n}\right)$$

$$+ \frac{2}{3}C_{hi,n+1}\frac{\hat{h}_{i,n+1} - \hat{h}_{i,n}}{\Delta t_{n+1}} - C_{Ri}\left[\frac{1}{3}\left(H_{i,n} - \hat{h}_{i,n}\right) + \frac{2}{3}\left(H_{i,n+1} - \hat{h}_{i,n+1}\right)\right]$$

$$= \left[\frac{C_{hi,n+1}}{\Delta t_{n+1}} + C_{Ri}\right]\delta_i - \frac{1}{3}\left(P_{Hi,n} + 2Q_{Hi,n}\right) - \frac{2}{3}C_{Hi,n+1}\frac{H_{i,n+1} - H_{i,n}}{\Delta t_{n+1}}$$

$$+ \frac{1}{3}\left(P_{hi,n} + 2Q_{hi,n}\right) - \frac{1}{3}C_{Ri}\left[H_{i,n} - \hat{h}_{i,n} + 2\left(H_{i,n+1} - \hat{h}_{i,n}\right)\right]$$
(206)

The leakage process described by equation (206) is time dependent, but linear. Therefore, equation (206) is added into equation (58), unless the predictor-corrector method is required to include other phenomena, in which case equation (206) is added into the appropriate predictor and corrector equations. The coefficient of δ_i is added into $\underline{V}$ for every node i where

leakage occurs, and the remaining terms are subtracted from the right-hand side.

FINITE-ELEMENT FORMULATION IN AXISYMMETRIC CYLINDRICAL COORDINATES

GOVERNING FLOW EQUATION AND BOUNDARY CONDITIONS

Axially symmetric ground-water flow in an aquifer is assumed to be governed by the following unsteady-state flow equation written in axisymmetric cylindrical coordinates (Bear, 1979, p. 116):

$$\frac{1}{r}\frac{1}{\partial r}\left(K_{rr}r\frac{\partial h}{\partial r}\right) + \frac{\partial}{\partial z}\left(K_{zz}\frac{\partial h}{\partial z}\right) = S_s\frac{\partial h}{\partial t},$$
(207)
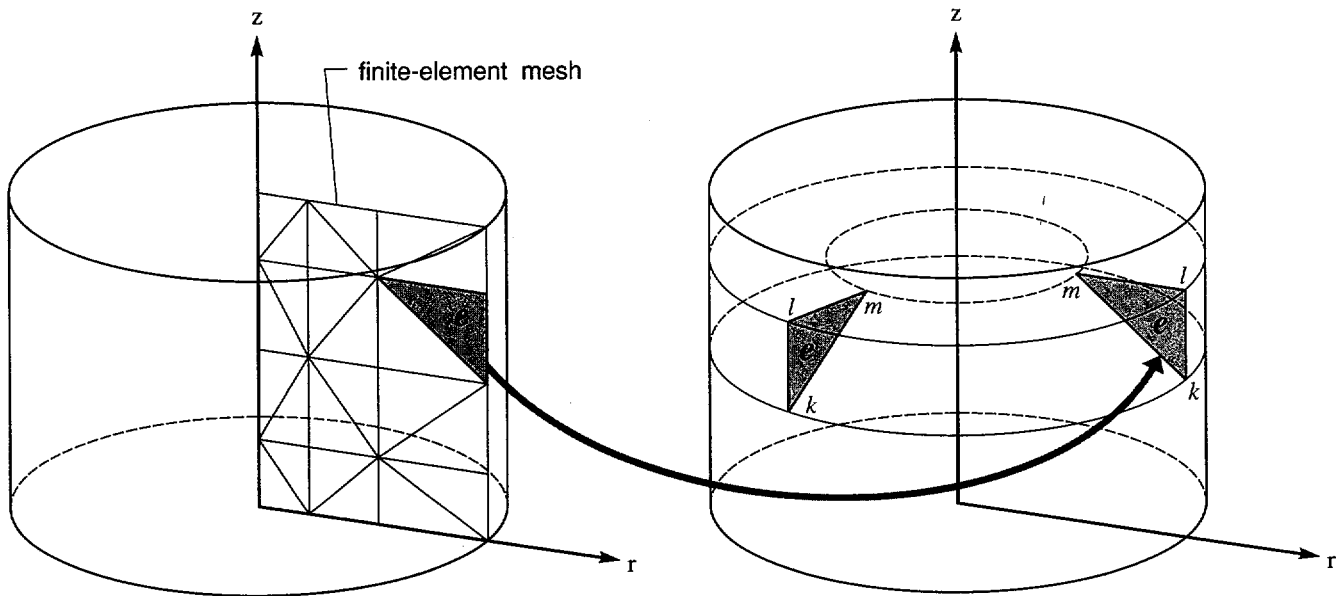
Figure 17. Axisymmetric aquifer subdivided into spatial finite elements.

where the new symbols used are

r = radial (horizontal) coordinate direction [length] from the axis of symmetry, which is vertical,

z = vertical coordinate direction [length],

$K_{rr}(r,z), K_{zz}(r,z)$ = the principal components of the hydraulic conductivity tensor [length/time] in the radial and vertical coordinate directions, respectively, and

$S_s(r,z)$ = specific storage [length^{-1}].

The orientation of the r and z axes is shown in figure 17.

The principal directions of the hydraulic conductivity tensor are assumed to be parallel to the r and z axes in equation (207). Equation (207) was not written in full component form like equation (1), because any rotation of the principal directions from the r and z axes (see figure 5) must be revolved around the z axis to maintain axial symmetry. This produces the physically unusual case of an axially symmetric rotation of the principal directions, which seemed to the author to be an unnecessary complication to include.

Equation (207) is subject to boundary and initial conditions analogous to equations (2) through (5) used for equation (1). However, equations (3) and (4) must be changed to reflect the change from flow integrated over aquifer thickness in equation (1) to flow in a radial cross-section in equation (207). Thus, equation (3), which expresses flow continuity across a discontinuity in the porous medium, is replaced by

$$v_n\big|_a = v_n\big|_b, \tag{208}$$

where $v_n(r,z,t)$ is the normal component of specific discharge, and equation (4), which expresses either a specified-head or Cauchy-type boundary condition, is replaced by

$$v_n = v_B + \alpha'(H_B - h), \tag{209}$$

65

where

$v_B(r,z,t)$ = specified specific discharge normal to the boundary [length/time] (positive for inflow), and

$\alpha'(r,z,t)$ = a parameter that, like α of equation (4), approaches infinity for a specified-head (Dirichlet) condition, is zero for a specified flow (Neumann) condition, and is finite and positive for a general or mixed (Cauchy) condition [time^{-1}].

FINITE-ELEMENT DISCRETIZATION

Finite-element discretization in axisymmetric cylindrical coordinates is accomplished in the same manner as for the Cartesian case. An r-z plane is subdivided into triangular elements (figure 17) over each of which the approximate solution h is assumed to vary linearly. Because of axial symmetry, each element is revolved around the symmetry axis so that it is a ringlike volume with triangular cross section. The time domain is subdivided into time elements over each of which the solution is also assumed to vary linearly. Therefore, the approximate solution is stated as an equation that is analogous to equation (14):

$$\hat{h} = \sum_i \left[\hat{h}_{i,n}\sigma_n + \hat{h}_{i,n+1}\sigma_{n+1} \right] N_i^e, \quad i = k,l,m, \qquad (210)$$

where

$$N_i^e = \left[a_i^e + b_i^e r + c_i^e z \right] / 2\Delta^e, \quad i = k,l,m, \qquad (211)$$

$$a_k^e = r_1 z_m - r_m z_1,$$

$$b_k^e = z_1 - z_m,$$

$$c_k^e = r_m - r_1,$$

$$a_1^e = r_m z_k - r_k z_m,$$

$$b_1^e = z_m - z_k, \qquad (212)$$

$$c_1^e = r_k - r_m,$$

$$a_m^e = r_k z_1 - r_1 z_k,$$

$$b_m^e = z_k - a_1,$$

$$c_m^e = r_1 - r_k,$$

$$2\Delta^e = \left(r_k - r_m \right) \left(z_1 - z_m \right) - \left(r_m - r_1 \right) \left(z_m - z_k \right), \qquad (213)$$

and σ_n and σ_{n+1} are defined by equation (13).

66

DERIVATION OF FINITE-ELEMENT EQUATIONS

The error-functional equation in axisymmetric cylindrical coordinates is analogous to equation (15). It is written for an r-z plane as

$$I(\hat{e}) = \sum_e \int_0^{\Delta t_{n+1}} \left\{ \int\int_{\Delta^e} \left[\frac{\partial \hat{e}}{\partial r} K_{rr} \frac{\partial \hat{e}}{\partial r} + \frac{\partial \hat{e}}{\partial z} K_{zz} \frac{\partial \hat{e}}{\partial z} + S_s \left(\frac{\partial \hat{e}}{\partial t} \right)^2 t' \right] r\,dr\,dz \right.$$

$$\left. + \int_{C_2^e} \alpha' \hat{e}^2 r\,dC \right\} dt' . \tag{214}$$

Minimization of equation (214) with respect to $\hat{h}_{i,n+1}$ and separation of the result into two parts as for equation (17) gives

$$\sum_{e_i} \int_0^{\Delta t_{n+1}} \sigma_{n+1} \left\{ \int\int_{\Delta^e} \left[N_i^e S_s \frac{\partial \hat{h}}{\partial t} + \frac{\partial N_i^e}{\partial r} K_{rr} \frac{\partial \hat{h}}{\partial r} + \frac{\partial N_i^e}{\partial z} K_{zz} \frac{\partial \hat{h}}{\partial z} \right] r\,dr\,dz \right.$$

$$\left. - \int_{C_2^e} N_i^e \left[v_B + \alpha' \left(H_B - \hat{h} \right) \right] r\,dC \right\} dt' - \sum_{e_i} \int_0^{\Delta t_{n+1}} \sigma_{n+1} \left\{ \int\int_{\Delta^e} \left[N_i^e S_s \frac{\partial h}{\partial t} + \frac{\partial N_i^e}{\partial r} K_{rr} \frac{\partial h}{\partial r} \right. \right.$$

$$\left. \left. + \frac{\partial N_i^e}{\partial z} K_{zz} \frac{\partial h}{\partial z} \right] r\,dr\,dz - \int_{C_2^e} N_i^e \left[v_B + \alpha' \left(H_B - h \right) \right] r\,dC \right\} dt' = 0. \tag{215}$$

By following the procedures used in appendix A, the second part of equation (215) can be shown to equal zero. Therefore, when the integrals involving $S_s \partial \hat{h}/\partial t$ and $\alpha'(H_B - \hat{h})$ are converted to diagonal form using approximations analogous to equation (19), the following integral form of the finite-element equations results:

$$\sum_{e_i} \int_0^{\Delta t_{n+1}} \sigma_{n+1} \left\{ \int\int_{\Delta^e} \left[N_i^e S_s \frac{d\hat{h}_i}{dt} + \frac{\partial N_i^e}{\partial r} K_{rr} \frac{\partial \hat{h}}{\partial r} + \frac{\partial N_i^e}{\partial z} K_{zz} \frac{\partial \hat{h}}{\partial z} \right] r\,dr\,dz \right.$$

$$\left. - \int_{C_2^e} N_i^e \left[v_B + \alpha' \left(H_{Bi} - \hat{h}_i \right) \right] r\,dC \right\} dt' = 0, \quad i = 1, 2, \ldots, N. \tag{216}$$

Equation (216) is analogous to equation (21). However, because the principal axes of the hydraulic conductivity tensor were originally assumed to be parallel to the r and z coordinate axes, no coordinate rotations are performed.

Equation (216) is integrated to obtain the final finite-element equations. As before, the spatial integrations are performed first. It is assumed that S_s, K_{rr}, and K_{zz} are all constant in each spatial element, and that v_B and α' are constant along any Cauchy-type boundary side of each element. The extra r in the integrals presents a complication not present for the Cartesian case. However, by writing r as the identity

$$r = N_k^e r_k + N_l^e r_l + N_m^e r_m, \tag{217}$$

equations (24) and (25) can be used to perform the integrations. Therefore, by substituting the appropriate expressions for $\hat{h}$, N_i^e, $\partial N_i^e/\partial r$, and $\partial N_i^e/\partial z$, i = k,l,m, the spatial integrals in equation (216) are evaluated for i = k (for example) as

$$\int_{\Delta^e}\int N_k^e S_s \frac{d\hat{h}_k}{dt} r\,dr\,dz = \frac{1}{12}S_s^e\left(2r_k + r_l + r_m\right)\Delta^e\frac{d\hat{h}_k}{dt}, \tag{218}$$

$$\int_{\Delta^e}\int \frac{\partial N_k^e}{\partial r} K_{rr} \frac{\partial \hat{h}}{\partial r} r\,dr\,dz = \int_{\Delta^e}\int \frac{\partial N_k^e}{\partial r} K_{rr}^e\left[\frac{\partial N_k^e}{\partial r}\hat{h}_k + \frac{\partial N_l^e}{\partial r}\hat{h}_l + \frac{\partial N_m^e}{\partial r}\hat{h}_m\right] r\,dr\,dz$$

$$= \frac{K_{rr}^e}{4\Delta^e}\left[b_k^e b_k^e \hat{h}_k + b_k^e b_l^e \hat{h}_l + b_k^e b_m^e \hat{h}_m\right]\bar{r}, \tag{219}$$

$$\int_{\Delta^e}\int \frac{\partial N_k^e}{\partial z} K_{zz} \frac{\partial \hat{h}}{\partial z} r\,dr\,dz = \int_{\Delta^e}\int \frac{\partial N_k^e}{\partial z} K_{zz}^e\left[\frac{\partial N_k^e}{\partial z}\hat{h}_k + \frac{\partial N_l^e}{\partial z}\hat{h}_l + \frac{\partial N_m^e}{\partial z}\hat{h}_m\right] r\,dr\,dz$$

$$= \frac{K_{zz}^e}{4\Delta^e}\left[c_k^e c_k^e \hat{h}_k + c_k^e c_l^e \hat{h}_l + c_k^e c_m^e \hat{h}_m\right]\bar{r}, \tag{220}$$

$$\int_{C_2^e} N_k^e\left[v_B + \alpha'\left(H_{Bk} - \hat{h}_k\right)\right] r\,dC = \frac{1}{6}\left(2r_k + r_l\right)\left[\left(v_B L\right)_{kl} + \left(\alpha' L\right)_{kl}\left(H_{Bk} - \hat{h}_k\right)\right]$$

$$+ \frac{1}{6}\left(2r_k + r_m\right)\left[\left(v_B L\right)_{km} + \left(\alpha' L\right)_{km}\left(H_{Bk} - \hat{h}_k\right)\right], \tag{221}$$

where S_s^e, K_{rr}^e, and K_{zz}^e are the constant values of S_s, K_{rr}, and K_{zz} in element e,

$$\bar{r} = \frac{1}{3}\left(r_k + r_l + r_m\right), \tag{222}$$

and other terms were defined earlier.

The spatially integrated finite-element equation for node k is obtained by substituting equations (218) through (221) into equation (216) and using equations analogous to equations (41) and (42). The result is

$$
\sum_{e_i} \int_0^{\Delta t_{n+1}} \sigma_{n+1} \left\{ c_{kk}^e \frac{d\hat{h}_k}{dt} + \left(g_{kk}^e + v_{kk}^e \right) \hat{h}_k + g_{kl}^e \hat{h}_l + g_{km}^e \hat{h}_m \right.
$$

$$
- \frac{1}{6} \left[\left(2r_k + r_l \right) \left(v_B L \right)_{kl} + \left(2r_k + r_m \right) \left(v_B L \right)_{km} \right] - \frac{1}{6} \left[\left(2r_k + r_l \right) \left(\alpha' L \right)_{kl} \right.
$$

$$
\left. \left. + \left(2r_k + r_m \right) \left(\alpha' L \right)_{km} \right] H_{Bk} \right\} dt' = 0, \tag{223}
$$

where

$$
c_{kk}^e = \frac{1}{12} S_s^e \left(2r_k + r_l + r_m \right) \Delta^e, \tag{224}
$$

$$
v_{kk}^e = \frac{1}{6} \left[\left(2r_k + r_l \right) \left(\alpha' L \right)_{kl} + \left(2r_k + r_m \right) \left(\alpha' L \right)_{km} \right], \tag{225}
$$

$$
g_{kk}^e = -g_{kl}^e - g_{km}^e, \tag{226}
$$

$$
g_{kl}^e = \left[\frac{K_{rr}^e}{4\Delta^e} b_k b_l + \frac{K_{zz}^e}{4\Delta^e} c_k c_l \right] \bar{r}, \tag{227}
$$

$$
g_{km}^e = \left[\frac{K_{rr}^e}{4\Delta^e} b_k b_m + \frac{K_{zz}^e}{4\Delta^e} c_k c_m \right] \bar{r}, \tag{228}
$$

As before, equation (223) must apply to all N nodes of the finite-element mesh. These N equations are written in matrix form as equation (45), where C_{ij}, V_{ij}, and G_{ij} are given by equations (47) through (49), and

$$
B_i = \sum_{e_i} \left\{ \frac{1}{6} \sum_{j'} \left(2r_i + r_{j'} \right) \left[\left(v_B L \right)_{ij'} + \left(\alpha' L \right)_{ij'} H_{Bi} \right] \right\}. \tag{229}
$$

Specified-head boundaries are handled using equation (51).

To perform the final time integration, parameters S_s, K_{rr}, K_{zz}, and α' are all assumed to be constant in time, and specified boundary flux v_B is assumed to be linearly variable through each time element. Thus, the time integrated finite-element equations for axisymmetric flow are given by equation (58) with B replaced by $\bar{B}$ defined by equation (62). No nonlinear or other extensions are used.

69

FINITE-ELEMENT FORMULATION FOR STEADY-STATE FLOW

By definition, steady-state flow occurs when hydraulic heads do not vary with time, or $\partial h/\partial t = 0$ in equation (1) or (207). Steady-state flow equations are obtained by setting S or S_s to zero and letting all quantities in equations (1) through (4) or (207) through (209) be time invariant.

LINEAR CASE

The finite-element equations for steady-state confined flow in the absence of any of the nonlinear source-sink functions may be derived from equation (56) by setting $\underline{\underline{C}}$ (which contains S or S_s) to zero and setting $\hat{\underline{h}}_{n+1} = \hat{\underline{h}}_n = \hat{\underline{h}}$. The resulting equation is

$$\underline{\underline{A}}\hat{\underline{h}} = \underline{B}, \tag{230}$$

where $\underline{\underline{A}} = \underline{\underline{G}} + \underline{\underline{V}}$. As for unsteady flow, round-off error may be minimized by solving for head change rather than head. Thus, by defining this head change, $\underline{\delta}_o$, as

$$\underline{\delta}_o = \hat{\underline{h}} - \underline{h}_o, \tag{231}$$

where $\underline{h}_o$ is an arbitrary initial set of heads close to $\hat{\underline{h}}$, equation (230) is modified to become

$$\underline{\underline{A}}\underline{\delta}_o = \underline{B} - \underline{\underline{A}}\underline{h}_o. \tag{232}$$

To solve a linear, steady-state flow problem, first equation (232) is solved for $\underline{\delta}_o$, then equation (231) is solved for $\hat{\underline{h}}$. Mass balance components are obtained from equation (232) using analogous procedures to those used for unsteady-state flow.

NONLINEAR CASE

If steady-state flow is unconfined or a nonlinear source-sink function is employed, then $\underline{\underline{A}}$ and $\underline{B}$ are functions of $\hat{\underline{h}}$ and a nonlinear problem results. In the case of unconfined flow, an off-diagonal entry of $\underline{\underline{G}}$ is given by equation (74) in which $b_i = \hat{h}_i - z_{bi}$. The particular form of the nonlinear source-sink term incorporated into $\underline{V}$ and $\underline{B}$ is dependent on the type of function: point head-dependent discharge, areal head-dependent leakage, areal head-dependent discharge, or line head-dependent leakage.

The general algorithm used to solve the nonlinear problem is derived before stating the specific terms used for unconfined flow and nonlinear sources and sinks. For nonlinear problems, equation (230) is restated as

$$\underline{\underline{A}}\left(\hat{\underline{h}}\right)\hat{\underline{h}} = \underline{B}\left(\hat{\underline{h}}\right), \tag{233}$$

where the notation $\underline{\underline{A}}\left(\hat{\underline{h}}\right)$ and $\underline{B}\left(\hat{\underline{h}}\right)$ signifies that $\underline{\underline{A}}$ and $\underline{B}$ are functions of $\hat{\underline{h}}$.

An iterative solution method for equation (233) may be derived by adding and

subtracting $\underline{\underline{A}}\left[\hat{\underline{h}}\right]$ to the right-hand side, then restating the result as an iteration equation of the form

$$\underline{\underline{A}}\left(\hat{\underline{h}}_\ell\right)\hat{\underline{h}}_{\ell+1} = \underline{\underline{A}}\left(\hat{\underline{h}}_\ell\right)\hat{\underline{h}}_\ell + \underline{B}\left(\hat{\underline{h}}_\ell\right) - \underline{\underline{A}}\left(\hat{\underline{h}}_\ell\right)\hat{\underline{h}}_\ell,$$

or

$$\underline{\underline{A}}_\ell \delta_\ell = \underline{r}_\ell \qquad (234)$$

where ℓ is the iteration index, and

$$\underline{\underline{A}}_\ell = \underline{\underline{A}}\left(\hat{\underline{h}}_\ell\right), \qquad (235)$$

$$\underline{B}_\ell = \underline{B}\left(\hat{\underline{h}}_\ell\right), \qquad (236)$$

$$\underline{\delta}_\ell = \hat{\underline{h}}_{\ell+1} - \hat{\underline{h}}_\ell, \qquad (237)$$

$$\underline{r}_\ell = \underline{B}_\ell - \underline{\underline{A}}_\ell \hat{\underline{h}}_\ell. \qquad (238)$$

Head-change vector $\underline{\delta}_\ell$ in equation (234) frequently requires damping to reduce undesirable oscillations from one iteration to the next. Addition of a damping parameter ρ_ℓ $(0 < \rho_\ell \le 1)$ to equation (237) yields the following iteration algorithm:

$$\left.\begin{array}{l} \underline{r}_\ell = \underline{B}_\ell - \underline{\underline{A}}_\ell \hat{\underline{h}}_\ell \\[6pt] \underline{\delta}_\ell = \underline{\underline{A}}_\ell^{-1} \underline{r}_\ell \\[6pt] \hat{\underline{h}}_{\ell+1} = \rho_\ell \underline{\delta}_\ell + \hat{\underline{h}}_\ell \end{array}\right\} \quad \ell = 0,1,\ldots \qquad (239)$$

The iterations are terminated when

$$\max_i \left|\delta_i^\ell\right| \le \epsilon_s, \qquad (240)$$

where δ_i^ℓ is a component of $\underline{\delta}_\ell$ and ϵ_s is a small number about an order of magnitude smaller than the desired accuracy in $\hat{h}$.

An effective empirical scheme for computing ρ_ℓ was developed by Cooley (1983, p. 1274). It is given in three steps. Let e_ℓ be the value of δ_i^ℓ that is largest in absolute value for all $i = 1,2,\ldots,N$, and let e_{max} be the largest value of $\left|e_\ell\right|$ permitted on any iteration. Then,

Step 1

$$p = \frac{e_\ell}{\rho_{\ell-1} e_{\ell-1}}, \quad \ell > 1$$

$$p = 1 \qquad\quad, \quad \ell = 1 \qquad\qquad (241)$$

Step 2

$$\rho^* = \frac{3 + p}{3 + |p|}, \quad p \ge -1$$

$$\rho^* = \frac{1}{2|p|} \quad\quad, \quad p < -1 \qquad\qquad (242)$$

71

Step 3

$$\rho_\ell = \rho^*, \quad \rho^* \left| e_\ell \right| \le e_{max}$$

$$\rho_\ell = \frac{e_{max}}{\left| e_\ell \right|}, \quad \rho^* \left| e_\ell \right| > e_{max} \tag{243}$$

A good trial value for e_{max} is about half the maximum value of $\left| \underline{h}_{oi} - \hat{\underline{h}}_i \right|$ expected (where $\underline{h}_{oi}$ is a component of the initial head vector). Much smaller values may be needed for highly nonlinear problems.

At the beginning of each iteration ℓ, $\underline{\underline{A}}_\ell$ and $\underline{B}_\ell$ must be recomputed using the newly computed values of $\hat{\underline{h}}_\ell$. The way in which $\underline{\underline{A}}_\ell$ and $\underline{B}_\ell$ are recomputed depends on the source of the nonlinearity. Nonlinearity from unconfined flow results when $\hat{h}_i < z_{ti}$. To allow for the possibility of unconfined flow, off-diagonals of $\underline{\underline{G}}$ are computed from equation (74) written for iteration ℓ as

$$G_{ij}^\ell = \frac{1}{2}\left[b_i^\ell + b_j^\ell \right] D_{ij}, \tag{244}$$

where

$$b_i^\ell = b_i^{\ell-1} + \rho_\ell \delta_i^{\ell-1}, \quad \hat{h}_i^\ell < z_{ti}$$

$$b_i^\ell = z_{ti} - z_{bi}, \qquad \hat{h}_i^\ell \ge z_{ti} \tag{245}$$

Nodes that go dry are treated in the same manner as explained for unsteady-state flow. The head is allowed to decline below the base of the aquifer at a dry node i, but horizontal flow in the aquifer is allowed between adjacent nodes i and j unless node j also goes dry. If a dry node i becomes surrounded by dry nodes during the iterative solution process, then $G_{ii} = 0$ and flow can only take place vertically through an underlying confining unit or across a Cauchy-type boundary at the dry node. If there is no confining unit or Cauchy-type boundary so that V_{ii} is also zero, then A_{ii}, which is $G_{ii} + V_{ii}$, is zero so that the head at node i is undefined and must be removed from the solution. This is accomplished by setting A_{ii} to 10^{30} and setting the right-hand side of the equation to zero, which holds the head constant at the last computed value. If the sum of the known fluxes is negative at a dry node, this sum is reduced by 1/2 at each iteration until the node remains saturated. As discussed earlier, this tells the user the approximate discharge that can be sustained at the node.

Nonlinear source-sink functions require reevaluation of $\underline{\underline{V}}_\ell$ and $\underline{B}_\ell$. For point head-dependent discharge functions, reevaluation is based on equation (106) written for iteration ℓ as

$$Q_{pi}^\ell = \begin{cases} C_{pi}\left(z_{pi} - \hat{h}_i^{\ell+1} \right), & \hat{h}_i^\ell > z_{pi} \\ 0 & , \hat{h}_i^\ell \le z_{pi} \end{cases} . \tag{246}$$

72

The terms that add into equation (234) are found by converting Q_{pi} to residual form using equation (237). That is, if $\hat{h}_i^\ell > z_{pi}$, then

$$Q_{pi}^\ell = C_{pi}\left[z_{pi} - \hat{h}_i^{\ell+1}\right] = -C_{pi}\delta_i^\ell + C_{pi}\left[z_{pi} - \hat{h}_i^\ell\right] \qquad (247)$$

so that C_{pi} is added into V_{ii}^ℓ and $C_{pi}\left[z_{pi} - \hat{h}_i^\ell\right]$ is added to the right-hand side. If $\hat{h}_i^\ell \leq z_{pi}$, then no terms are added.

Use of the other nonlinear source-sink functions is analogous. For areal head-dependent leakage, equation (118) for iteration ℓ is

$$Q_{ai}^\ell = \begin{cases} C_{ai}\left[H_{ai} - \hat{h}_i^{\ell+1}\right], & \hat{h}_i^\ell > z_{ti} \\ C_{ai}\left[H_{ai} - z_{ti}\right], & \hat{h}_i^\ell \leq z_{ti} \end{cases} \qquad (248)$$

so that, when $\hat{h}_i^\ell > z_{ti}$,

$$Q_{ai}^\ell = C_{ai}\left[H_{ai} - \hat{h}_i^{\ell+1}\right] = -C_{ai}\delta_i^\ell + C_{ai}\left[H_{ai} - \hat{h}_i^\ell\right]. \qquad (249)$$

Therefore, when $\hat{h}_i^\ell > z_{ti}$, equation (234) is modified by adding C_{ai} into V_{ii}^ℓ and adding $C_{ai}\left[H_{ai} - \hat{h}_i^\ell\right]$ to the right-hand side. When $\hat{h}_i^\ell \leq z_{ti}$, $C_{ai}\left[H_{ai} - z_{ti}\right]$ is added to the right-hand side and V_{ii}^ℓ is not modified.

Likewise, for areal head-dependent discharge functions, equation (131) is written for iteration ℓ as

$$Q_{ei}^\ell = \begin{cases} C_{ei}\left[z_{ei} - z_{ti}\right], & \hat{h}_i^\ell \geq z_{ti} \\ C_{ei}\left[z_{ei} - \hat{h}_i^{\ell+1}\right], & z_{ei} < \hat{h}_i^\ell < z_{ti} \\ 0, & \hat{h}_i^\ell \leq z_{ei} \end{cases} \qquad (250)$$

so that, when $z_{ei} < \hat{h}_i^\ell < z_{ti}$,

$$Q_{ei}^\ell = C_{ei}\left[z_{ei} - \hat{h}_i^{\ell+1}\right] = -C_{ei}\delta_i^\ell + C_{ei}\left[z_{ei} - \hat{h}_i^\ell\right]. \qquad (251)$$

Substitutions into equation (234) are analogous to the previous case. Finally, for line head-dependent leakage functions, equation (154) is written for iteration ℓ as

$$Q_{ri}^\ell = \begin{cases} C_{ri}\left[H_{ri} - \hat{h}_i^{\ell+1}\right], & \hat{h}_i^\ell > z_{ri} \\ C_{ri}\left[H_{ri} - z_{ri}\right], & \hat{h}_i^\ell \leq z_{ri} \end{cases} \qquad (252)$$

so that, when $\hat{h}_i^\ell > z_{ri}$,

$$Q_{ri}^\ell = C_{ri}\left[H_{ri} - \hat{h}_i^{\ell+1}\right] = -C_{ri}\delta_i^\ell + C_{ri}\left[H_{ri} - \hat{h}_i^\ell\right]. \qquad (253)$$

73

SOLUTION OF MATRIX EQUATIONS

Some of the symbols used in previous sections are redefined in this section to avoid complex or nonstandard matrix-solution terminology. Thus, symbols defined in this section are for use in this section only.

DEFINITION OF MATRIX EQUATION

Equation (58) must be solved for each time level of a linear, unsteady-state flow problem, and equations (76) and (80) (the predictor-corrector equations) must be solved sequentially at each time level of a nonlinear, unsteady-state flow problem. Likewise, equation (232) must be solved for a linear, steady-state flow problem, and equation (234) must be solved for each iteration of a nonlinear, steady-state flow problem. All of these equations are linear and of the form

$$\underline{\underline{A}}\,\underline{x} = \underline{d}, \tag{254}$$

where definitions of the coefficient matrix $\underline{\underline{A}}$, the known vector $\underline{d}$, and the unknown vector $\underline{x}$ depend on the equation being solved. For example, for equation (58),

$$\underline{\underline{A}} = \frac{\underline{\underline{C}}}{(2/3)\Delta t_{n+1}} + \underline{\underline{G}} + \underline{\underline{V}}, \tag{255}$$

$$\underline{x} = \underline{\delta}, \tag{256}$$

$$\underline{d} = \underline{B} - \left(\underline{\underline{G}} + \underline{\underline{V}}\right)\hat{\underline{h}}_n. \tag{257}$$

Thus, $\underline{\underline{A}}$ is an $N \times N$ matrix, and $\underline{x}$ and $\underline{d}$ are N-vectors.

The location of nonzero entries in matrix $\underline{\underline{A}}$ depends on the finite-element mesh. Each row i of $\underline{\underline{A}}$ contains nonzero entries only corresponding to nodes in the patch of elements for node i. Therefore, unless N is very small, $\underline{\underline{A}}$ is sparse in that most entries in any row are zero. Also, if the nodes are numbered so that the difference between the largest and smallest node numbers in the patch is small compared to N, then $\underline{\underline{A}}$ is banded, which means that all nonzero entries in each row are clustered near the main diagonal. Because $\underline{\underline{A}}$ is derived from the positive definite forms in equation (15) or (214), it is symmetric and positive definite. Finally, as discussed previously, if all internal angles of the spatial elements are acute, $\underline{\underline{A}}$ is a Stieltjes matrix. Additional information on finite-element matrices can be found in Desai and Abel (1972, chap. 2).

If node i is a specified-head node, equation i of equation (254) is $x_i = \frac{2}{3}\left[H_{Bi,n+1} - \hat{h}_{i,n}\right]$ for unsteady-state flow and $x_i = H_{Bi} - \hat{h}_i$ for steady-state flow. Because x_i is known at all specified-head nodes, all of the corresponding equations may be eliminated from equation (254). This may be accomplished as a partitioning operation by numbering all specified-head nodes in the finite-element mesh last, which is accomplished automatically in the code. Terms in the remaining equations that contain values of x_i for the specified-head nodes are then transferred to the right-hand sides of these equations to become part of the known vector. In the remainder of this section, equation (254) is regarded as the reduced equation resulting from this partitioning operation.

74

SYMMETRIC-DOOLITTLE METHOD

The first of two alternative matrix-solution procedures is discussed in this section. This method is referred to as symmetric-Doolittle decomposition (Fox, 1965, p. 99-102, 104-105) and is generally the preferred direct solution method for finite-element equations (Desai and Abel, 1972, p. 21). It is a direct method because the solution is found directly in three steps as opposed to iteratively in an unspecified number of steps required by the second method. Direct solution is usually efficient whenever there are fewer than about 500 nodes (Gambolati and Volpi, 1982).

The symmetric-Doolittle method is based on the fact that the symmetric matrix $\underline{\underline{A}}$ can be uniquely factored into the product of three matrices (Fox, 1965, p. 105), so that

$$\underline{\underline{A}} = \underline{\underline{U}}^T \underline{\underline{D}}\, \underline{\underline{U}}, \tag{258}$$

where superscript T stands for transpose, $\underline{\underline{U}}$ is upper triangular of the form

$$\underline{\underline{U}} = \begin{bmatrix} \alpha_{11} & u_{12} & u_{13} & \cdots & u_{1N} \\ 0 & \alpha_{22} & u_{23} & \cdots & u_{2N} \\ 0 & 0 & \alpha_{33} & \cdots & u_{3N} \\ & & \cdots & & \\ 0 & 0 & 0 & \cdots & \alpha_{NN} \end{bmatrix}, \tag{259}$$

and $\underline{\underline{D}}$ is diagonal of the form

$$\underline{\underline{D}} = \begin{bmatrix} 1/\alpha_{11} & 0 & 0 & \cdots & 0 \\ 0 & 1/\alpha_{22} & 0 & \cdots & 0 \\ 0 & 0 & 1/\alpha_{33} & \cdots & 0 \\ & & \cdots & & \\ 0 & 0 & 0 & \cdots & 1/\alpha_{NN} \end{bmatrix}. \tag{260}$$

Factorization, which is the first step of the three-step solution, is accomplished by forming the product matrix $\underline{\underline{U}}^T\underline{\underline{D}}\,\underline{\underline{U}}$, setting each entry of this matrix equal to the corresponding entry of $\underline{\underline{A}}$, then solving for the unknown values of u_{ij} and α_{ii}, entry by entry.

Solution of equation (254) using the factorization given by equation (258) is accomplished as follows. By defining a vector $\underline{y}$ by

$$\underline{\underline{U}}\,\underline{x} = \underline{y}, \tag{261}$$

the combination of equations (254) and (258) can be written as

$$\underline{\underline{U}}^T\underline{\underline{D}}\,\underline{y} = \underline{d}. \tag{262}$$

The lower triangular form of $\underline{\underline{U}}^T\underline{\underline{D}}$ and the upper triangular form of $\underline{\underline{U}}$ permit equations (262) and (261) to easily be solved for $\underline{y}$ and $\underline{x}$, respectively, as the second and third steps of the solution procedure. By forming the product $\underline{\underline{U}}^T\underline{\underline{D}}\,\underline{y}$, it can be seen that the first equation in equation (262) contains only y_1 as an unknown, the second y_1 and y_2, and so forth, so that

the first equation may be solved for y_1, which is used in the second to solve for y_2, and so forth. Solution vector $\underline{x}$ is found in exactly the opposite way. The last equation in equation (261) contains only the last unknown, x_N, the second from the last x_N and x_{N-1}, and so forth, so that the last equation is solved for x_N, which is used in the second from the last to solve for x_{N-1}, and so forth.

For N equations with N unknown values in $\underline{x}$, the calculations may be stated in algorithmic form as

$$\left.\begin{array}{l} \alpha_{ii} = a_{ii} - \displaystyle\sum_{k=1}^{i-1} u_{ki}u_{ki}/\alpha_{kk} \\[3mm] u_{ij} = a_{ij} - \displaystyle\sum_{k=1}^{i-1} u_{ki}u_{kj}/\alpha_{kk} \\[3mm] u'_{ij} = u_{ij}/\alpha_{ii} \end{array}\right\} \begin{array}{l} i = 1,2,\ldots,N \\ j = i+1,i+2,\ldots,N, \end{array} \tag{263}$$

$$\left.\begin{array}{l} y_i = d_i - \displaystyle\sum_{k=1}^{i-1} u'_{ki}y_k \\[3mm] y'_i = y_i/\alpha_{ii} \end{array}\right\} \quad i = 1,2,\ldots,N, \tag{264}$$

$$x_i = y'_i - \sum_{k=i+1}^{N} u'_{ik}x_k, \qquad i = N,N-1,\ldots,1. \tag{265}$$

Equation (263) is known as the factorization step, equation (264) is the forward substitution step, and equation (265) is the backward substitution step.

When the above algorithm is applied to the banded matrix $\underline{A}$, entries in $\underline{U}$ outside of the band are always found to be zero. However, $\underline{U}$ has mostly nonzero entries within the band, even if the corresponding entries in $\underline{A}$ are mostly zeros. Therefore, the algorithm can be coded to operate on and store only entries within the band. Storage of $\underline{A}$ for efficient computer application of the solution algorithm is explained in part 3.

As with any direct solution method, the above method can generate inaccurate solutions for poorly conditioned equation systems, such as can occur when $\underline{A}$ is not diagonally dominant, or has weak diagonal dominance, and(or) has highly variable entries. Matrix $\underline{A}$ can have weak diagonal dominance if all internal angles in spatial elements are acute but R, S, and α in equations (1) and (4) (or S_s and α' in equations (207) and (209)) are

zero and there are few specified-head nodes. Matrix $\underline{A}$ may not be diagonally dominant if R, S, and α are zero, and one or more elements has an obtuse internal angle. Entries in $\underline{A}$ can be highly variable if values of transmissivity or element shapes are highly variable. An inaccurate solution generally results in a large mass imbalance.

76

MODIFIED INCOMPLETE-CHOLESKY CONJUGATE-GRADIENT METHOD

If N is large or the direct solution method produces large mass balance errors, then an iterative method should be used. An iterative solution method called the generalized conjugate-gradient method (GCGM) has been found by Gambolati and Volpi (1982) to be more efficient than the direct method for solving sets of finite-element equations when N ≥ 500. The iterative method used here is a combination of a variant of GCGM by Manteuffel (1980) with a preconditioning method by Wong (1979) designed to enhance the convergence rate.

Generalized conjugate-gradient method

The iteration equation for the GCGM method is derived by replacing $\underline{\underline{A}}$ with a coefficient matrix $\underline{\underline{M}}$ that is similar to $\underline{\underline{A}}$ but much easier to invert (Concus and others, 1975). Matrix $\underline{\underline{M}}$, known as a preconditioning matrix, is defined from the fact that $\underline{\underline{A}}$ can always be split into the sum of two matrices, $\underline{\underline{M}}$ and $\underline{\underline{N}}$ (Varga, 1962, p. 87-93), so that

$$\underline{\underline{A}} = \underline{\underline{M}} + \underline{\underline{N}}. \tag{266}$$

Therefore, because the combination of equations (254) and (266) gives

$$\underline{\underline{M}}\underline{x} = \underline{d} - \underline{\underline{N}}\underline{x}, \tag{267}$$

the iteration equation is

$$\underline{\underline{M}}\underline{x}_{k+1} = \underline{d} - \underline{\underline{N}}\underline{x}_k \tag{268}$$

or, written in residual form,

$$\underline{\underline{M}}\underline{s}_{k+1} = \underline{r}_k, \tag{269}$$

where

$$\underline{s}_{k+1} = \underline{x}_{k+1} - \underline{x}_k, \tag{270}$$

$$\underline{r}_k = \underline{d} - \underline{\underline{A}}\underline{x}_k.$$

The GCGM algorithm based on iteration equation (269) can be stated as (Concus and others, 1975, p. 7-8)

$$\left.\begin{array}{l} \underline{s}_k = \underline{\underline{M}}^{-1}\underline{r}_k \\[2mm] \underline{p}_k = \underline{s}_k \end{array}\right\} \ k = 0,$$

$$\left.\begin{array}{l} \underline{s}_k = \underline{\underline{M}}^{-1}\underline{r}_k \\[4mm] \beta_k = \dfrac{\underline{s}_k^T \underline{r}_k}{\underline{s}_{k-1}^T \underline{r}_{k-1}} \\[4mm] \underline{p}_k = \underline{s}_k + \beta_k \underline{p}_{k-1} \end{array}\right\} \ k = 1,2,\ldots, \tag{271}$$

77

$$\left.\begin{array}{c} \alpha_k = \dfrac{s_k^T r_k}{p_k^T A p_k} \\[2ex] x_{k+1} = x_k + \alpha_k p_k \\[1ex] r_{k+1} = r_k + \alpha_k A p_k \end{array}\right\} \quad k = 0,1,2,\ldots,$$

Equations (271) can be derived using the idea that, if a set of linearly independent vectors p_k, $k = 1,2,\ldots,N$, can be obtained, then the solution x can be written as a linear combination of the p_k's because this set of vectors spans the N-dimensional space. Such a set of linearly independent vectors can be obtained by constructing them to be A-conjugate, that is, so that $p_i^T A p_j = 0$ if and only if $i \neq j$ (Beckman, 1967, p. 63). Coefficients β_k are calculated to construct this set of vectors. The proper linear combination of the p_k vectors to give the solution x is obtained by minimizing the error in the solution along the line $x_k + a p_k$ at each iteration (Beckman, 1967, p. 64). The value of "a" that minimizes this error is given by α_k.

In the absence of round-off error, the exact solution x is obtained in N iterations. Thus, if nearly N iterations were actually needed to obtain a good approximation of x, the method would not be useful for large systems of equations. Concus and others (1975) argued that the method can be considered to be a general iteration method that permits the gradual loss of A-orthogonality from round-off error and never converges to the exact solution. They showed that the weighted error function $\left[x - x_k\right]^T A \left[x - x_k\right]$ is reduced at each iteration if M and A are symmetric and positive definite, and that the method has certain optimality properties, so that, for a good choice of M, it usually converges to the desired accuracy in far fewer than N iterations.

Modified incomplete-Cholesky factorization

Modified incomplete-Cholesky factorization is an extension of a method introduced by Meijerink and van der Vorst (1977) known as incomplete-Cholesky factorization.[1] The extension is a combination of methods from Wong (1979) and Manteuffel (1980).

Wong's (1979) method, known as row-sums agreement factorization, is developed from incomplete-Cholesky factorization as follows. Let matrix entries located at $(\tilde{i},\tilde{j})$ be those entries corresponding to nonzero entries of A, and let $\tilde{U}$ be an upper triangular matrix with the same form as U, except that the only nonzero entries of $\tilde{U}$ are located at $(\tilde{i},\tilde{j})$. Finally, let $\tilde{D}$ be a diagonal matrix with the same form as D. Then an approximate (incomplete) factorization of A is defined by

[1]Meijerink and van der Vorst (1977) used an approximate factorization that is more like the symmetric-Dolittle method than the Cholesky method. However, it is still called incomplete-Cholesky factorization.

$$\underline{\underline{M}} = \underline{\underline{\tilde{U}}}^T \underline{\underline{\tilde{D}}} \, \underline{\underline{\tilde{U}}}, \qquad (272)$$

where $\underline{\underline{M}}$ will generally contain nonzero entries in addition to nonzero entries at the $(\tilde{i}, \tilde{j})$ locations because of fill-in generated by forming the product $\underline{\underline{\tilde{U}}}^T \underline{\underline{\tilde{D}}} \, \underline{\underline{\tilde{U}}}$. Both incomplete-Cholesky and row-sums agreement factorization are based on equation (272).

For incomplete-Cholesky factorization, entries of $\underline{\underline{\tilde{D}}}$ and $\underline{\underline{\tilde{U}}}$ are obtained by equating entries of $\underline{\underline{\tilde{U}}}^T \underline{\underline{\tilde{D}}} \, \underline{\underline{\tilde{U}}}$ with nonzero entries a_{ij} of $\underline{\underline{A}}$ and rearranging the results, so that

$$\tilde{\alpha}_{ii} = a_{ii} - \sum_{k=1}^{i-1} \tilde{u}_{ki}\tilde{u}_{ki}/\tilde{\alpha}_{kk}, \quad i = 1, 2, \ldots, N, \qquad (273)$$

$$\tilde{u}_{ij} = \begin{cases} a_{ij} - \sum\limits_{k=1}^{i-1} \tilde{u}_{ki}\tilde{u}_{kj}/\tilde{\alpha}_{kk}, & (i,j) \text{ belongs to } (\tilde{i},\tilde{j}) \\ 0 & , (i,j) \text{ does not belong to } (\tilde{i},\tilde{j}). \end{cases} \qquad (274)$$

It can be verified by direct calculation that entries of $\underline{\underline{M}}$ and $\underline{\underline{A}}$ located at $(\tilde{i}, \tilde{j})$ are identical. The two matrices differ because of fill-in in $\underline{\underline{M}}$. By assigning the negatives of the fill-in entries in $\underline{\underline{M}}$ to $\underline{\underline{N}}$ and letting all other entries in $\underline{\underline{N}}$ be zero, $\underline{\underline{A}} = \underline{\underline{M}} + \underline{\underline{N}}$, as required.

An ideal modification would make $\underline{\underline{N}}$ near zero, but this is not possible using equation (272) without adding nonzero entries to $\underline{\underline{\tilde{U}}}$. It is possible to modify $\underline{\underline{N}}$ to have the property of a zero matrix that the sum of entries of each row (a row sum) of $\underline{\underline{N}}$ equal zero. This is Wong's (1979) row-sum agreement factorization. To develop this method, each row sum of $\underline{\underline{A}}$ is set equal to each row sum of $\underline{\underline{M}}$ using equations (273) and (274) to define entries of $\underline{\underline{M}}$. Because $a_{ij} = a_{ji}$, entries of $\underline{\underline{A}}$ below the main diagonal where $\tilde{u}_{ij} = 0$ are given from equation (274) as $a_{ij} = \tilde{u}_{ji} + \sum\limits_{k=1}^{i-1} \tilde{u}_{kj}\tilde{u}_{ki}/\tilde{\alpha}_{kk}$ so that a row sum is

$$a_{ii} + \sum_{j=1}^{i-1} a_{ij} + \sum_{j=i+1}^{N} a_{ij}$$

$$= \tilde{\alpha}_{ii} + \sum_{k=1}^{i-1} \tilde{u}_{ki}\tilde{u}_{ki}/\tilde{\alpha}_{kk} + \sum_{j=1}^{i-1} \left[\tilde{u}_{ji} + \sum_{k=1}^{i-1} \tilde{u}_{kj}\tilde{u}_{ki}/\tilde{\alpha}_{kk} \right]$$

$$+ \sum_{j=i+1}^{N} \left[\tilde{u}_{ij} + \sum_{k=1}^{i-1} \tilde{u}_{ki}\tilde{u}_{kj}/\tilde{\alpha}_{kk} \right]. \qquad (275)$$

The factorization is obtained from equation (275) by computing all values of $\tilde{u}_{ij}$ using equation (274), so that all nonzero off-diagonal entries of $\underline{\underline{A}}$ cancel with their corresponding entries of $\underline{\underline{M}}$. Thus, the only remaining entries in the sums on j in equation (275) result from fill-in, for which $\tilde{u}_{ij} = 0$, so that

$$a_{ii} = \tilde{\alpha}_{ii} + \sum_{k=1}^{i-1} \tilde{u}_{ki}\tilde{u}_{ki}/\tilde{\alpha}_{kk} + \sum_{j=1}^{i-1} f_{ji} + \sum_{j=i+1}^{N} f_{ij}, \qquad (276)$$

where f_{ij} is a fill-in entry of $\underline{\underline{M}}$ defined by

$$f_{ij} = \begin{cases} \sum_{k=1}^{i-1} \tilde{u}_{ki}\tilde{u}_{kj}/\tilde{\alpha}_{kk}, & (i,j) \text{ does not belong to } (\tilde{i},\tilde{j}) \\ 0, & (i,j) \text{ belongs to } (\tilde{i},\tilde{j}). \end{cases} \qquad (277)$$

Diagonal entry $\tilde{\alpha}_{ii}$ is calculated from equation (276) as

$$\tilde{\alpha}_{ii} = a_{ii} - \sum_{k=1}^{i-1} \tilde{u}_{ki}\tilde{u}_{ki}/\tilde{\alpha}_{kk} - \sum_{j=1}^{i-1} f_{ji} - \sum_{j=i+1}^{N} f_{ij}. \qquad (278)$$

Comparison of equations (273) and (278) shows that diagonal entries of $\underline{A}$ no longer equal diagonal entries of $\underline{\underline{M}}$ defined by equation (272).

The method from Manteuffel (1980) forces $\underline{\underline{M}}$ to be positive definite, as required by the generalized conjugate-gradient method. If $\underline{A}$ is not a Stieltjes matrix, then $\underline{\underline{M}}$ as defined using incomplete-Cholesky factorization may not be positive definite (Meijerink and van der Vorst, 1977), which means that α_{ii} computed by equation (273) will not be positive. In this case, $\underline{\underline{M}}$ as computed for row-sums agreement factorization also may not be positive definite because $\tilde{\alpha}_{ii}$ calculated using (278) may be even smaller.

Manteuffel (1980) showed that computation of $\tilde{\alpha}_{ii} \le 0$ for incomplete-Cholesky factorization of finite-element matrices can be prevented by adding the product of an empirically determined, small positive number, δ, and a_{ii} to the right-hand side of equation (273). The analogous modification of equation (278) is

$$\tilde{\alpha}_{ii} = (1 + \delta)a_{ii} - \sum_{k=1}^{i-1} \tilde{u}_{ki}\tilde{u}_{ki}/\tilde{\alpha}_{kk} - \sum_{j=1}^{i-1} f_{ji} - \sum_{j=i+1}^{N} f_{ij}. \qquad (279)$$

The matrix approximately factored by this modification of row-sums agreement factorization is thus $\underline{A} + \delta\underline{I}$ (where $\underline{\underline{I}}$ is the identity matrix), which is more diagonally dominant than $\underline{\underline{A}}$.

The above method is implemented here as follows. If $\tilde{\alpha}_{ii} \le 0$ is detected during factorization, then factorization is stopped and a new value of δ, δ_{new}, is computed from the old value, δ_{old}, using the empirical equation

$$\delta_{new} = \frac{3}{2}\delta_{old} + 0.001, \qquad (280)$$

where the initial value of δ_{old} is zero. Factorization is then reinitiated, and equation (280) is applied again if $\tilde{\alpha}_{ii} \le 0$ is detected again, and so forth. This process is continued until a large enough value of δ is computed that all $\tilde{\alpha}_{ii} > 0$.

Gustafsson (1978, 1979) also presented a method that yields equations having forms similar to those of equations (274), (277), and (279). However, his method applies to finite-difference approximations for which $\underline{A}$ is a Stieltjes matrix so that the motivation and method of choosing δ are different.

Based on equation (272), the solution of equation (269) is obtained using the forward and backward substitution steps of the symmetric-Doolittle method as

$$\tilde{\underline{U}}^T \tilde{\underline{D}} \underline{y}_k = \underline{r}_k,$$
$$\tilde{\underline{U}}^T \underline{s}_k = \underline{y}_k, \tag{281}$$

where entries of $\tilde{\underline{U}}$ and $\tilde{\underline{D}}$ are computed using equations (274) and (279), respectively. The remaining part of the algorithm implied by equations (281) is

$$\tilde{u}'_{ij} = \tilde{u}_{ij}/\tilde{\alpha}_{ii}, \quad (i,j) \text{ belongs to } (\tilde{I}, \tilde{J}), \tag{282}$$

$$\left.\begin{array}{l} y_i^k = r_i^k - \sum_{\ell=1}^{i-1} \tilde{u}'_{\ell i} y_\ell^k \\[2ex] y_i'^k = y_i^k/\tilde{\alpha}_{ii} \end{array}\right\} \quad i = 1, 2, \ldots, N, \tag{283}$$

$$s_i^k = y_i'^k - \sum_{\ell=i+1}^{N} \tilde{u}'_{i\ell} s_\ell^k, \quad i = N, N-1, \ldots, 1. \tag{284}$$

In applying the above algorithm, note that the factorization step to compute $\tilde{\underline{D}}$ and $\tilde{\underline{U}}$ is only done once before applying the generalized conjugate-gradient algorithm (equations (271)). At each iteration, $\underline{s}_k$ is computed

using only equations (283) and (284). The factorization, forward substitution, and backward substitution steps are all fast and require little computer storage because $\tilde{\underline{U}}$ is sparse like $\tilde{\underline{A}}$. This combination of GCGM and modified incomplete-Cholesky factorization is called the modified incomplete-Cholesky conjugate-gradient method (MICCG method).

Stopping criteria

One stopping criterion is to terminate the algorithm whenever the maximum value of $\left| x_i^{k+1} - x_i^k \right|$ becomes small, or whenever

$$\max_i \left| x_i^{k+1} - x_i^k \right| = \max_i \left| \alpha_k p_i^k \right| \le \epsilon, \tag{285}$$

where x_i^k is an entry of $\underline{x}_k$, p_i^k is an entry of $\underline{p}_k$, and ϵ is a small positive number, such as 10^{-4}. The value of $\max_i \left| x_i^{k+1} - x_i^k \right|$ is usually assumed to be a rough measure of the error $\max_i \left| x_i - x_i^k \right|$ in the solution. However, the

conjugate-gradient algorithm can sometimes yield a value of $\max\limits_{i} \left| x_i^{k+1} - x_i^k \right|$

that is small even when the solution is inaccurate. Thus, another criterion

that is also a rough measure of $\max\limits_{i} \left| x_i - x_i^k \right|$ is employed.

The residual given by equation (270) can be written for any row i as

$$a_{i1}\left(x_1 - x_1^k\right) + a_{i2}\left(x_2 - x_2^k\right) + \ldots + a_{iN}\left(x_N - x_N^k\right) = r_i^k \qquad (286)$$

Thus, because a_{ii} is positive,

$$\frac{\left|a_{i1}\right|}{a_{ii}}\left|x_1 - x_1^k\right| + \frac{\left|a_{i2}\right|}{a_{ii}}\left|x_2 - x_2^k\right| + \ldots + \frac{\left|a_{iN}\right|}{a_{ii}}\left|x_N - x_N^k\right| \geq \frac{\left|r_i^k\right|}{a_{ii}}, \qquad (287)$$

or

$$\frac{1}{a_{ii}} \sum_{j=1}^{N} \left|a_{ij}\right| \max_{i}\left|x_i - x_i^k\right| \geq \frac{\left|r_i^k\right|}{a_{ii}}. \qquad (288)$$

The sum $\sum\limits_{j=1}^{N} \left|a_{ij}\right|/a_{ii}$ is generally in the range of 1 to 2, so is assumed to

be unity. Therefore, a rough measure of $\max\limits_{i} \left| x_i - x_i^k \right|$ is $\left|r_i^k\right|/a_{ii}$, and the
additional stopping criterion is

$$\max_{i}\left|r_i^k\right|/a_{ii} \leq \epsilon. \qquad (289)$$

Note that if MICCG is used to solve the nonlinear equation (234), then
there will be an inner MICCG iteration loop and an outer loop on the
nonlinearity. An efficient way of employing MICCG for these problems is to
set the convergence criterion ϵ to be larger than normal (say, larger than
ϵ_s by about an order of magnitude) to reduce the number of inner iterations
taken at each outer iteration. Good accuracy is achieved by requiring close
convergence of the outer iteration sequence.

COMPARISONS OF NUMERICAL RESULTS WITH ANALYTICAL SOLUTIONS

Results of simulating some simple ground-water flow problems for which
analytical solutions have been presented in the literature are given here to
demonstrate the accuracy of the finite-element code (MODFE). Each simula-
tion is designed to test specific computational features that were discussed
in preceding sections and to verify that MODFE can accurately represent the
physical processes. To demonstrate that any consistent system of units may
be used with MODFE, both English and metric systems of units are used in the
example problems.

THEIS SOLUTION OF UNSTEADY RADIAL FLOW TO A PUMPED WELL

MODFE is used with axisymmetric cylindrical coordinates to compute
unsteady flow to a well located in a confined nonleaky aquifer having homo-
geneous and isotropic hydraulic properties and an infinite areal extent.

82

The pumped well fully penetrates the aquifer thickness (100 feet) and its diameter (1 foot) is not significant for the simulation. The analytical solution for drawdown is given by the Theis equation (Lohman, 1972, p. 15) as

$$s = \frac{Q}{4\pi T} W(u),$$

where s is drawdown [length], Q is volumetric discharge [length³/time], T is transmissivity [length²/time], and W(u) is the well function

$$\int_u^\infty \frac{e^{-v}}{v} dv,$$

where $u = r^2 S/4Tt$, r is radial distance from the well [length], and S is the storage coefficient [0].

Because of radial symmetry about the well bore, the problem can be simulated as an r - z plane section through the aquifer with the well located at the z axis (figure 18). The radial extent of the simulated-aquifer region is 8,000 feet, although the analytical solution was developed for an aquifer of infinite areal extent. This distance is beyond the influence of the pumped well during the simulation period so that the computed solution near the well is not affected by the boundary condition

at r = 8,000 feet. Radial node spacing was expanded by a factor of $\sqrt{2}$ starting with r = 125 feet to obtain the finite-element mesh composed of 52 triangular elements and 42 nodes shown in figure 18. The initial time-element size was 3×10^{-5} days, and an expansion factor of 1.25 was used to generate subsequent elements, to yield a total of 20 time elements. Other characteristics of the problem are:

$$T = 10^5 \text{ ft}^2/d,$$

$$S = 0.001,$$

$$Q = 160,000 \text{ ft}^3/d,$$

$$h(r,z,0) = 0 \text{ ft},$$

$$h(8,000,z,t) = 0 \text{ ft}.$$

These characteristics and time-element sizes are the same as used by Wilson and others (1979, p. 85-88) to test their finite-element code, except that
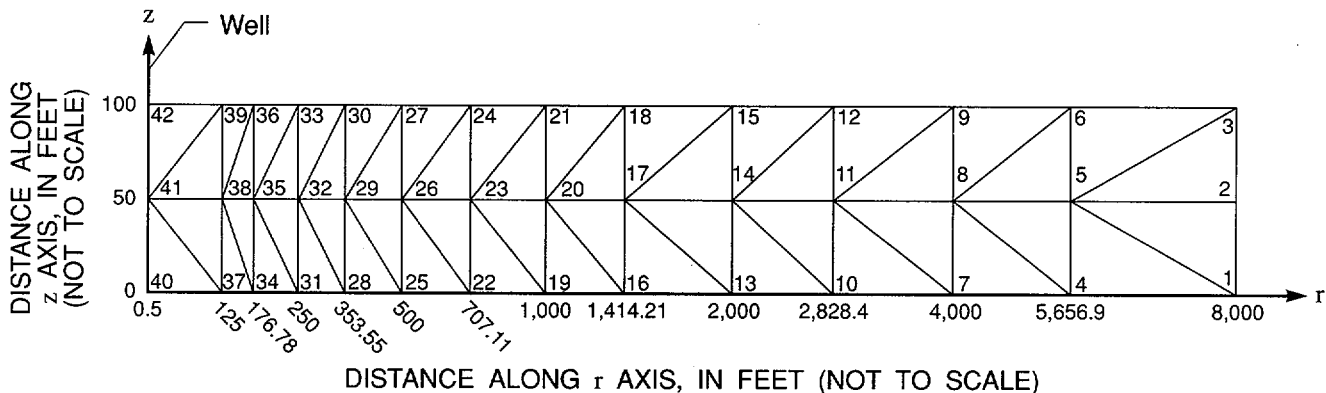

Figure 18. Finite-element mesh used to simulate unsteady-state radial flow to a pumped well.
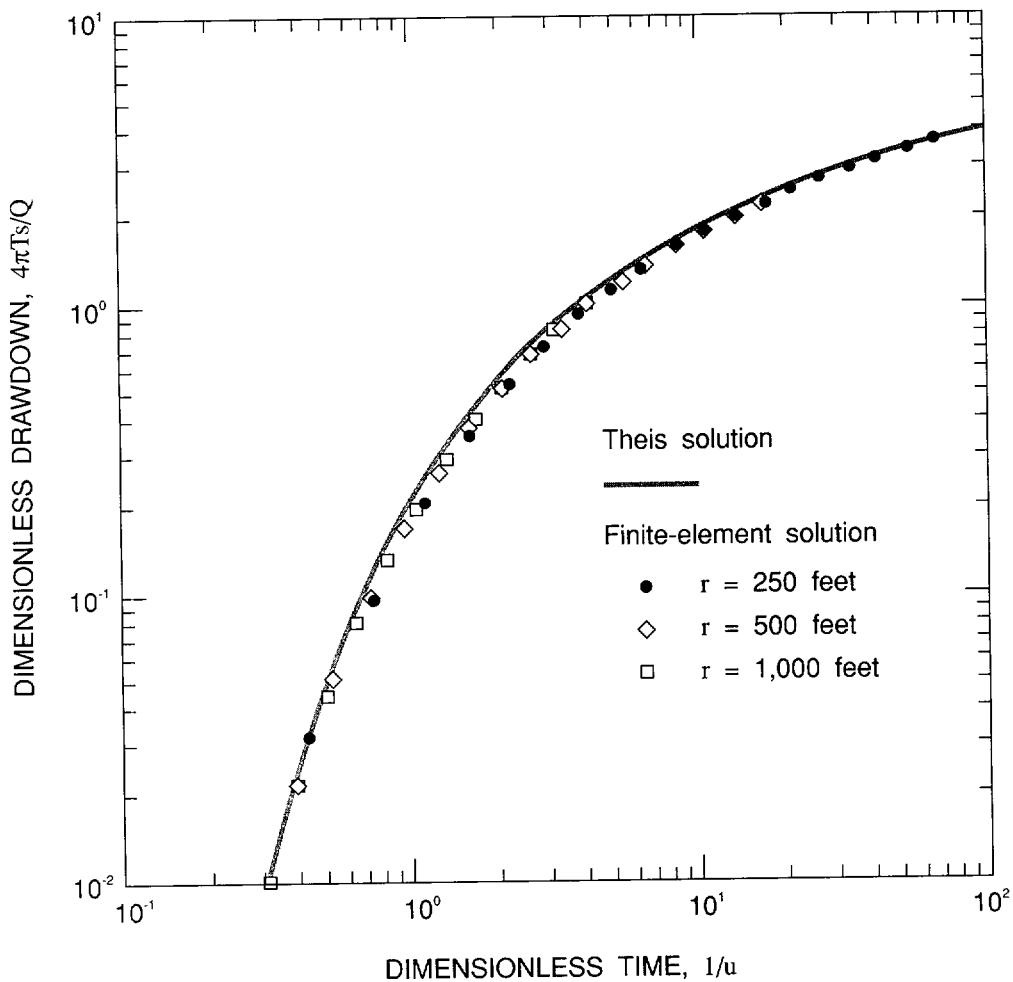
Figure 19. Theis solution (Lohman, 1972, p. 15) and finite-element results for unsteady radial flow to a pumped well.

lengths are designated as feet here rather than meters. Radial node spacing is also the same, but Wilson and others solved the problem using Cartesian coordinates.

To simulate confined flow in axisymmetric cylindrical coordinates, no-flow boundaries are placed along the aquifer top and bottom. Well discharge is simulated as a line sink at the well radius $r_w = 0.5$ feet using the specified-flow part $\left[v_B\right]$ of the Cauchy-type boundary condition (equation (209)). Because v_B is specific discharge (volumetric discharge per unit area), it is obtained from Q as follows:

$$v_B = \frac{-Q}{2\pi r_w b} = \frac{-16,000}{2\pi(0.5)(100)} = -509.296 \text{ ft/d}$$

Computed values of $4\pi Ts/Q$ versus $1/u$ for the radial distances of 250, 500, and 1,000 feet are compared with the type curve of the Theis solution in figure 19. The numerical results show good agreement with the analytical solution and are nearly the same as obtained by Wilson and others (1979, p. 87) in Cartesian coordinates.

HANTUSH SOLUTION OF UNSTEADY RADIAL FLOW TO A PUMPED WELL
IN A LEAKY AQUIFER

The effects of release of water stored in an elastic confining layer
(transient leakage) in the vicinity of a well pumped at a constant rate in a
confined, homogeneous, isotropic aquifer of infinite areal extent are
contained in an analytical solution by Hantush (1960) (figure 20). The
analytical solution for drawdown in this flow system is stated as (see
Hantush, 1960, figure 5)

$$s = \frac{Q}{4\pi T} H(u, \beta'),$$

where $H(u, \beta')$ is an infinite integral that equals the well function $W(u)$
when $\beta' = 0$, and

$$\beta' = \frac{1}{4} \frac{r}{B} \sqrt{\frac{S'}{S}},$$

where S' is the storage coefficient $(S_s' b')$ of the confining unit and B is
$\sqrt{Tb'/K'}$ [length].

The flow problem could be conceptualized with axisymmetric cylindrical
coordinates, as in the first simulation. However, in order to test the
transient-leakage algorithm, the flow system is represented by Cartesian
coordinates. Radial symmetry is used to reduce the size of the flow domain
by simulating a 22.5-degree wedge of the total flow system (see figure 21).

The finite-element mesh used in this simulation consists of 86
triangular elements and 67 nodes (figure 21) and extends 32,000 feet from
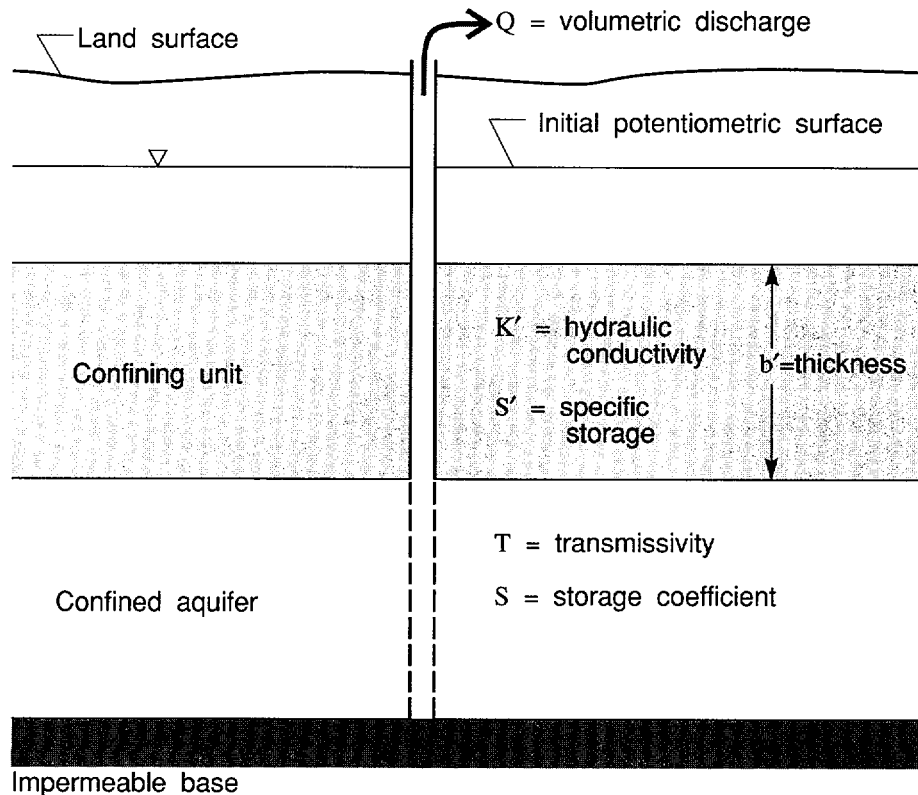the pumped well, which is placed at node 1. Node spacing increases radially

Figure 20. Geometry used to simulate the effects of transient leakage on
drawdown near a pumped well.

85
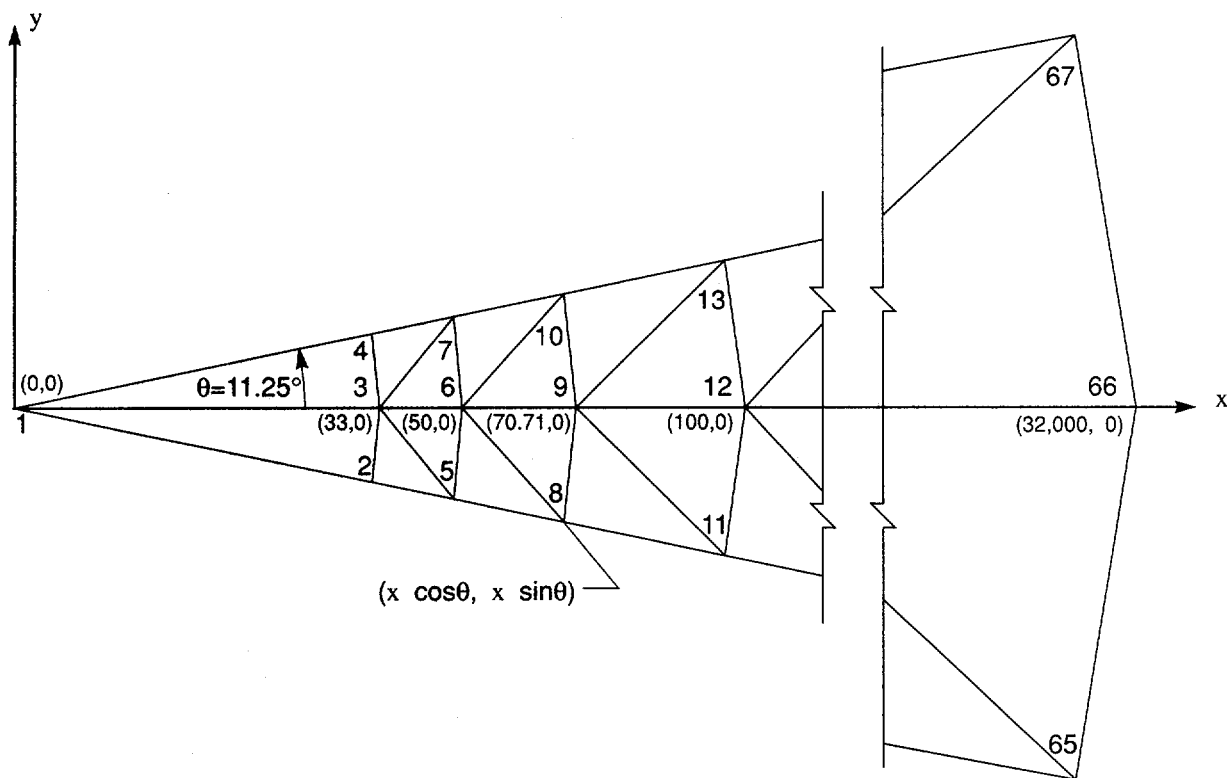
Figure 21. Finite-element mesh used to simulate the effects of transient leakage on drawdown near a pumped well.

from the pumped well by the factor $\sqrt{2}$ starting at 50 feet (figure 21). The (x,y) coordinates of nodes that are offset from the x axis by 11.25 degrees (figure 21) are computed from the x coordinate of the nodes located on the x axis as (x cos θ, x sin θ), where $\theta = \pm 11.25$ degrees.

Hydraulic heads are specified at nodes 65, 66, and 67 along the external model boundary that is 32,000 feet from the pumped well. This boundary is beyond the radius of influence of the pumped well during the simulation period. Because the flow system exhibits radial symmetry, element sides that are oriented in the radial direction from the well represent flow lines; hence, there is no flow across these element sides. Other characteristics of the problem are

$$T = 10^5 \text{ ft}^2/\text{d},$$
$$S = 1.25 \times 10^{-4},$$
$$Q = 1,256,637 \text{ ft}^3/\text{d},$$
$$h(x,y,0) = 0 \text{ ft},$$
$$h(r = 32,000 \text{ ft},t) = 0 \text{ ft}$$

for the aquifer and

$$K' = 10 \text{ ft/d},$$
$$b' = 400 \text{ ft},$$
$$S' = 0.008$$

for the confining unit. The head above the confining unit is held constant at 0 feet for the simulation period.

Pumpage is simulated for 0.10417 day (about 15 minutes) using 87 time elements. An initial time-element size of 2×10^{-8} day was selected, and the other time-element sizes were generated by multiplying previous values by factors ranging from 1.0 to 1.5.

Computed values of $4\pi Ts/Q$ versus $1/u$ at distances of 100, 300, 500, and 2,000 feet from the pumped well are compared with the type curves $H(u,\beta')$ versus $1/u$ using β' values of 0.1, 0.3, 0.5, and 2, respectively, in figure 22. The Theis solution plotted on this figure indicates the extent to which transient leakage affects drawdown. The numerical results show good agreement with the analytical solution.

MOENCH AND PRICKETT SOLUTION FOR CONVERSION FROM CONFINED TO UNCONFINED FLOW NEAR A PUMPED WELL

An analytical solution of Moench and Prickett (1972) is used to test the accuracy of MODFE for the problem of drawdown in an aquifer that converts from confined to unconfined conditions. A fully penetrating well of negligible diameter placed in a nonleaky, confined, homogeneous, and isotropic aquifer that is infinite in areal extent pumps at a constant rate Q sufficient to partially dewater the aquifer near the well (figure 23). Ground-water flow is assumed to be horizontal and obeys the Dupuit assumptions (Bear, 1979, p. 74-78) in the unconfined part of the aquifer. Changes in aquifer thickness, b, with drawdown in the unconfined part of the aquifer are assumed to be small and do not cause significant changes in transmissivity.
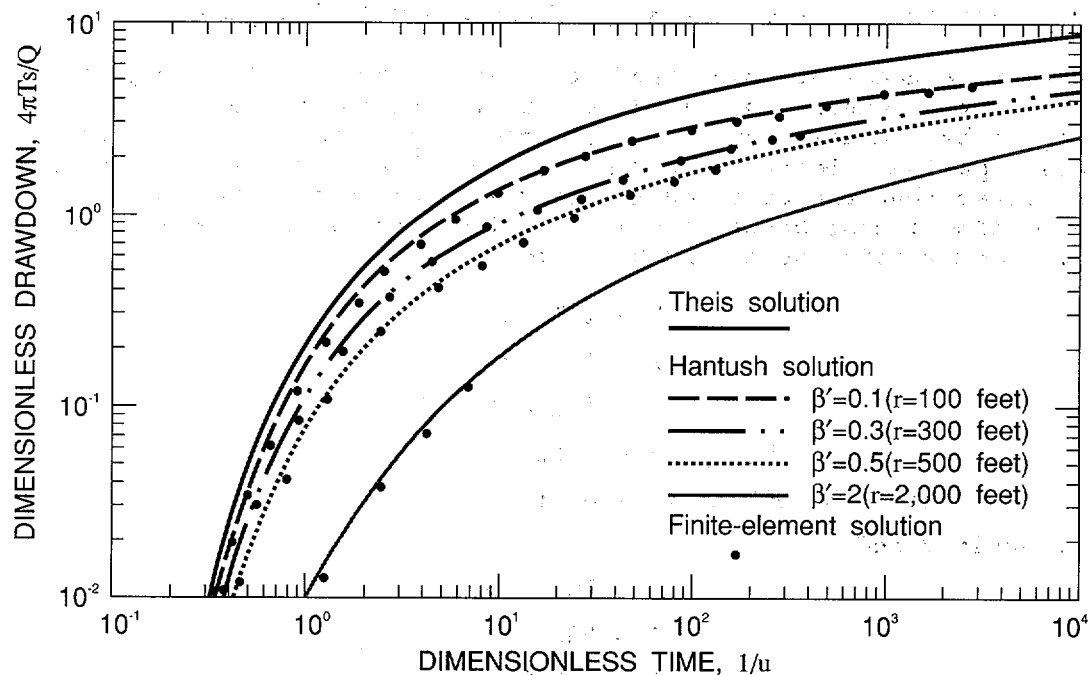
Figure 22. Hantush (1960) solution and finite-element results for the effects of transient leakage on drawdown near a pumped well.
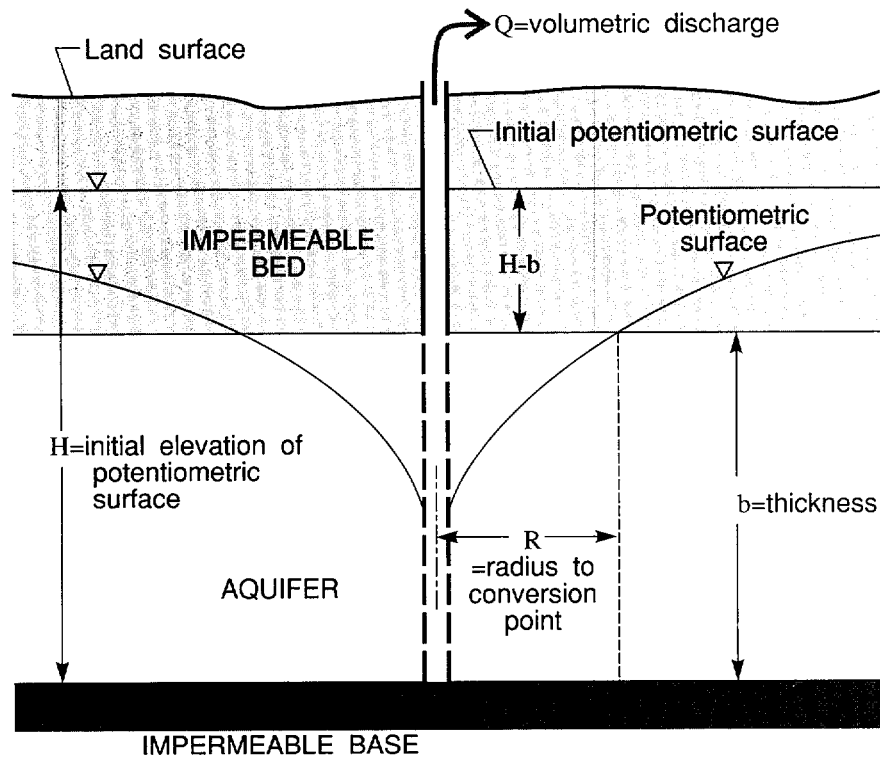
Figure 23. Geometry used to simulate the effects of conversion from confined to unconfined flow near a pumped well.

Solutions for total drawdown in the unconfined part of the aquifer, s_1, and in the confined part, s_2, are given by

$$s_1 = \frac{Q}{4\pi T} \left[W(u_1,v) + \frac{4\pi T(H-b)}{Q} \right],$$

and

$$s_2 = \frac{Q}{4\pi T} \left\{ e^{v[(\alpha_1/\alpha_2)-1]} W(u_2) \right\},$$

where

$$u_1 = \frac{r^2 S_y}{4Tt},$$

$$u_2 = \frac{r^2 S}{4Tt},$$

$$v = \frac{R^2 S_y}{4Tt},$$

S_y is the specific yield, S is the storage coefficient, α_1/α_2 [0] is the aquifer-diffusivity ratio $(T/S_y)/(T/S)$, or S/S_y, and R is the radial distance from the pumped well to where conversion takes place. $W(u_2)$ is the well function used for the Theis solution and $W(u_1,v) = W(u_1)-W(v)$.

The aquifer problem is simulated using Cartesian coordinates, and the finite-element mesh is the same as used by Wilson and others (1979, p. 95-101) for a similar test problem, except that their mesh terminated at

88

r = 8,000 feet whereas the mesh used here extends to r = 32,000 feet. The 22.5-degree wedge of the aquifer region is subdivided into 68 triangular elements and 52 nodes (figure 24) such that the node spacing expands in the radial direction by a factor of $\sqrt{2}$ starting at 125 feet from the well. Time elements range in size from the initial value of 5×10^{-5} days to a final value of 30 days and are expanded by factors of 1.0 for the first four elements to approximately 1.5 afterward; 44 time elements were used. Other characteristics of the problem are

$$K = 26.73 \text{ ft/d,}$$

$$b = 100 \text{ ft,}$$

$$S_y = 0.1,$$

$$S = 0.0001,$$

$$Q = 33,591 \text{ ft}^3\text{/d,}$$

$$h(x,y,0) = 0 \text{ ft,}$$

$$H = 0 \text{ ft.}$$

A Cauchy-type boundary is placed along the element sides at 32,000 feet. Because the influence of the pumped well on the aquifer extends beyond this radial distance and the analytical solution assumes that the aquifer has an infinite areal extent, the Cauchy-type boundary is used to simulate the part of the aquifer that is influenced by the pumped well but is not represented by the finite-element mesh. It allows flow across the artificial model boundary from the aquifer region that is external to the mesh, and allows drawdowns to be computed at the model boundary.
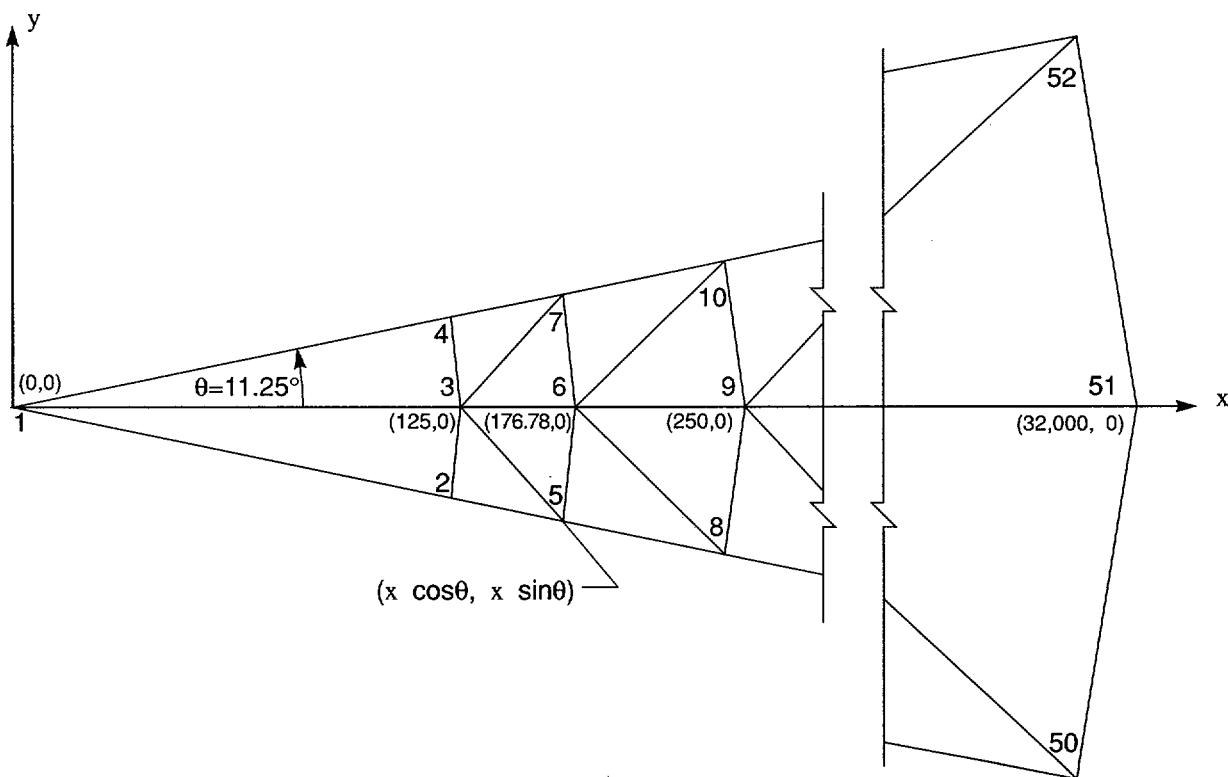

Figure 24. Finite-element mesh used to simulate the effects of conversion from confined to unconfined flow near a pumped well.
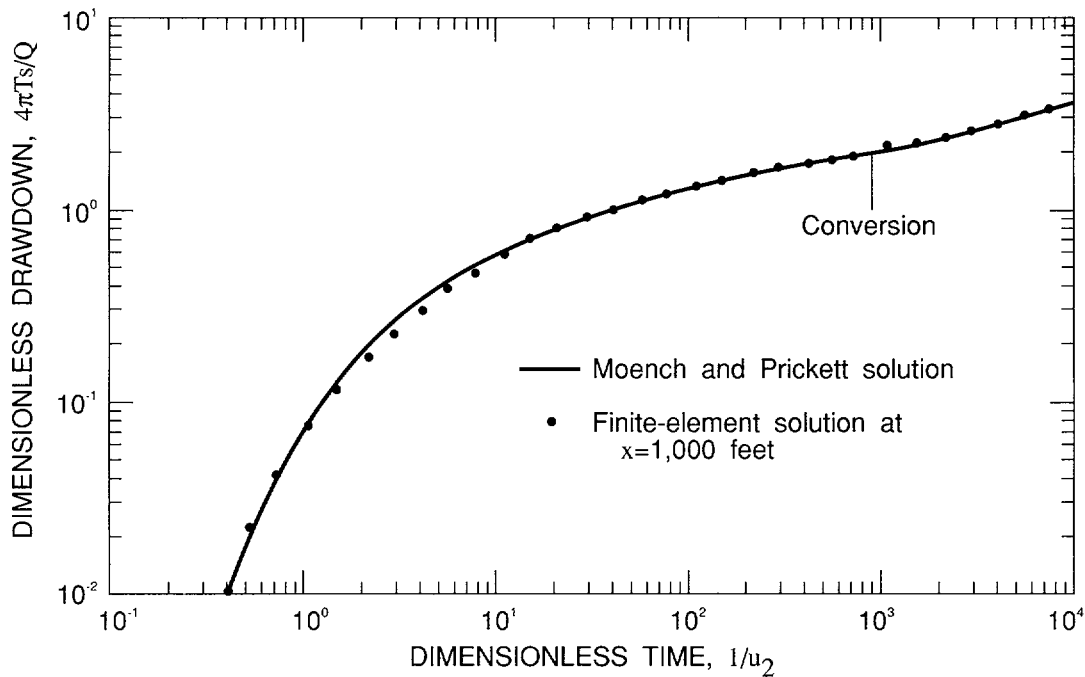
Figure 25. Moench and Prickett (1972) solution and finite-element results for conversion from confined to unconfined flow near a pumped well.

The specified head H_B (equation (4)) for the Cauchy-type boundary was located 200,000 feet from the pumped well. It is assumed that all drawdown in the infinite aquifer occurs within this distance. The coefficient α was obtained by assuming that flow beyond 32,000 feet is governed by the steady-state flow equation with known-head boundary conditions of $\hat{h}(t)$ at $r = 32,000$ feet and $H_B = 0$ at $r = 200,000$ feet. Therefore, by using the appropriate solution to the steady-state flow equation (Bear, 1979, p. 306, equation (8-7)) and equation (4),

$$q_n = T \left.\frac{\partial h}{\partial r}\right|_{r = 32,000} = \frac{T\left[\hat{h} - H_B\right]}{\ln\left[\frac{32,000}{200,000}\right]} \frac{1}{32,000}$$

$$= \alpha\left[H_B - \hat{h}\right],$$

so that, because $T = 2,673$ ft^2/d, $\alpha = 0.04558$ ft/d.

Computed drawdowns at a radial distance of 1,000 feet from the pumped well were compared with the analytical solution. Values of dimensionless drawdown, $4\pi Ts/Q$, and dimensionless time, $1/u_2$, were computed from the simulation results and are plotted in figure 25 along with the type curves of the analytical solution. Values for $e^{v[(\alpha_1/\alpha_2)-1]}W(u_2)$ versus $1/u_2$ were plotted for drawdowns less than 2 feet (before conversion), and values of $\left[W(u_1,v) + 2\right]$ versus $1/u_2$ were plotted for drawdowns greater than 2 feet (after conversion). The numerical results are in good agreement with the analytical solution, and are better than the results of Wilson and others (1979, p. 99) because they specified $H_B = 0$ at $r = 8,000$ feet, which did not allow drawdown to propagate beyond 8,000 feet as it should have.

90
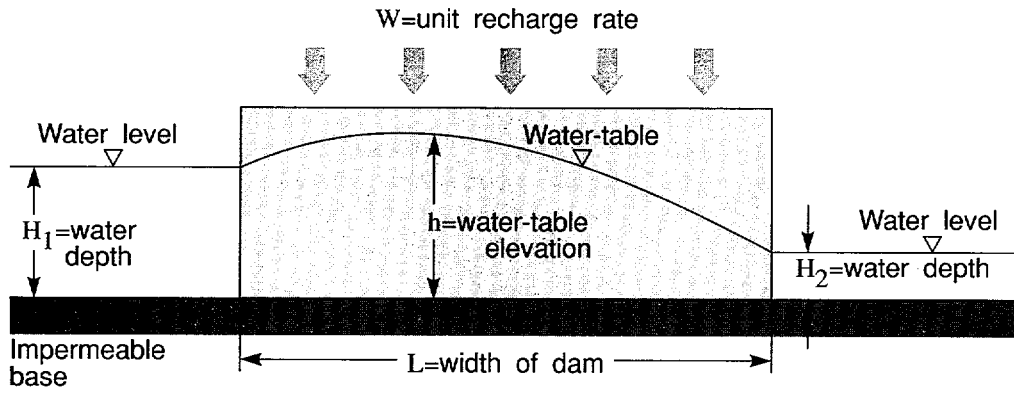
Figure 26. Cross section of steady-state flow through a dam with areal recharge.

STEADY-STATE FLOW THROUGH A DAM WITH AREAL RECHARGE

A straight dam with vertical faces 50 meters wide and 100 meters long maintains a water level of 8 meters on one side and 2 meters on the other side (figure 26). The hydraulic conductivity, K, of the earth material in the dam is 10^{-6} m/s and areal recharge, W, is applied to the surface of the dam at the rate of 4.8×10^{-8} m/s (figure 27). By making the Dupuit assumptions for unconfined flow, the solution for the height, h, of the water table in the dam can be obtained as the Dupuit parabola (see Verruijt, 1970, p. 51-57),

$$h^2 = H_1^2 - \left[H_1^2 - H_2^2\right]\frac{x}{L} + \frac{W}{K}x(L - x),$$

where x is the horizontal distance along the width L of the dam and the water levels H_1 and H_2 on either side of the dam are

$$H_1 = 8 \text{ meters}, \quad x = 0 \text{ meters},$$

$$H_2 = 2 \text{ meters}, \quad x = L = 50 \text{ meters}.$$
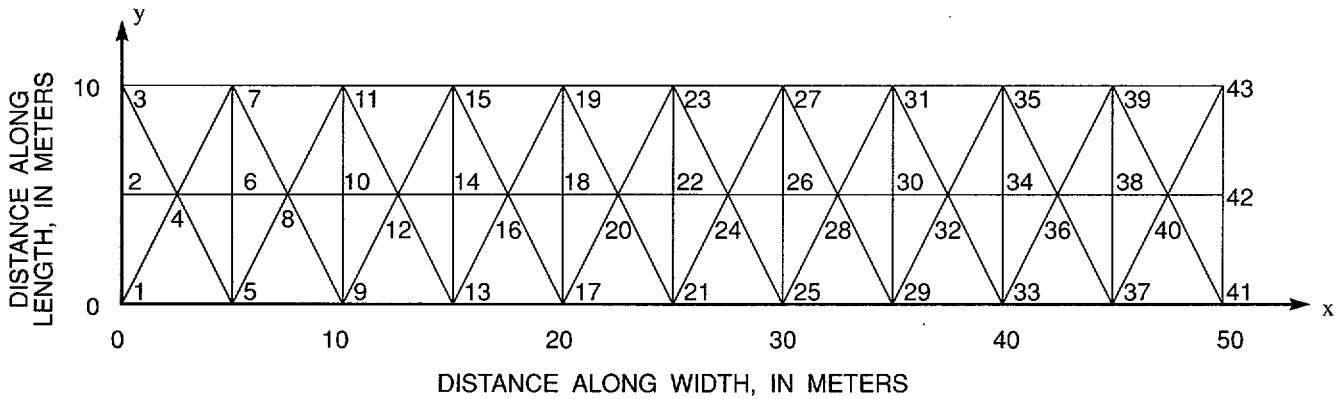
Figure 27. Finite-element mesh used to simulate steady-state flow through a dam with areal recharge.

Because ground-water flow is virtually one-dimensional through the dam, the entire 100-meter length need not be simulated. Instead, a 10-meter-long section of the dam is represented by a finite-element mesh consisting of 60 triangular elements and 43 nodes (figure 27). No-flow boundaries are placed along the element sides that are parallel to the x axis at y = 0 meters and y = 10 meters, because ground-water flow is parallel to these boundaries. Specified-head boundaries are located along the lines x = 0 meters and x = 50 meters in order to maintain the height of the water levels at the values given for H_1 and H_2, respectively. Areally distributed recharge is applied over the entire model area.

The steady-state solution for the water-table height provided by MODFE is plotted along with the analytical solution in figure 28. The solution by MODFE is in close agreement with the analytical solution.
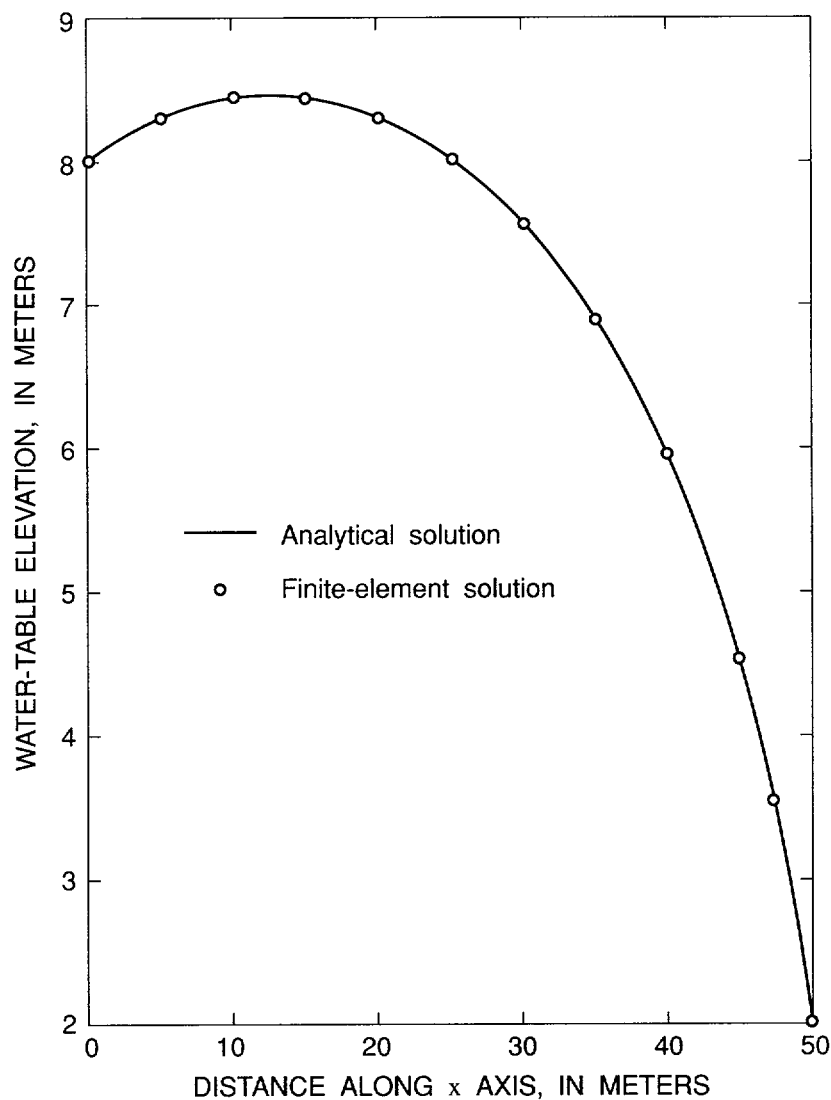
Figure 28. Analytical solution (Dupuit parabola) and finite-element results for steady-state flow through a dam with areal recharge.
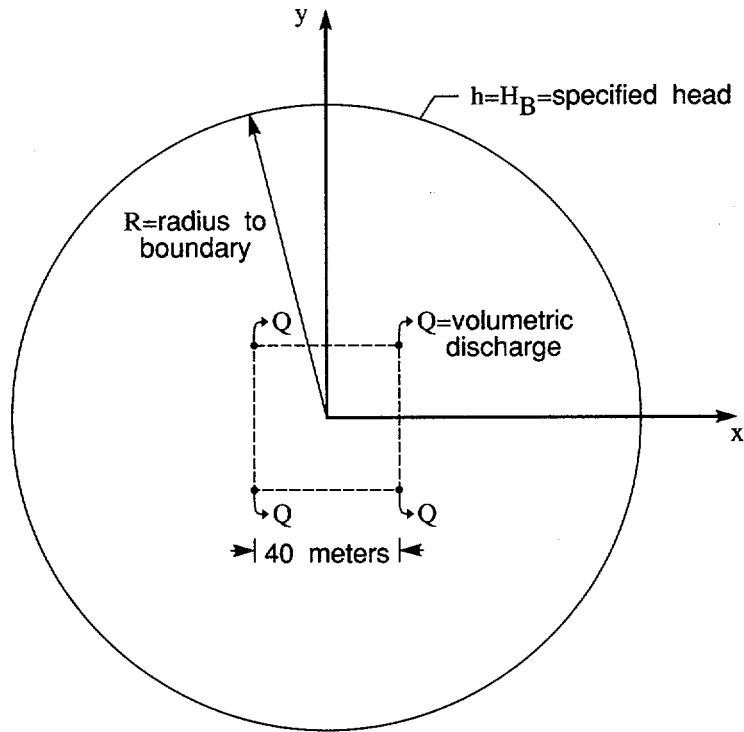
Figure 29. Geometry for two-dimensional steady-state flow in an unconfined aquifer.

TWO-DIMENSIONAL STEADY-STATE FLOW IN AN UNCONFINED AQUIFER

The flow problems described previously for testing the accuracy of MODFE have one-dimensional solutions, even though MODFE solved these problems in two dimensions. This flow problem tests the ability of MODFE to accurately compute steady-state water levels in an unconfined aquifer for a problem that has an analytical solution in two dimensions.

The problem used in this simulation is taken from Verruijt (1970, Problem 6.2) where four wells pumped at the same rate (1.196×10^{-6} m³/s) are located at the corners of a 40-meter square in an unconfined aquifer (figure 29). The square represents a construction site where the water level is to be maintained 4 meters below the original water table, which is 10 meters above the impermeable base. The aquifer is homogeneous and isotropic with a hydraulic conductivity, K, of 10^{-7} m/s, an initial thickness of 10 meters, and an external radius, R, of 2,000 meters, measured from the center of the square. Beyond R the drawdown is zero.

The solution for the aquifer head, h, is given by Verruijt (1970, p. 66) as

$$ h = \left[H_B^2 + \frac{1}{\pi K} \sum_{j=1}^{4} Q_j \ln \frac{\sqrt{(x - x_j)^2 + (y - y_j)^2}}{R} \right]^{\frac{1}{2}} $$

where H_B is the specified head (10 meters) at the radius R and Q_j is the volumetric discharge from well j located at the point $\left(x_j, y_j\right)$ in the aquifer.

93

Because the wells are pumped at the same rate and are regularly spaced about the center of the square (see figure 29), the flow problem can be simulated in the 45-degree wedge shown in figure 30. The origin of the wedge is the center of the square, and one of the four pumped wells is placed at the point (20 meters, 20 meters) in the wedge. No-flow boundaries are located along the x axis and the line y = x, and a specified-head boundary (H_B = 10 meters) is located at a distance of 2,000 meters from the center of the square.

The aquifer region is subdivided by a finite-element mesh consisting of 94 triangular elements and 67 nodes (figure 30). Results from the nonlinear steady-state simulation and the analytical solution are presented in figure 31. The simulated results show good agreement with the analytical solution.

DISTANCE ALONG x AXIS, IN METERS
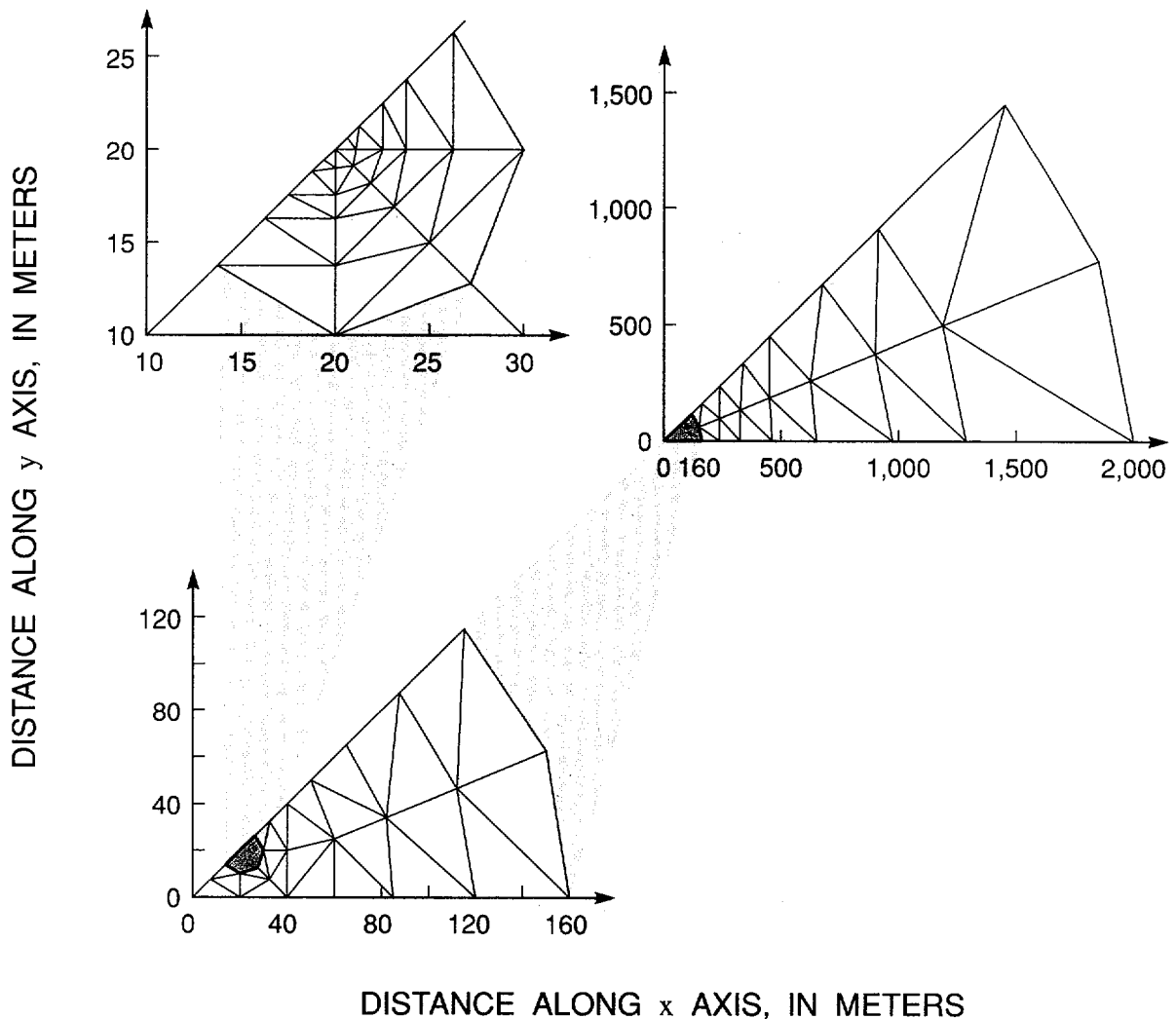
Figure 30. Finite-element mesh used to simulate two-dimensional steady-state flow in an unconfined aquifer.

Figure 31. Analytical solution and finite-element results for
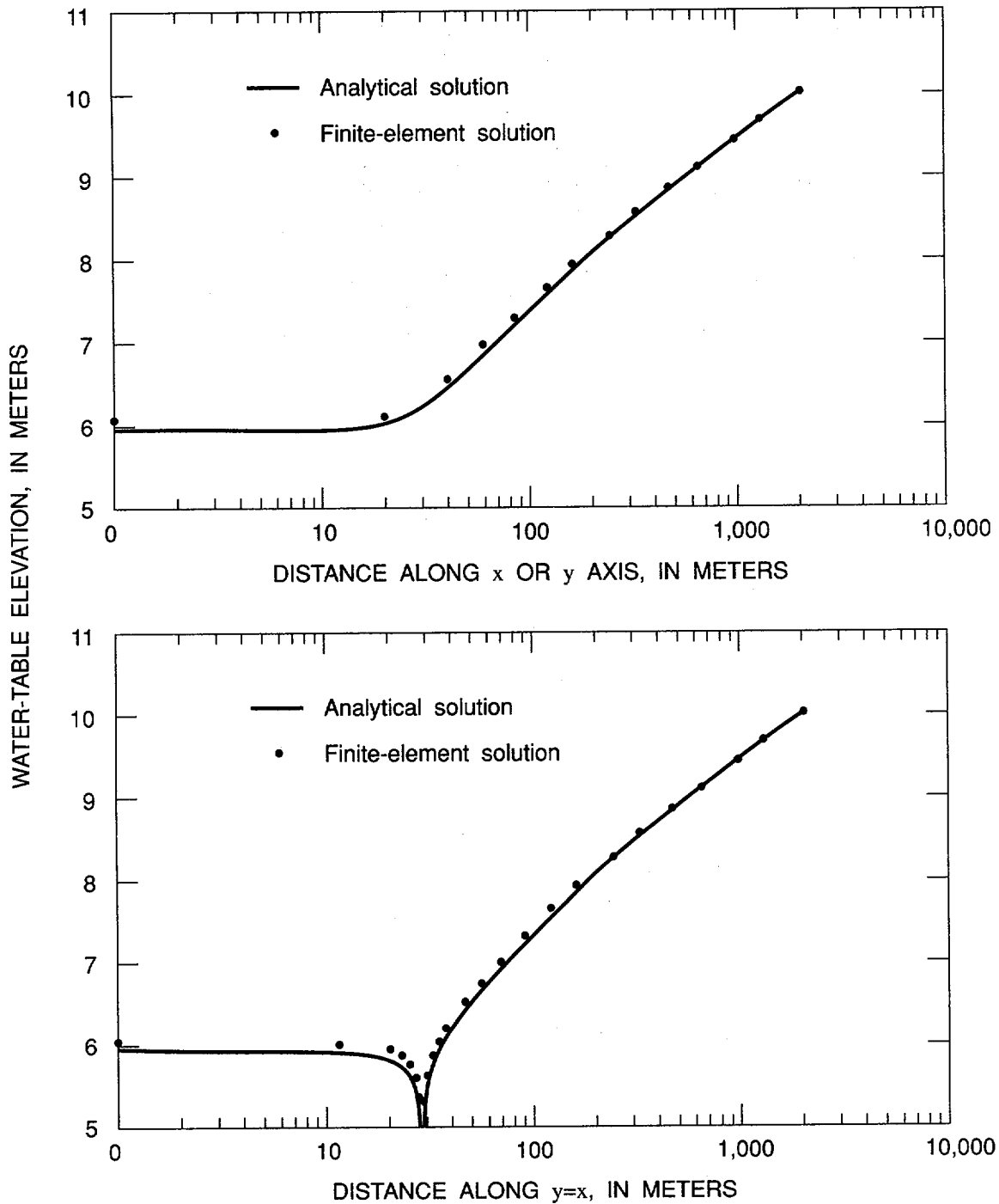two-dimensional steady-state flow in an unconfined aquifer.

SUMMARY

The two-dimensional steady- and unsteady-state equations for ground-
water flow in a heterogeneous, anisotropic aquifer were approximately solved
with finite-element techniques. Spatial finite elements are triangular with
two-dimensional linear basis functions and time elements are linear, with

95

one-dimensional linear basis functions. Finite-element equations were derived by minimizing a functional of the difference between true and approximate hydraulic heads and are equivalent to finite-element equations obtained by either classical variational or Galerkin methods. Variable directions of anisotropy are incorporated by rotating the coordinate system locally so that the rotated coordinates are aligned with the local principal directions of the transmissivity tensor. For unsteady-state problems, a mass balance is computed at the end of each time element. Computed flow components include accumulation or depletion of water in storage, flow across confining beds, flow across specified-head boundaries, flow across specified-flow boundaries, and flow across head-dependent flow boundaries. For steady-state problems, the mass balance, excluding the storage component, is computed at the end of the simulation.

The basic finite-element equations include the following processes: confined flow; leakage through rigid confining beds; specified areal and point recharge and discharge; and specified-flow, specified-head, or head-dependent boundary conditions. Extensions of these equations allow for unconfined flow using the Dupuit assumption, decreases of aquifer thickness to zero and increases from zero (termed drying of nodes), conversions from confined to unconfined flow and vice versa, point head-dependent discharge from springs and drainage wells, areal head-dependent leakage combined with aquifer dewatering, areal head-dependent discharge from evapotranspiration, line head-dependent leakage combined with aquifer dewatering for narrow rivers, and transient leakage from confining beds. Except for transient leakage, all of these extensions are nonlinear.

The finite-element equations were also formulated using axisymmetric cylindrical coordinates to allow analysis of problems involving axisymmetric flow in multiaquifer systems. Boundary conditions are the same as for the two-dimensional Cartesian versions, but for radial flow the principal directions of the hydraulic conductivity tensor are assumed to be the radial and vertical directions. None of the extensions can be used in this case.

Matrix solution techniques for the finite-element equations include the direct symmetric-Doolittle method, which can be efficient for small to medium problems (less than about 500 nodes), and the iterative modified incomplete-Cholesky conjugate-gradient (MICCG) method, which is more efficient for larger problems (more than about 500 nodes). Nonlinear unsteady-state problems are solved using a predictor-corrector method that can employ either the direct or MICCG method to solve both the predictor and corrector equations. Nonlinear steady-state problems are solved using an iterative method that can also employ either matrix solution method. Use of MICCG for nonlinear steady-state problems yields an inner MICCG iteration loop and an outer iteration loop on the nonlinearity. Because the inner loop converges in progressively fewer iterations as the outer loop converges, MICCG is recommended for these problems.

The accuracy of the finite-element solution method was evaluated using five test problems for which analytical solutions are available: radial flow to a well in a homogeneous, infinite, nonleaky, confined aquifer; radial flow to a well in a homogeneous, infinite, confined aquifer with transient leakage; radial flow to a well in a homogeneous, infinite, nonleaky aquifer undergoing conversion from confined to unconfined flow; one-dimensional, unconfined, steady-state flow through a dam with areal recharge; and two-dimensional, unconfined, steady-state flow to a group of drainage wells. All problems except the first were solved using Cartesian coordinates. All numerical solutions are in good agreement with the analytical solutions.

96

REFERENCES CITED

Bear, Jacob, 1979, Hydraulics of groundwater: New York, McGraw-Hill, 569 p.

Beckman, F.S., 1967, The solution of linear systems by the conjugate gradient method, in Ralston, Anthony, and Wilf, H.S., Mathematical methods for digital computers, v. 1: New York, John Wiley, p. 62-72.

Bettencourt, J.M., Zienkiewicz, O.C., and Cantin, G., 1981, Consistent use of finite elements in time and the performance of various recurrence schemes for the heat diffusion equation: International Journal for Numerical Methods in Engineering, v. 17, p. 931-938.

Briggs, J.E., and Dixon, T.N., 1968, Some practical considerations in the numerical solution of two-dimensional reservoir problems: Society of Petroleum Engineers Journal, v. 8, no. 2, p. 185-194.

Carslaw, H.S., and Jaeger, J.C., 1959, Conduction of heat in solids: Oxford, Great Britain, Clarendon, 510 p.

Concus, Paul, Golub, G.H., and O'Leary, D.P., 1975, A generalized conjugate gradient method for the numerical solution of elliptic partial differential equations: Berkeley, Calif., Lawrence Berkeley Publication LBL-4604, 24 p.

Cooley, R.L., 1971, A finite difference method for unsteady flow in variably saturated porous media: Application to a single pumping well: Water Resources Research, v. 7, no. 6, p. 1607-1625.

----- 1972, Numerical simulation of flow in an aquifer overlain by a water-table aquitard: Water Resources Research, v. 8, no. 4, p. 1046-1050.

----- 1974, Finite element solutions for the equations of ground-water flow: Center for Water Resources Research, Desert Research Institute, University of Nevada System, Reno, Technical Report Series H-W, Hydrology and Water Resources Publication No. 18, 134 p.

----- 1983, Some new procedures for numerical solution of variably saturated flow problems: Water Resources Research, v. 19, no 5, p. 1271-1285.

Cooley, R.L., and Naff, R.L., 1990, Regression modeling of ground-water flow: U.S. Geological Survey Techniques of Water-Resources Investigations, Book 3, Chapter B4, 232 p.

Crank, J., and Nicolson, P., 1947, A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type: Proceedings of the Cambridge Philosophical Society, v. 43, p. 50-67.

Desai, C.S., and Abel, J.F., 1972, Introduction to the finite element method: New York, Van Nostrand Reinhold, 477 p.

Douglas, Jim, Jr., and Jones, B.F., 1963, On predictor-corrector methods for nonlinear parabolic differential equations: Journal for the Society of Industrial and Applied Mathematics, v. 11, p. 195-204.

Fox, Leslie, 1965, An introduction to numerical linear algebra: New York, Oxford, 327 p.

Gambolati, Giuseppe, and Volpi, Giampiero, 1982, Analysis of performance of the modified conjugate gradient method for solution of sparse linear sets of finite element equations, in Gallagher, R.H., Norrie, D.H., Oden, J.T., and Zienkiewicz, O.C., eds., Finite Elements in Fluids, v. 4: 3rd International Conference on Finite Elements in Flow Problems held at Banff, Alberta, Canada, 10-13 June, 1980, John Wiley, p. 256-264.

Gresho, P.M., Lee, R.L., and Sani, R.L., 1976, Advection-dominated flows, with emphasis on the consequences of mass lumping: International Centre for Computer Aided Design, Second International Symposium on Finite Element Methods in Flow Problems held at Santa Margherita Ligure, Italy, June 14-June 18, 1976, p. 745-756.

Gustafsson, Ivar, 1978, A class of first order factorization methods: BIT, v. 18, no. 2, p. 142-156.

----- 1979, On modified incomplete Cholesky factorization methods for the solution of problems with mixed boundary conditions and problems with discontinuous material coefficients: International Journal for Numerical Methods in Engineering, v. 14, p. 1127-1140.

Hantush, M.S., 1960, Modification of the theory of leaky aquifers: Journal of Geophysical Research, v. 65, no. 11, p. 3713-3725.

Herrera, Ismael, 1970, Theory of multiple leaky aquifers: Water Resources Research, v. 6, no. 1, p. 185-193.

----- 1974, Integrodifferential equations for systems of leaky aquifers and applications, 2, Error analysis of approximate theories: Water Resources Research, v. 10, no. 4, p. 811-820.

Herrera, Ismael, and Rodarte, Leopoldo, 1973, Integrodifferential equations for systems of leaky aquifers and applications, 1, The nature of approximate theories: Water Resources Research, v. 9, no. 4, p. 995-1005.

Herrera, Ismael, and Yates, Robert, 1977, Integrodifferential equations for systems of leaky aquifers and applications, 3, A numerical method of unlimited applicability: Water Resources Research, v. 13, no. 4, p. 725-732.

Hohn, F.E., 1964, Elementary matrix algebra (2d ed.): New York, Macmillan, 395 p.

Lohman, S.W., 1972, Ground-water hydraulics: U.S. Geological Survey Professional Paper 708, 70 p.

Manteuffel, T.A., 1980, An incomplete factorization technique for positive definite linear systems: Mathematics of Computation, v. 34, no. 150, p. 473-497.

McCracken, D.D., and Dorn, W.S., 1964, Numerical methods and FORTRAN programming: New York, John Wiley, 457 p.

Meijerink, J.A., and van der Vorst, H.A., 1977, An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix: Mathematics of Computation, v. 31, no. 137, p. 148-162.

Moench, A.F., and Prickett, T.A., 1972, Radial flow in an infinite aquifer undergoing conversion from artesian to water table conditions: Water Resources Research, v. 8, no. 2, p. 494-499.

Narasimhan, T.N., Neuman, S.P., and Witherspoon, P.A., 1978, Finite element method for subsurface hydrology using a mixed explicit-implicit scheme: Water Resources Research, v. 14, no. 5, p. 863-877.

Neuman, S.P., and Witherspoon, P.A., 1969, Theory of flow in a confined two-aquifer system: Water Resources Rsearch, v. 5, no. 4, p. 803-816.

Norrie, D.H., and de Vries, Gerard, 1973, The finite element method: New York, Academic, 322 p.

Pinder, G.F., and Gray, W.G., 1977, Finite element simulation in surface and subsurface hydrology: New York, Academic, 295 p.

Prickett, T.A., and Lonnquist, L.G., 1971, Selected digital computer techniques for groundwater resource evaluation: Illinois State Water Survey, Bulletin 55, 62 p.

Remson, Irwin, Hornberger, G.M., and Molz, F.J., 1971, Numerical methods in subsurface hydrology: New York, Wiley-Interscience, 389 p.

Segerlind, L.J., 1976, Applied finite element analysis: New York, John Wiley, 422 p.

Smith, G.D., 1965, Numerical solution of partial differential equations: New York, Oxford, 179 p.

Spiegel, M.R., 1959, Theory and problems of vector analysis: New York, Schaum, 225 p.

Torak, L.J., 1992a, A MODular Finite-Element model (MODFE) for areal and axisymmetric ground-water flow problems, part 1: Model description and user's manual: U.S. Geological Survey Open-File Report 90-194.

----- 1992b, A MODular Finite-Element model (MODFE) for areal and axisymmetric ground-water flow problems, part 3: Design philosophy and programming details: U.S. Geological Survey Open-File Report 91-471.

Trescott, P.C., Pinder, G.F., and Larson, S.P., 1976, Finite-difference model for aquifer simulation in two dimensions with results of numerical experiments: U.S. Geological Survey Techniques of Water-Resources Investigations, Book 7, Chapter C1, 116 p.

Varga, R.S., 1962, Matrix iterative analysis: Englewood Cliffs, N.J., Prentice-Hall, 322 p.

Verruijt, A., 1970, Theory of groundwater flow: New York, Gordon and Breach, 190 p.

Wang, H.F., and Anderson, M.P., 1982, Introduction to groundwater modeling; Finite difference and finite element methods: San Francisco, W.H. Freeman, 237 p.

Wilson, J.L., Townley, L.R., and Sa da Costa, A., 1979, Mathematical development and verification of a finite element aquifer flow model AQUIFEM-1: Massachusetts Institute of Technology, Technological Planning Program, TAP Report 79-2, 114 p.

Wong, Y.S., 1979, Pre-conditioned conjugate gradient methods for large sparse matrix problems, in Lewis, R.W., and Morgan, K., eds., Numerical methods in thermal problems: Proceedings of the First International Conference held at University College, Swansea, England, 2-6 July, 1979, Pineridge Press, p. 967-979.

Zienkiewicz, O.C., 1971, The finite element method in engineering science: London, McGraw-Hill, 521 p.

Zienkiewicz, O.C., Mayer, Paul, and Cheung, Y.K., 1966, Solution of anisotropic seepage by finite elements: Journal of the Engineering Mechanics Division, Proceedings of the American Society of Civil Engineers, v. 92, no. EM1, p. 111-120.

APPENDIX A

It is shown here that

$$\sum_{e_i} \int_0^{\Delta t_{n+1}} \sigma_{n+1} \left\{ \int\int_{\Delta^e} \left[N_i^e \left(S\frac{\partial h}{\partial t} - R(H - h) - W - P \right) + \frac{\partial N_i^e}{\partial x} \left(T_{xx}\frac{\partial h}{\partial x} + T_{xy}\frac{\partial h}{\partial y} \right) \right. \right.$$

$$\left. \left. + \frac{\partial N_i^e}{\partial y} \left(T_{yx}\frac{\partial h}{\partial x} + T_{yy}\frac{\partial h}{\partial y} \right) \right] dxdy - \int_{C_2^e} N_i^e \left[q_B + \alpha\left(H_B - h\right) \right] dC \right\} dt' = 0. \quad \text{(A1)}$$

To simplify notation, the generalized Darcy's law (Bear, 1979, p. 71) is used to replace the terms involving transmissivity. That is,

$$q_x = -T_{xx}\frac{\partial h}{\partial x} - T_{xy}\frac{\partial h}{\partial y} \quad \text{(A2)}$$

and

$$q_y = -T_{yx}\frac{vh}{\partial x} - T_{yy}\frac{vh}{\partial y} \quad \text{(A3)}$$

are used to write equation (A1) as

$$\sum_{e_i} \int_0^{\Delta t_{n+1}} \sigma_{n+1} \left\{ \int\int_{\Delta^e} \left[N_i^e \left(S\frac{\partial h}{\partial t} - R(H - h) - W - P \right) - \frac{\partial N_i^e}{\partial x}q_x - \frac{\partial N_i^e}{\partial y}q_y \right] dxdy \right.$$

$$\left. - \int_{C_2^e} N_i^e \left[q_B + \alpha\left(H_B - h\right) \right] dC \right\} dt' = 0. \quad \text{(A4)}$$

To initiate the proof, equations (A2) and (A3) are substituted into equation (1), which is then multiplied by $\sigma_{n+1}N_i^e$ and integrated over spatial element e and time element n+1 to obtain

$$\int_0^{\Delta t_{n+1}} \sigma_{n+1} \left\{ \int\int_{\Delta^e} N_i^e \left(S\frac{\partial h}{\partial t} - R(H - h) - W - P \right) dxdy \right\} dt'$$

$$= - \int_0^{\Delta t_{n+1}} \sigma_{n+1} \left\{ \int\int_{\Delta^e} N_i^e \left(\frac{\partial q_x}{\partial x} + \frac{\partial q_y}{\partial y} \right) dxdy \right\} dt'. \quad \text{(A5)}$$

Next, a result from vector calculus known as Green's first identity (Spiegel, 1959, p. 107) is used to modify the right side of equation (A5).

If q_x, q_y, and N_i^e are continuous and have continuous derivatives in element e, then

$$\int_{\Delta^e}\int N_i^e\left(\frac{\partial q_x}{\partial x} + \frac{\partial q_y}{\partial y}\right)dxdy = -\int_{\Delta^e}\int\left(\frac{\partial N_i^e}{\partial x}q_x + \frac{\partial N_i^e}{\partial y}q_y\right)dxdy - \int_{C^e}N_i^e q_n dC, \qquad (A6)$$

where q_n is the component of the flow vector $\left(q_x, q_y\right)$ that is normal to the element boundary and is positive for inflow, and C^e is the boundary of element e. Substitution of equation (A6) into equation (A5) and rearrangement yields

$$\int_0^{\Delta t_{n+1}}\sigma_{n+1}\left\{\int_{\Delta^e}\int\left[N_i^e\left(S\frac{\partial h}{\partial t} - R(H - h) - W - P\right) - \frac{\partial N_i^e}{\partial x}q_x - \frac{\partial N_i^e}{\partial y}q_y\right]dxdy\right.$$
$$\left. - \int_{C^e}N_i^e q_n dC\right\}dt' = 0. \qquad (A7)$$

The integral over the element boundary C^e can be split into two integrals: the integral over a Cauchy-type boundary and the integral over the remaining side(s), so that

$$\int_{C^e}N_i^e q_n dC = \int_{C_1^e}N_i^e q_n dC + \int_{C_2^e}N_i^e\left[q_B + \alpha\left(H_B - h\right)\right]dC, \qquad (A8)$$

where C_1^e designates the side(s) that are not Cauchy-type boundaries and equation (4) was used to replace q_n in the integral over a Cauchy-type boundary.

When equation (A7) is summed over all elements in the patch for node i, all boundary integrals over C_1^e for adjacent element sides cancel because of equation (3) and the continuity of N_i^e across an element boundary. Furthermore, all boundary integrals for element sides forming the outer boundary of the patch are zero, because N_i^e is zero on these sides. Therefore, equation (A7) yields

$$\sum_{e_i}\int_0^{\Delta t_{n+1}}\sigma_{n+1}\left\{\int_{\Delta^e}\int\left[N_i^e\left(S\frac{\partial h}{\partial t} - R(H - h) - W - P\right) - \frac{\partial N_i^e}{\partial x}q_x - \frac{\partial N_i^e}{\partial y}q_y\right]dxdy\right.$$
$$\left. - \int_{C_2^e}N_i^e\left[q_B + \alpha\left(H_B - h\right)\right]dC\right\}dt' = 0, \qquad (A9)$$

which, when written using equations (A2) and (A3) to replace q_x and q_y, is equation (A1).

NOTATION

The principal notation used in this report is given below. Certain symbols used only locally are omitted from the list for brevity.

A^e, B^e, C^e
- Coefficients defined by equation (6) and used to compute the approximate solution, $\hat{h}$.

A'_m, B'_m
- Coefficients defined by equations (176) and (177) and used to approximate the infinite series for transient leakage calculations.

A_m, B_m
- $A_m = A'_m/\alpha_m$; $B_m = B'_m/\beta_m$.

(a'_j, b'_j)
- (x,y) location of the jth point source or sink. Overbars signify that the location is given in $(\bar{x}, \bar{y})$ coordinates.

a^e_i, b^e_i, c^e_i
- Coefficients used in basis functions N^e_i; defined by equation (10) for Cartesian coordinates and by equation (212) for axisymmetric cylindrical coordinates. Overbars signify evaluation using $(\bar{x}, \bar{y})$ coordinates.

$\underline{\underline{A}}$
- Matrix $\underline{\underline{G}} + \underline{\underline{V}}$ or the coefficient matrix defined by equation (254), depending on context. Entries are A_{ij} for $\underline{\underline{A}} = \underline{\underline{G}} + \underline{\underline{V}}$ and a_{ij} for $\underline{\underline{A}}$ as the coefficient matrix. Subscript ℓ ($\underline{\underline{A}}_\ell$) signifies evaluation of the coefficient matrix at iteration ℓ for a nonlinear steady-state problem.

$\underline{B}$
- Vector of known flows and boundary conditions, defined by equations (45) and (50) for Cartesian coordinates and by equations (45) and (229) for axisymmetric cylindrical coordinates. Entries are B_i. Subscript ℓ ($\underline{B}_\ell$) signifies evaluation at iteration ℓ for a nonlinear steady-state problem.

$\bar{\underline{B}}$
- Weighted average $\underline{B}$ over time element n+1 defined by equation (62). Entries are $\bar{B}_i$.

b
- Aquifer thickness; b_i is aquifer thickness at node i, and $b_{i,n}$ is aquifer thickness at node i and time level n.

b'
- Confining-unit thickness; b'_i is harmonic mean confining-unit thickness at node i defined by equation (170).

C^e_2
- Side(s) of element e that are Cauchy-type boundaries.

C_{ai}
- Coefficient for areal head-dependent leakage function that applies for aquifer dewatering at node i; defined by equation (119).

C_{ei}
- Coefficient for areal head-dependent discharge function at node i; defined by equation (132).

C_{pi}
- Coefficient for point head-dependent discharge function at node i; defined by equation (104).

C_{ri}
- Coefficient for line head-dependent leakage function at node i; defined by equation (155).

c^e_{ii}
- Storage coefficient term for node i of element e; defined by equation (36) for Cartesian coordinates and by equation (224) for axisymmetric cylindrical coordinates.

$c^{(1)}_{ii}, c^{(2)}_{ii}$
- Entry C_{ii} of $\underline{\underline{C}}$ for before (1) and after (2) conversion from confined to unconfined flow or vice versa.

$\underline{\underline{C}}$
- Diagonal matrix with diagonal entries defined by $C_{ii} = \sum_{e_i} c^e_{ii}$.

D_{ij}
- $\sum_{e_i} d^e_{ij}$.

$\underline{\underline{D}}$
- Diagonal matrix for symmetric-Doolittle factorization of $\underline{\underline{A}}$; defined by equation (260). Entries are $1/\alpha_{ii}$.

$\underset{=}{\tilde{D}}$ • Diagonal matrix for incomplete-Cholesky or modified incomplete-Cholesky factorization of $\underset{=}{A}$; defined by equation (272) and calculated using equation (273) for incomplete-Cholesky factorization or equation (279) for modified incomplete-Cholesky factorization. Entries are $1/\tilde{\alpha}_{ii}$.

d_{ij}^e • Hydraulic-conductivity term for unconfined flow; defined by equation (72).

$\underset{-}{d}$ • Right-hand side vector for finite-element matrix equation (254). Entries are d_i.

$\hat{e}$ • Error $h - \hat{h}$ in the approximate solution of equation (1).

e_i • Index indicating summation over elements in the patch for node i.

f_{ij} • Element of fill-in in $\underset{=}{M}$; defined by equation (277).

$\underset{=}{G}$ • Matrix defined by entries $G_{ij} = \sum_{e_i} g_{ij}^e$ for confined flow and by equation (74) for unconfined flow.

$\underset{=}{\tilde{G}}$ • Weighted average $\underset{=}{G}$ over time element n+1 defined by equation (69).

$\underset{=}{\bar{G}}$ • Weighted average $\underset{=}{G}$ over time element n+1 defined by equation (70).

$\underset{=}{\tilde{G}}{}^*$ • Matrix $\underset{=}{\tilde{G}}$ computed using predicted head vector $\underset{-}{\hat{h}}{}^*$.

$\underset{=}{\bar{G}}{}^*$ • Matrix $\underset{=}{\bar{G}}$ computed using predicted head vector $\underset{-}{\hat{h}}{}^*$.

g_{ij}^e • Transmissivity term defined by equations (38) (or (43)), (39), and (40) for Cartesian coordinates and by equations (226), (227), and (228) for axisymmetric cylindrical coordinates.

H • Hydraulic head at the distal side of a confining unit.

H_B • Specified head at a boundary.

H_o • Initial head (at t = 0).

H_a • Hydraulic head at the distal side of a confining unit or the stage elevation of wide river overlying an aquifer being dewatered.

H_r • Controlling head for line head-dependent leakage functions (for rivers it is the river-stage elevation).

h • True hydraulic head in the aquifer.

$\hat{h}$ • Approximate hydraulic head in the aquifer defined by equation (6); $\hat{h}_i$ is $\hat{h}$ at node i and $\hat{h}_{i,n}$ is $\hat{h}$ at node i and time level n.

$\hat{h}_i'$ • Predicted head at node i during a conversion from confined to unconfined flow.

$\underset{-}{\hat{h}}$ • Vector of entries $\hat{h}_i$; $\underset{-}{\hat{h}}_n$ is a vector of entries $\hat{h}_{i,n}$.

$\underset{-}{\bar{h}}$ • Weighted mean of $\hat{h}$ over time element n+1; defined by equation (63).

$\underset{-}{\hat{h}}{}^*$ • Predicted head vector for time level n+1 for predictor-corrector method.

$\underset{-}{h}_o$ • Arbitrary initial head vector for steady-state flow problems.

$\hat{I}(e)$ • Error functional defined by equation (15) for Cartesian coordinates and by equation (214) for axisymmetric cylindrical coordinates.

$I_{mi,n}$ • The mth term of the infinite series for transient leakage at time level n resulting from time variation of head in the aquifer at node i; defined by equation (172).

$\hat{I}_{mi,n}$ • The mth term of the finite series to approximate transient leakage at time level n resulting from time variation of head in the aquifer at node i; defined by equation (179).

105

$(\tilde{\imath},\tilde{\jmath})$	• Row and column location of nonzero entries in $\underline{\underline{A}}$.
$J_{mi,n}$	• The mth term of the infinite series for transient leakage at time level n resulting from time variation of head at the distal side of a confining unit at node i; defined by equation (175).
$\hat{J}_{mi,n}$	• The mth term of the finite series to approximate transient leakage at time level n resulting from time variation of head at the distal side of the confining unit at node i; defined by equation (181).
$\left(K_{xx},\ K_{xy},\right.$	• Components of the hydraulic conductivity tensor for the
$\left.K_{yx},\ K_{yy}\right)$	aquifer written using Cartesian coordinates (x,y). Principal components in the $(\bar{x},\bar{y})$ coordinate system are $\left(K_{\bar{x}\bar{x}},\ K_{\bar{y}\bar{y}}\right)$.
$\left(K_{rr},\ K_{zz}\right)$	• Principal components of the hydraulic conductivity tensor written using axisymmetric cylindrical coordinates (r,z).
K'_{zz}	• Vertical hydraulic conductivity in a confining unit; K'^{e}_{zz} is the constant value of K'_{zz} for spatial element e.
$L_{ij'}$	• Length of the side of an element between nodes i and j'.
$M_1\left(\Delta t_D\right)$	• Finite series approximation of $S_1\left(\Delta t_D\right)$ for transient leakage.
$M_2\left(\Delta t_D\right)$	• Finite series approximation of $S_2\left(\Delta t_D\right)$ for transient leakage.
$\underline{\underline{M}}$	• Preconditioning matrix that is an approximation of $\underline{\underline{A}}$ but is much easier to invert; defined by equation (266).
N	• Number of nodes in the finite-element mesh, or the number of unknowns in equation (254), depending on context.
$N_1,\ N_2$	• Number of terms in $M_1\left(\Delta t_D\right)$ and $M_2\left(\Delta t_D\right)$, respectively.
N^e_i	• Basis functions for spatial finite elements defined by equation (9) for Cartesian coordinates and equation (211) for axisymmetric cylindrical coordinates. Overbar signifies evaluation using $(\bar{x},\bar{y})$ coordinates.
$\underline{\underline{N}}$	• Matrix $\underline{\underline{A}}-\underline{\underline{M}}$.
P	• $\sum\limits_{j=1}^{p}\delta\left(x-a'_j\right)\delta\left(y-b'_j\right)Q_j(t)$, which is the designation of p sources or sinks, each of strength Q_j, defined for equation (1).
Q_j	• Volumetric flow rate for point source or sink j; defined for equation (1).
Q_{ai}	• Volumetric flow rate at node i from leakage through a confining unit or river overlying an aquifer being dewatered.
Q_{ei}	• Volumetric flow rate at node i from areal head-dependent discharge.
Q_{pi}	• Volumetric flow rate at node i from point head-dependent discharge.
Q_{ri}	• Volumetric flow rate at node i from line head-dependent discharge.
q_B	• Specified flow (specific discharge times aquifer thickness) normal to a boundary.
q_n	• Normal component of flow (specific discharge times aquifer thickness) at a boundary.
R	• Hydraulic conductance of a confining unit; R^e is the constant value of R for spatial element e.
(r,z)	• Radial and vertical coordinates of the axisymmetric cylindrical coordinate system.
r_ℓ	• Residual vector $\underline{B}_\ell - \underline{\underline{A}}_\ell \underline{h}_\ell$ for the finite-element matrix equations at iteration ℓ.
S	• Storage coefficient of the aquifer; S^e is the constant value of S for spatial element e.

S_s • Specific storage; S_s^e is the constant value of S_s for spatial element e.

S_s' • Specific storage of a confining unit; $S_s'^e$ is the constant value of S_s' for spatial element e.

S_y • Specific yield; S_y^e is the constant value of S_y for spatial element e.

$S_1\left(\Delta t_D\right)$ • Infinite series for transient leakage; defined by equation (186).

$S_2\left(\Delta t_D\right)$ • Infinite series for transient leakage; defined by equation (187).

$\underline{s}_k$ • Displacement vector $\underline{x}_{k+1} - \underline{x}_k$ for the iterative GCGM method.

$\left(T_{xx}, T_{xy},\right.$ • Components of the transmissivity tensor for the aquifer

$\left.T_{yx}, T_{yy}\right)$ written using Cartesian coordinates (x,y). Principal components in the $(\bar{x},\bar{y})$ coordinate system are $\left(T_{\bar{x}\bar{x}}, T_{\bar{y}\bar{y}}\right)$.

t • Time.

t' • Time since time-level n, $t - t_n$.

$\underline{\underline{U}}$ • Upper triangular matrix for symmetric-Doolittle factorization of $\underline{\underline{A}}$; defined by equation (259). Entries are u_{ij}, $i < j$, and α_{ii}.

$\underline{\underline{\tilde{U}}}$ • Upper triangular matrix for incomplete-Cholesky or modified incomplete-Cholesky factorization of $\underline{\underline{A}}$; defined by equation (272). Nonzero entries are $\tilde{u}_{ij}$, $i < j$, and $\tilde{\alpha}_{ii}$.

$\underline{\underline{V}}$ • Diagonal matrix defined by entries $V_{ii} = \sum_{e_i} v_{ii}^e$.

v_B • Specified specific discharge normal to a boundary.

v_n • Normal component of specific discharge at a boundary.

v_{ii}^e • Hydraulic conductance and Cauchy-type boundary condition term defined by equation (37) for Cartesian coordinates and by equation (225) for axisymmetric cylindrical coordinates.

W • Unit areal recharge or discharge rate for the aquifer; W^e is the value of W for spatial element e.

(x,y) • Global Cartesian coordinates.

$(\bar{x},\bar{y})$ • Local, rotated Cartesian coordinates along the principal directions of the transmissivity tensor.

$\underline{x}$ • Solution vector for the finite-element matrix equation (254).

$\underline{y}$ • Intermediate vector for the symmetric-Doolittle factorization solution of equation (254) or equation (269).

z • Vertical coordinate direction, positive upward.

z_b • Elevation of the aquifer base; z_{bi} is z_b at node i.

z_e • Elevation below which the areal head-dependent discharge function vanishes; z_{ei} is z_e at node i.

z_p • Elevation below which the point head-dependent discharge function vanishes; z_{pi} is z_p at node i.

z_r • Elevation at which discharge to the aquifer from a line head-dependent source or sink is at a maximum; z_{ri} is z_r at node i.

z_t • Elevation of the top of the aquifer (base of the confining unit); z_{ti} is z_t at node i.

Greek

α • Parameter for Cauchy-type boundary conditions in Cartesian coordinates; defined by equation (4).

α' • Parameter for Cauchy-type boundary conditions in axisymmetric cylindrical coordinates; defined by equation (209).

α_{ii} • Entry of matrix $\underline{\underline{D}}^{-1}$.

$\tilde{\alpha}_{ii}$ • Entry of matrix $\underline{\underline{\tilde{D}}}^{-1}$.

α_m, β_m • Exponents defined by equations (176) and (177) used to approximate the infinite series for transient leakage calculations.

γ_i • Transient leakage parameter for node i; defined by equation (167).

Δ^e • Area of element e; defined by equation (11) for Cartesian coordinates and by equation (213) for axisymmetric cylindrical coordinates.

Δt_n • Time interval $t_n - t_{n-1}$ for time element n.

Δt_D • Dimensionless time interval $\gamma_i \Delta t_n$.

$\underline{\delta}$ • Head change vector $\frac{2}{3}\left(\underline{\hat{h}}_{n+1} - \underline{\hat{h}}_n\right)$ over time interval $\frac{2}{3}\Delta t_{n+1}$ for unsteady-state problems; $\underline{\delta}_\ell$ is head change $\underline{\hat{h}}_{\ell+1} - \underline{\hat{h}}_\ell$ from iteration ℓ to iteration $\ell+1$ for nonlinear, steady-state problems.

$\underline{\delta}^*$ • Predicted head change vector over time interval $\frac{2}{3}\Delta t_{n+1}$ for predictor step of the predictor-corrector method.

$\underline{\delta}_o$ • Head change vector $\underline{h} - \underline{\hat{h}}_o$ computed for linear steady-state problems.

ϵ • Convergence criterion for the MICCG method; defined for equation (285).

ϵ_s • Convergence criterion for the iterative solution of nonlinear steady-state flow problems; defined for equation (240).

θ^e • Counter-clockwise rotation angle from (x,y) coordinates to $(\bar{x},\bar{y})$ coordinates in element e.

θ_i • Proportionate point in time element n+1 when node i converts from confined to unconfined flow or vice versa.

$\hat{\theta}_i$ • The estimate of θ_i given by equation (96).

σ_n, σ_{n+1} • Basis functions for time elements; defined by equation (13).

ϕ_i • Proportionate point in time element n+1 when a point head-dependent discharge function, an areal head-dependent leakage function, or line head-dependent leakage function changes form at node i.

ϕ'_i • $\phi_i\left(\phi_i + 1\right)/2$.

ϕ_{ei} • Proportionate point in time element n+1 when an areal head-dependent discharge function changes form at node i; defined by equation (134).

ϕ'_{ei} • $\phi_{ei}\left(\phi_{ei} + 1\right)/2$.

ϕ_{ti} • Proportionate point in time element n+1 when an areal head-dependent discharge function changes form at node i; defined by equation (133).

ϕ'_{ti} • $\phi_{ti}\left(\phi_{ti} + 1\right)/2$.