

ITERATIVE SOLUTION OF PIECEWISE LINEAR SYSTEMS AND APPLICATIONS TO FLOWS IN POROUS MEDIA*

LUIGI BRUGNANO[†] AND VINCENZO CASULLI[‡]

Abstract. The correct numerical modeling of free-surface hydrodynamic problems often requires to have the solution of special linear systems whose coefficient matrix is a piecewise constant function of the solution itself. In doing so, one may fulfill relevant physical constraints. The existence, the uniqueness, and two constructive iterative methods to solve a *piecewise linear system* of the form $\max[\mathbf{l}, \min(\mathbf{u}, \mathbf{x})] + T\mathbf{x} = \mathbf{b}$ are analyzed. The methods are shown to have a finite termination property; i.e., they converge to an exact solution in a finite number of steps and, actually, they converge very quickly, as confirmed by a few numerical tests, which are derived from the mathematical modeling of flows in porous media.

Key words. piecewise linear systems, free-surface hydrodynamics, wetting and drying, flows in porous media, confined-unconfined aquifers

AMS subject classifications. 90C33, 90C53, 90C06, 76M20, 76S05

DOI. 10.1137/08072749X

1. Introduction. The problem which is analyzed here is that of finding a solution to

$$(1) \quad \max[\mathbf{l}, \min(\mathbf{u}, \mathbf{x})] + T\mathbf{x} = \mathbf{b},$$

where the max and min functions are applied componentwise, $\mathbf{l}, \mathbf{u}, \mathbf{b} \in \mathbb{R}^n$ are known vectors, $\mathbf{l} = (l_i) \leq \mathbf{u} = (u_i)$, and T is a symmetric and (at least) positive semidefinite matrix of dimension n satisfying either one of the following properties:

T1. T is a Stieltjes matrix, i.e., a symmetric M -matrix (then nonsingular; see, e.g., [10]); or

T2. $\text{null}(T) \equiv \text{span}(\mathbf{v})$ (see, e.g., [9]) with $\mathbf{v} > \mathbf{0}$, and $T + D$ satisfies T1 for all *diagonal* matrices $D \not\leq O$ (i.e., $D \geq O$ and $D \neq O$), where hereafter $\mathbf{0}$ and O denote the zero vector and the zero matrix of appropriate size, respectively.

By denoting with I the identity matrix, if \mathbf{x} is a solution of the linear system $(I + T)\mathbf{x} = \mathbf{b}$ and $\mathbf{l} \leq \mathbf{x} \leq \mathbf{u}$, then \mathbf{x} is the unique solution of (1) and can be efficiently determined by a preconditioned conjugate gradient method.

If \mathbf{x} is a solution of the problem $\max(\mathbf{l}, \mathbf{x}) + T\mathbf{x} = \mathbf{b}$ and $\mathbf{x} \leq \mathbf{u}$, then \mathbf{x} is a solution of (1). Similarly, if \mathbf{x} is a solution of the problem $\min(\mathbf{u}, \mathbf{x}) + T\mathbf{x} = \mathbf{b}$ and $\mathbf{x} \geq \mathbf{l}$, then \mathbf{x} is a solution of (1). In both of these cases \mathbf{x} can be conveniently determined with a simple Newton-type method as described in [3].

In general, system (1) can be written, equivalently, as

$$(2) \quad P(\mathbf{x})(\mathbf{x} - \mathbf{l}) - Q(\mathbf{x})(\mathbf{x} - \mathbf{u}) + T\mathbf{x} = \mathbf{b} - \mathbf{l},$$

*Received by the editors June 16, 2008; accepted for publication (in revised form) November 26, 2008; published electronically March 13, 2009.

<http://www.siam.org/journals/sisc/31-3/72749.html>

[†]Dipartimento di Matematica “U. Dini,” Università di Firenze, Viale Morgagni 67/A, 50134 Firenze, Italy (luigi.brugnano@unifi.it).

[‡]Dipartimento di Ingegneria Civile e Ambientale, Università di Trento, Via Mesiano 77, 38050 Trento, Italy (vincenzo.casulli@unitn.it).

where $P(\mathbf{x})$ and $Q(\mathbf{x})$ are diagonal matrices whose diagonal entries are step functions given by

$$(3) \quad p(x_i) = \begin{cases} 1 & \text{if } x_i \geq l_i, \\ 0 & \text{otherwise,} \end{cases} \quad q(x_i) = \begin{cases} 1 & \text{if } x_i > u_i, \\ 0 & \text{otherwise.} \end{cases}$$

Because of the characterization (3) of system (2), this will be said to be a *piecewise linear system* (see also [3]).

The following easy properties are stated here for later reference.

LEMMA 1. *With reference to (2)–(3) $\forall \mathbf{x} \in \mathbb{R}^n$ one has*

- $I \geq P(\mathbf{x}) \geq Q(\mathbf{x}) \geq O$;
- *for any diagonal matrix R with binary diagonal entries (0 or 1),*

$$[P(\mathbf{x}) - R](\mathbf{x} - \mathbf{l}) \geq \mathbf{0}, \quad [Q(\mathbf{x}) - R](\mathbf{x} - \mathbf{u}) \geq \mathbf{0}.$$

Problems in the form (1) arise in the numerical solution of free-surface hydrodynamic problems so that their efficient solution is of interest in applications (see, e.g., [7, 8, 12, 13]; see also [5]).

By suitably increasing its dimensionality, problem (1) could be written as a non-smooth system of equations [14]. In particular, it can be reformulated as a box constrained semismooth problem, whose numerical solution is an active field of investigation (see, e.g., [2, 11] and the references contained therein). Alternatively, if the dimension of the problem is left unchanged, system (2) could be solved by means of some kind of Picard iteration. Nevertheless, for all such methods, convergence to the solution, in general, occurs only in the limit of an infinite number of iterations.

Moreover, a straightforward application of the Newton method could even lead to undefined and nonconverging iterations. Consider, for example, problem (2) with

$$(4) \quad \mathbf{l} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} -3 \\ 4 \end{pmatrix}, \quad T = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

and the following Newton-type iteration for its solution

$$(5) \quad \begin{aligned} P^0 &= I, & Q^0 &= O, \\ [P^{k-1} - Q^{k-1} + T] \mathbf{x}^k &= \mathbf{b} - (I - P^{k-1}) \mathbf{l} - Q^{k-1} \mathbf{u}, & k &= 1, 2, \dots, \end{aligned}$$

where, hereafter, $P^k = P(\mathbf{x}^k)$ and $Q^k = Q(\mathbf{x}^k)$ for all $k \geq 1$. It turns out that such an iteration becomes undefined after just one iteration. In fact,

$$\mathbf{x}^1 = \frac{1}{3} \begin{pmatrix} -2 \\ 5 \end{pmatrix} \implies P^1 = Q^1 = \begin{pmatrix} 0 & \\ & 1 \end{pmatrix}.$$

Consequently, iteration (5) is no longer defined and \mathbf{x}^1 is clearly not a solution. On the other hand, if we consider the following modified Newton method

$$\begin{aligned} \mathbf{x}^k &= \mathbf{x}^{k-1} - (I + T)^{-1} [P^{k-1}(\mathbf{x}^{k-1} - \mathbf{l}) - Q^{k-1}(\mathbf{x}^{k-1} - \mathbf{u}) \\ &\quad + T\mathbf{x}^{k-1} - \mathbf{b} + \mathbf{l}], & k &= 1, 2, \dots, \end{aligned}$$

global convergence can be readily shown. However, starting from $\mathbf{x}^0 = \mathbf{0}$, this problem requires 34 iterations to obtain an approximate solution within 15 digits of accuracy.

ALGORITHM 1.

```

Set  $Q^0 = O$ 
Do  $k = 1, n$ 
  Set  $P^{k,0} = I$ 
  Do  $\nu = 1, n$ 
    Solve  $P^{k,\nu-1}(\mathbf{x}^{k,\nu} - \mathbf{l}) - Q^{k-1}(\mathbf{x}^{k,\nu} - \mathbf{u}) + T\mathbf{x}^{k,\nu} = \mathbf{b} - \mathbf{l}$ 
    If  $(P^{k,\nu} - P^{k,\nu-1})(\mathbf{x}^{k,\nu} - \mathbf{l}) = \mathbf{0}$ 
      Set  $\mathbf{x}^k = \mathbf{x}^{k,\nu}$  and Exit
    End If
  End Do
  If  $(Q^k - Q^{k-1})(\mathbf{x}^k - \mathbf{u}) = \mathbf{0}$ 
    Set  $\mathbf{x} = \mathbf{x}^k$  and Exit
  End If
End Do

```

In order to overcome these potential drawbacks, two simple nested (*inner-outer*) iterative algorithms will be defined. Often, inner-outer iterations require more computation than one-level iterations. Nevertheless, the two nested iterations proposed here are able to guarantee convergence to the exact solution in a *finite* number of iterations. For the above example, a total of two iterations are sufficient.

The existence of a solution for problem (2) is studied in section 2. In section 3 the conditions for its uniqueness are analyzed. Section 4 describes some numerical examples derived from the mathematical modeling of flows in porous media. Finally, some concluding remarks are contained in section 5.

2. Existence of a solution. The existence of a solution for system (2) will be established constructively in this section by means of two nested iterative methods.

First, by setting $Q^0 = O$, a sequence of iterates $\{\mathbf{x}^k\}$ is determined from

$$(6) \quad P^k(\mathbf{x}^k - \mathbf{l}) - Q^{k-1}(\mathbf{x}^k - \mathbf{u}) + T\mathbf{x}^k = \mathbf{b} - \mathbf{l}, \quad k = 1, 2, \dots$$

A straightforward stopping criterion for this iteration is provided by the following result.

LEMMA 2. *An exact solution of (2) is obtained from (6) when, for some $k \geq 1$, one has*

$$(7) \quad (Q^k - Q^{k-1})(\mathbf{x}^k - \mathbf{u}) = \mathbf{0}.$$

For all $k = 1, 2, \dots$, a vector \mathbf{x}^k solving (6) can be obtained by *inner* iterations by setting $P^{k,0} = I$ and

$$(8) \quad P^{k,\nu-1}(\mathbf{x}^{k,\nu} - \mathbf{l}) - Q^{k-1}(\mathbf{x}^{k,\nu} - \mathbf{u}) + T\mathbf{x}^{k,\nu} = \mathbf{b} - \mathbf{l}, \quad \nu = 1, 2, \dots,$$

where $P^{k,\nu} = P(\mathbf{x}^{k,\nu})$ for all $k, \nu \geq 1$. A straightforward stopping criterion for the inner iterations (8) is provided by the following lemma.

LEMMA 3. *An exact solution of (6) is obtained from (8) when, for some $\nu \geq 1$, one has*

$$(9) \quad (P^{k,\nu} - P^{k,\nu-1})(\mathbf{x}^{k,\nu} - \mathbf{l}) = \mathbf{0}.$$

This nested iterative scheme can be summarized into Algorithm 1. As one may observe, a finite upper bound for both the inner and the outer iterations has been

specified: in what follows it will be shown that these bounds are high enough to guarantee convergence to an exact solution. The following two results prove that Algorithm 1 is well defined.

THEOREM 1. *If T satisfies T2 and*

$$(10) \quad \mathbf{v}^T \mathbf{l} \leq \mathbf{v}^T \mathbf{b} \leq \mathbf{v}^T \mathbf{u},$$

then for all k and ν , until convergence, $P^{k,\nu-1} - Q^{k-1} + T$ is Stieltjes and, consequently, Algorithm 1 is well defined.

Proof. One proceeds by double induction. For $k = 1$ one has $P^{1,0} = I$ and $Q^0 = O$. Thus, $P^{1,0} - Q^0 + T$ is Stieltjes and the first inner iterate is well defined. Then, for $\nu \geq 1$ one assumes that $P^{1,\nu-1} - Q^0 + T$ is Stieltjes so that $\mathbf{x}^{1,\nu}$ can be uniquely determined from (8) and $P^{1,\nu} \geq O$. Now, since $Q^0 = O$, if $P^{1,\nu} \neq O$, one has that $P^{1,\nu} - Q^0 + T$ is Stieltjes and, consequently, the subsequent $(\nu+1)$ st inner iterate is well defined. Alternatively, if $P^{1,\nu} = O$, (8) can be written as

$$-(P^{1,\nu} - P^{1,\nu-1})(\mathbf{x}^{1,\nu} - \mathbf{l}) + T\mathbf{x}^{1,\nu} = \mathbf{b} - \mathbf{l}.$$

Thus, from Lemma 1 and (10) one obtains

$$0 \geq -\mathbf{v}^T (P^{1,\nu} - P^{1,\nu-1})(\mathbf{x}^{1,\nu} - \mathbf{l}) = \mathbf{v}^T (\mathbf{b} - \mathbf{l}) \geq 0$$

so that $(P^{1,\nu} - P^{1,\nu-1})(\mathbf{x}^{1,\nu} - \mathbf{l}) = \mathbf{0}$, which is the exit condition for the inner loop. Hence, when $k = 1$, the inner iterates (8) are well defined for all ν until convergence.

Next, for $k > 1$, one assumes that the k th external cycle has been successfully completed so that \mathbf{x}^k is a solution of (6) and $Q^k \leq I$. Now, (6) can also be written as

$$(11) \quad (P^k - Q^k)(\mathbf{x}^k - \mathbf{l}) + (Q^k - Q^{k-1})(\mathbf{x}^k - \mathbf{u}) + T\mathbf{x}^k = \mathbf{b} - \mathbf{l} - Q^k(\mathbf{u} - \mathbf{l}),$$

which, from Lemma 1, implies

$$(12) \quad \mathbf{v}^T [\mathbf{b} - \mathbf{l} - Q^k(\mathbf{u} - \mathbf{l})] \geq 0.$$

Moreover, if $Q^k = I$, one has $P^k = I$ and (11) reduces to

$$(Q^k - Q^{k-1})(\mathbf{x}^k - \mathbf{u}) + T\mathbf{x}^k = \mathbf{b} - \mathbf{u}.$$

Thus, from Lemma 1, one has

$$0 \leq \mathbf{v}^T (Q^k - Q^{k-1})(\mathbf{x}^k - \mathbf{u}) = \mathbf{v}^T (\mathbf{b} - \mathbf{u}) \leq 0$$

so that $(Q^k - Q^{k-1})(\mathbf{x}^k - \mathbf{u}) = \mathbf{0}$, which is the exit condition for the outer loop. Alternatively, if $Q^k \neq I$, one has $P^{k+1,0} = I \not\geq Q^k$. Hence, $P^{k+1,0} - Q^k + T$ is Stieltjes, and the first one of the $(k+1)$ st inner iterates is well defined. Then, for $\nu \geq 1$, assuming that $P^{k+1,\nu-1} - Q^k + T$ is Stieltjes, the ν th inner iterate is well defined and $\mathbf{x}^{k+1,\nu}$ is uniquely determined from

$$(13) \quad P^{k+1,\nu-1}(\mathbf{x}^{k+1,\nu} - \mathbf{l}) - Q^k(\mathbf{x}^{k+1,\nu} - \mathbf{u}) + T\mathbf{x}^{k+1,\nu} = \mathbf{b} - \mathbf{l}$$

and satisfies

$$(14) \quad \mathbf{x}^{k+1,\nu} \geq \mathbf{x}^k.$$

In fact, by equating the left-hand sides of (6) and (13), one has

$$(P^{k+1,\nu-1} - Q^k + T) \mathbf{x}^{k+1,\nu} = (P^{k+1,\nu-1} - Q^k + T) \mathbf{x}^k + \boldsymbol{\xi}^{k,\nu},$$

where, from Lemma 1,

$$(15) \quad \boldsymbol{\xi}^{k,\nu} = (P^k - P^{k+1,\nu-1}) (\mathbf{x}^k - \mathbf{l}) + (Q^k - Q^{k-1}) (\mathbf{x}^k - \mathbf{u}) \geq \mathbf{0}.$$

Hence, because $(P^{k+1,\nu-1} - Q^k + T)^{-1} > O$, inequality (14) follows and implies $P^{k+1,\nu} \geq P^k \geq Q^k$. In order to complete the proof, note that if $P^{k+1,\nu} \not\geq Q^k$, then $P^{k+1,\nu} - Q^k + T$ is Stieltjes. Consequently, the corresponding $(\nu + 1)$ st inner iterate is well defined. Alternatively, if $P^{k+1,\nu} = Q^k$, (13) becomes

$$-(P^{k+1,\nu} - P^{k+1,\nu-1}) (\mathbf{x}^{k+1,\nu} - \mathbf{l}) + T \mathbf{x}^{k+1,\nu} = \mathbf{b} - \mathbf{l} - Q^k (\mathbf{u} - \mathbf{l}).$$

Consequently, from Lemma 1 and (12), one has

$$0 \geq -\mathbf{v}^T (P^{k+1,\nu} - P^{k+1,\nu-1}) (\mathbf{x}^{k+1,\nu} - \mathbf{l}) = \mathbf{v}^T [\mathbf{b} - \mathbf{l} - Q^k (\mathbf{u} - \mathbf{l})] \geq 0$$

so that $(P^{k+1,\nu} - P^{k+1,\nu-1})(\mathbf{x}^{k+1,\nu} - \mathbf{l}) = \mathbf{0}$, which is the exit condition for the inner loop. Hence, when $k > 1$, the inner iterates are well defined for all ν until convergence. \square

COROLLARY 1. *If T satisfies T1, then for all k and ν , until convergence, $P^{k,\nu-1} - Q^{k-1} + T$ is Stieltjes and, consequently, Algorithm 1 is well defined.*

Proof. For $k = 1$ one has $P^{1,\nu} \geq O = Q^0$ so that the corresponding inner iterations are well defined for all ν , since T satisfies T1.

Next, for $k > 1$, by using similar arguments as those used in the proof of Theorem 1, one obtains that (14) holds true and, consequently, $P^{k+1,\nu-1} \geq P^k \geq Q^k$. Hence, $P^{k+1,\nu-1} - Q^k + T$ is Stieltjes for all k and for all ν . \square

The next result provides the finite convergence property of Algorithm 1.

THEOREM 2. *Let T satisfy either T1 or T2. If T satisfies T2, assume also that (10) holds true. Then Algorithm 1 converges to an exact solution of problem (2) in at most $n(n + 1)/2$ (inner) steps.*

Proof. Recall first that $Q^0 = O$ and $\mathbf{x}^{k+1,\nu} \geq \mathbf{x}^k$ for all $k, \nu \geq 1$ (see (14)). Thus, $P^{k+1,\nu} \geq Q^k$. Moreover, within each external cycle, the inner iterates $\{\mathbf{x}^{k,\nu}\}$ are strictly decreasing (equality is excluded, because of the stopping criterion described in Lemma 3). Consider, in fact, two subsequent inner iterates, namely, (8) and

$$(16) \quad P^{k,\nu} (\mathbf{x}^{k,\nu+1} - \mathbf{l}) - Q^{k-1} (\mathbf{x}^{k,\nu+1} - \mathbf{u}) + T \mathbf{x}^{k,\nu+1} = \mathbf{b} - \mathbf{l}.$$

By equating the left-hand sides of (8) and (16), one obtains

$$(P^{k,\nu} - Q^{k-1} + T) \mathbf{x}^{k,\nu+1} = (P^{k,\nu} - Q^{k-1} + T) \mathbf{x}^{k,\nu} + \boldsymbol{\zeta}^{k,\nu},$$

where $\boldsymbol{\zeta}^{k,\nu} = -(P^{k,\nu} - P^{k,\nu-1})(\mathbf{x}^{k,\nu} - \mathbf{l}) \leq \mathbf{0}$. Thus, since $P^{k,\nu} - Q^{k-1} + T$ is Stieltjes, one has $(P^{k,\nu} - Q^{k-1} + T)^{-1} > O$, and, hence, $\{\mathbf{x}^{k,\nu}\}$ is strictly decreasing. Consequently, $\{P^{k,\nu}\}$ is decreasing as well, and because $P^{k,0} = I$ and $P^{k,\nu} \geq Q^{k-1}$, by denoting $\{q_i^k\}$ the diagonal entries of Q^k and setting

$$(17) \quad m_k = \sum_{i=1}^n q_i^k \geq k,$$

each inner cycle will converge in at most $n - m_{k-1}$ steps.

Next, inequality (14) implies that the external iterates $\{\mathbf{x}^k\}$ generated by (6) are not decreasing, i.e.,

$$(18) \quad \mathbf{x}^{k+1} \geq \mathbf{x}^k, \quad k \geq 1.$$

Consequently, the sequence $\{Q_k\}$ is not decreasing with $Q^0 = O$, and $Q^k \leq I$ for all k . Then, because of (17) and of the stopping criterion provided by Lemma 2, it follows that the outer cycle will converge in $K \leq n$ cycles. Therefore, in conclusion, the total number of inner iterations is bounded by

$$\sum_{k=1}^K (n - m_{k-1}) \leq \sum_{k=1}^n (n - k + 1) = \frac{n(n+1)}{2}. \quad \square$$

Remark 1. In practice, the determination of $\mathbf{x}^{k,\nu}$ from (8) can be accomplished quite efficiently by using a preconditioned conjugate gradient method (see, e.g., [9, 15]). This is particularly the case in applications where T is a sparse and very large matrix (see section 4). To this purpose, in light of the convergence property proved in Theorem 2, $\mathbf{x}^{k,\nu-1}$ (if $\nu > 1$) or \mathbf{x}^{k-1} (if $\nu = 1$) is conveniently used as a starting point for the conjugate gradient method (the effectiveness of this choice has been confirmed by a huge number of numerical tests).

Remark 2. Even though Theorem 2 provides a finite convergence property for the inner-outer iteration (6)–(8), the resulting upper bound is rather large when the dimension n of the problem is large. In practice, however, convergence is obtained in just a few iterations, as confirmed by several numerical tests.

COROLLARY 2. *If T satisfies T2, then (10) is a necessary and sufficient condition for problem (2) to have a solution.*

Proof. It has been already proved that (10) is a sufficient condition for problem (2) to have a solution. In order to prove that this condition is also necessary, from (2) one has

$$\mathbf{v}^T(\mathbf{b} - \mathbf{l}) = \mathbf{v}^T[P(\mathbf{x})(\mathbf{x} - \mathbf{l}) + Q(\mathbf{x})(\mathbf{u} - \mathbf{x})].$$

Moreover, from (3) and Lemma 1,

$$\mathbf{0} \leq P(\mathbf{x})(\mathbf{x} - \mathbf{l}) + Q(\mathbf{x})(\mathbf{u} - \mathbf{x}) \leq \mathbf{u} - \mathbf{l}.$$

The two inequalities in (10) then follow. \square

2.1. A dual algorithm. For the sake of completeness, a corresponding *dual* Algorithm 2, with respect to Algorithm 1, is also described, where $Q^{k,\nu} = Q(\mathbf{x}^{k,\nu})$ for all $k, \nu \geq 1$. Such an algorithm enjoys the same properties of Algorithm 1, which can be proved by means of similar arguments.

Observe that when the solution $\mathbf{x} \leq \mathbf{u}$, then $Q^0 = Q^1 = Q(\mathbf{x}) = O$ because of (18) and, consequently, Algorithm 1 requires only one outer iteration [3]. For the same reason, each outer iteration in Algorithm 2 requires only one inner iteration. Symmetrically, when the solution $\mathbf{x} \geq \mathbf{l}$, then $P^0 = P^1 = P(\mathbf{x}) = I$ and, consequently, Algorithm 2 converges in one outer iteration. For the same reason, each outer iteration in Algorithm 1 requires only one inner iteration.

3. Uniqueness of the solution. In this section the conditions for the uniqueness of the solution of problem (2) are studied. To begin with, the simpler case in which T satisfies T1 is analyzed.

ALGORITHM 2.

```

Set  $P^0 = I$ 
Do  $k = 1, n$ 
  Set  $Q^{k,0} = O$ 
  Do  $\nu = 1, n$ 
    Solve  $P^{k-1}(\mathbf{x}^{k,\nu} - \mathbf{l}) - Q^{k,\nu-1}(\mathbf{x}^{k,\nu} - \mathbf{u}) + T\mathbf{x}^{k,\nu} = \mathbf{b} - \mathbf{l}$ 
    If  $(Q^{k,\nu} - Q^{k,\nu-1})(\mathbf{x}^{k,\nu} - \mathbf{u}) = \mathbf{0}$ 
      Set  $\mathbf{x}^k = \mathbf{x}^{k,\nu}$  and Exit
    End If
  End Do
  If  $(P^k - P^{k-1})(\mathbf{x}^k - \mathbf{l}) = \mathbf{0}$ 
    Set  $\mathbf{x} = \mathbf{x}^k$  and Exit
  End If
End Do

```

THEOREM 3. *If T satisfies T1, then the solution of problem (2) exists and is unique.*

Proof. The existence of a solution \mathbf{x} of (2) has been already proved constructively by means of Algorithm 1. To prove its uniqueness, let \mathbf{y} be another solution of the same problem. Consequently,

$$(19) \quad P(\mathbf{x})(\mathbf{x} - \mathbf{l}) - Q(\mathbf{x})(\mathbf{x} - \mathbf{u}) + T\mathbf{x} = P(\mathbf{y})(\mathbf{y} - \mathbf{l}) - Q(\mathbf{y})(\mathbf{y} - \mathbf{u}) + T\mathbf{y}.$$

Moreover, one has

$$(20) \quad [P(\mathbf{y})(\mathbf{y} - \mathbf{l}) - Q(\mathbf{y})(\mathbf{y} - \mathbf{u})] - [P(\mathbf{x})(\mathbf{x} - \mathbf{l}) - Q(\mathbf{x})(\mathbf{x} - \mathbf{u})] = (\bar{P} - \bar{Q})(\mathbf{y} - \mathbf{x}),$$

where \bar{P} and \bar{Q} are diagonal matrices and whose diagonal entries $\{\bar{p}_i\}$ and $\{\bar{q}_i\}$ are, respectively, given by

$$(21) \quad \bar{p}_i = \begin{cases} 0 & \text{if } x_i, y_i < l_i, \\ 1 & \text{if } x_i, y_i \geq l_i, \\ \frac{x_i - l_i}{x_i - y_i} & \text{if } x_i \geq l_i > y_i, \\ \frac{y_i - l_i}{y_i - x_i} & \text{if } y_i \geq l_i > x_i, \end{cases} \quad \bar{q}_i = \begin{cases} 0 & \text{if } x_i, y_i \leq u_i, \\ 1 & \text{if } x_i, y_i > u_i, \\ \frac{x_i - u_i}{x_i - y_i} & \text{if } x_i > u_i \geq y_i, \\ \frac{y_i - u_i}{y_i - x_i} & \text{if } y_i > u_i \geq x_i \end{cases}$$

so that

$$(22) \quad \bar{P} \geq \bar{Q} \geq O.$$

From (19) and (20), one has

$$(23) \quad (\bar{P} - \bar{Q} + T)(\mathbf{y} - \mathbf{x}) = \mathbf{0}.$$

Since T satisfies T1, matrix $(\bar{P} - \bar{Q} + T)$ is Stieltjes and then nonsingular, and uniqueness ($\mathbf{y} = \mathbf{x}$) follows. \square

In order to discuss the remaining case, the following definition is given.

DEFINITION 1. *Let $\mathbf{x} = (x_i)$ be a solution of (2). This is said to be an interior solution if*

$$(24) \quad \exists i \in \{1, \dots, n\} : \quad l_i < x_i < u_i.$$

THEOREM 4. *If T satisfies T2 and (10) holds true, then problem (2) has at least one solution. Moreover,*

1. if such a solution is an interior solution, then this is the only solution of the problem;
2. if more than one solution exist, then the difference of any two solutions belongs to the null space of matrix T .

Proof. The existence of a solution \mathbf{x} of (2) has been already proved constructively by means of Algorithm 1 (see also Corollary 2). Its uniqueness is now proved when it is an interior solution. Let \mathbf{y} be another solution of the same problem. Consequently, (19)–(23) hold true. Since T satisfies T2, it suffices to show that $\bar{P} \neq \bar{Q}$. From (24) we know that there exists $i \in \{1, \dots, n\}$ such that $l_i < x_i < u_i$. For such index i , one has

$$\bar{p}_i - \bar{q}_i = \begin{cases} \frac{x_i - l_i}{x_i - y_i} > 0 & \text{if } y_i < l_i, \\ 1 & \text{if } l_i \leq y_i \leq u_i, \\ \frac{u_i - x_i}{y_i - x_i} > 0 & \text{if } y_i > u_i \end{cases}$$

so that $\bar{P} \neq \bar{Q}$. Therefore, matrix $\bar{P} - \bar{Q} + T$ is Stieltjes and uniqueness ($\mathbf{y} = \mathbf{x}$) follows from (23).

On the other hand, when \mathbf{x} is not unique, one has $\bar{P} = \bar{Q}$ so that (23) reduces to $T(\mathbf{y} - \mathbf{x}) = \mathbf{0}$. That is, $\mathbf{y} - \mathbf{x} \in \text{null}(T)$. \square

COROLLARY 3. If T satisfies T2 and (10) holds true, then a noninterior solution of problem (2) is generally not unique.

Proof. Let ω_u and ω_l be the two sets of indices defined by

$$\omega_u = \{i : x_i \geq u_i\}, \quad \omega_l = \{i : x_i \leq l_i\}.$$

If $\mathbf{v} = (v_i)$, set

$$\alpha = \begin{cases} -\min_{i \in \omega_u} \{(x_i - u_i)/v_i\} & \text{if } \omega_u \neq \emptyset, \\ -\infty & \text{otherwise,} \end{cases}$$

$$\beta = \begin{cases} \min_{i \in \omega_l} \{(l_i - x_i)/v_i\} & \text{if } \omega_l \neq \emptyset, \\ \infty & \text{otherwise.} \end{cases}$$

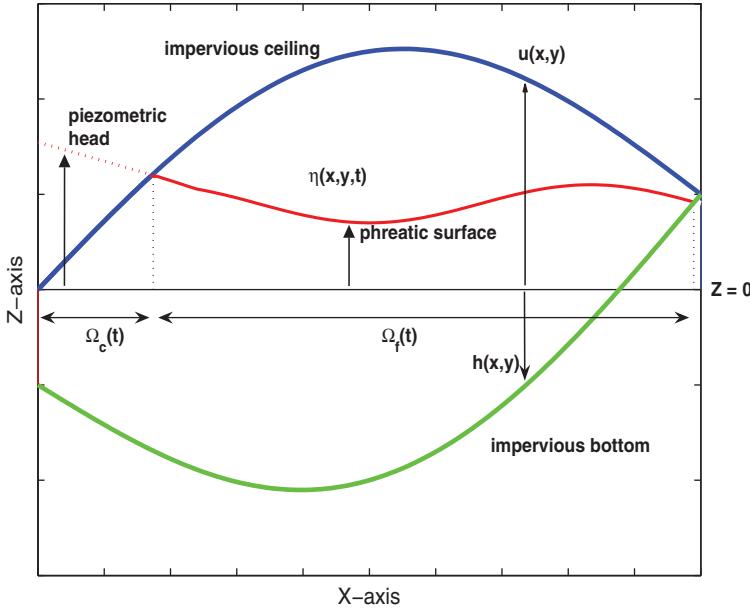
Clearly, $\alpha \leq 0 \leq \beta$. Moreover, one verifies that

$$(25) \quad \mathbf{x}(\theta) \equiv \mathbf{x} + \theta\mathbf{v}, \quad \theta \in [\alpha, \beta],$$

is a solution of the problem. \square

The following example elucidates the previous arguments.

Example 1. Let problem (2) be defined by (4). Consequently, T satisfies T2, with $\mathbf{v} = \mathbf{u}$, and (10) holds true. One verifies that $\mathbf{x} = (-1, 2)^T$ is a (noninterior) solution of (2). Moreover, according to Corollary 3, (25) is also a solution, with $\alpha = -1$ and $\beta = 1$. Therefore, in such a case we have infinite (noninterior) solutions to the problem. However, if we modify the right-hand side by setting $\mathbf{b} = (-1, 2)^T$, we have that (10) holds true. Moreover, one verifies that a solution to the problem (2) is given by $\mathbf{x} = (0, 1)^T$, which is a noninterior solution. Nevertheless, according to Corollary 3, such a solution turns out to be unique, because in this case one obtains $\alpha = \beta = 0$.

FIG. 1. *Confined-unconfined aquifer.*

4. Modeling confined-unconfined flows in porous media. Consider the mathematical modeling of a two-dimensional flow in a homogeneous and isotropic aquifer. When a free surface exists, the governing differential equation is given by the *Boussinesq equation* [1]

$$(26) \quad \varepsilon \eta_t = [\kappa(h + \eta)\eta_x]_x + [\kappa(h + \eta)\eta_y]_y + \varphi,$$

where (see Figure 1) x and y are coordinates in a horizontal reference frame; t is the time; ε and κ are the *porosity* and the *hydraulic conductivity*, respectively (which are assumed to be constant); $h(x, y)$ is the prescribed aquifer's bottom and $\eta(x, y, t)$ is the (unknown) free-surface location; $\varphi(x, y, t)$ represents the prescribed source or sink; and the time dependent domain is

$$\Omega(t) = \{(x, y) : h(x, y) + \eta(x, y, t) > 0\}, \quad t > 0.$$

Moreover, if the aquifer is confined from above by a ceiling set at distance $u(x, y)$ from the vertical reference level, one has that (26) actually holds true only in the subdomain

$$(27) \quad \Omega_f(t) = \{(x, y) \in \Omega(t) : -h(x, y) < \eta(x, y, t) < u(x, y)\}, \quad t > 0,$$

whereas, in

$$(28) \quad \Omega_c(t) = \Omega(t) \setminus \overline{\Omega}_f(t), \quad t > 0,$$

the governing equation is

$$(29) \quad 0 = [\kappa(h + \eta)\eta_x]_x + [\kappa(h + \eta)\eta_y]_y + \varphi.$$

In such a case, the unknown $\eta(x, y, t)$ represents the *piezometric head* [1]. The two subdomains (27) and (28) are obviously separated by the (unknown) line defined by

$$\Gamma_{fc}(t) = \{(x, y) : \eta = u\}.$$

The correct numerical solution of the mixed problem (26)–(29) is of interest in applications [6] and is nontrivial, because of the involvement of two moving boundaries [12]. For this purpose, consider a suitably large square containing $\Omega(t)$, $t > 0$, which is covered by a Cartesian grid having size Δx and Δy , $(x_i, y_j) = (i\Delta x, j\Delta y)$. Moreover, for a given time step Δt , let $t_\ell = \ell\Delta t$, $\ell \geq 0$, denote the ℓ th time level. It is assumed the knowledge of ε , κ , $h_{ij} = h(x_i, y_j)$, and $u_{ij} = u(x_i, y_j)$, which are independent from time, and

$$\varphi_{ij}^\ell = \varphi(x_i, y_j, t_\ell), \quad \ell \geq 0, \quad \eta_{ij}^0 = \eta(x_i, y_j, 0).$$

Finally, the total water depth is defined by

$$(30) \quad H_{ij}^\ell = \max \{0, \min \{h_{ij} + u_{ij}, h_{ij} + \eta_{ij}^\ell\}\},$$

which yields $H_{ij}^\ell = 0$ in the dry area, $H_{ij}^\ell = h_{ij} + \eta_{ij}^\ell$ in $\Omega_f(t)$, and $H_{ij}^\ell = h_{ij} + u_{ij}$ in $\Omega_c(t)$. Then, a finite difference discretization which is consistent with both equations (26) and (29) is taken to be (see also [3, 4])

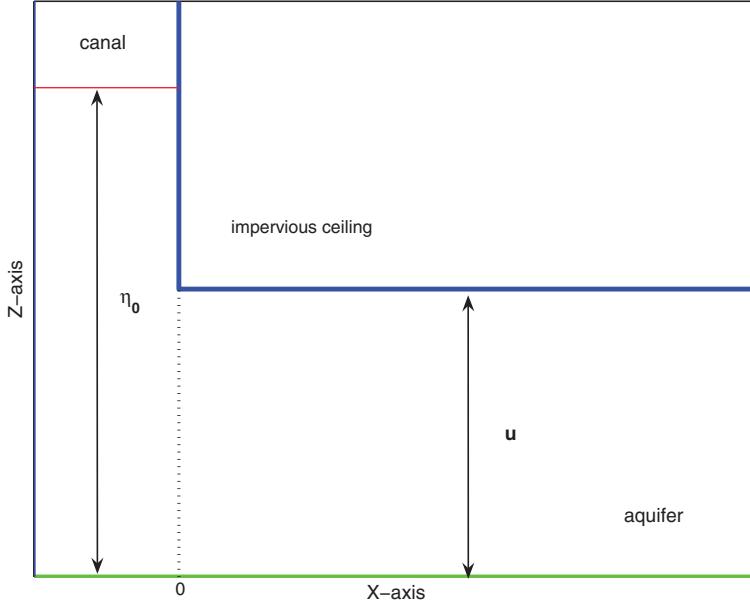
$$(31) \quad \begin{aligned} & \varepsilon \left(\frac{\max[-h_{ij}, \min(u_{ij}, \eta_{ij}^{\ell+1})] - \max[-h_{ij}, \min(u_{ij}, \eta_{ij}^\ell)]}{\Delta t} \right) \\ & - \kappa \left(\frac{H_{i+\frac{1}{2},j}^\ell (\eta_{i+1,j}^{\ell+1} - \eta_{ij}^{\ell+1}) - H_{i-\frac{1}{2},j}^\ell (\eta_{ij}^{\ell+1} - \eta_{i-1,j}^{\ell+1})}{\Delta x^2} \right. \\ & \left. + \frac{H_{i,j+\frac{1}{2}}^\ell (\eta_{i,j+1}^{\ell+1} - \eta_{ij}^{\ell+1}) - H_{i,j-\frac{1}{2}}^\ell (\eta_{ij}^{\ell+1} - \eta_{i,j-1}^{\ell+1})}{\Delta y^2} \right) = \varphi_{ij}^\ell. \end{aligned}$$

Between grid points, the aquifer thicknesses $H_{i\pm 1/2,j}^\ell$ and $H_{i,j\pm 1/2}^\ell$ are defined as averages from the nearest grid values.

It is to be noted that for those grid points (i, j) where $\varphi_{ij}^\ell = 0$, $H_{i\pm 1/2,j}^\ell = 0$, and $H_{i,j\pm 1/2}^\ell = 0$, (31) trivially implies (see (30)) $H_{ij}^{\ell+1} = H_{ij}^\ell = 0$ (dry area). In this case (31) does not contribute to the system that is being formulated.

The remaining set of equations (corresponding to those grid points where at least one of $H_{i\pm 1/2,j}^\ell$ and $H_{i,j\pm 1/2}^\ell$ is strictly positive) can be assembled into a single *piecewise linear* system. Upon multiplication by $\Delta t/\varepsilon$ and by setting $l_{ij} = -h_{ij}$, this system (which has to be solved *at every time step*) can be recognized as being in the form (1). The resulting matrix T is irreducible, sparse, symmetric, at least positive semidefinite, and of time dependent size $n_\ell \times n_\ell$, with n_ℓ being the number of grid points (i, j) where at least one of $H_{i\pm 1/2,j}^\ell$ and $H_{i,j\pm 1/2}^\ell$ is strictly positive. This matrix satisfies either T1, when the free-surface elevation is specified in at least one point of the boundary, or T2, when only the flow is specified at the boundary. According to Remark 1, a preconditioned conjugate gradient method is used for solving the resulting linear systems.

4.1. Bounded recharging aquifer. A first numerical test is presented, which is aimed to verify the accuracy of the discrete method (31). For this purpose, consider the simple case of a one-dimensional confined-unconfined aquifer reported in [12]. A canal (see Figure 2) confines with a bounded recharging aquifer whose ceiling height is u . The aquifer is initially dry. Let η_0 be the (known) constant height of

FIG. 2. *Bounded recharging aquifer.*

water in the canal. The following parameters are specified

$$\kappa = 10^{-3}\pi \text{ m/s}, \quad \varepsilon = 10^{-1}\pi, \quad u = 5 \text{ m}, \quad \eta_0 = 5.5 \text{ m},$$

for which the analytical solution has been reported in [12, p. 154].

A numerical simulation has been carried out by using a spatial step size $\Delta x = 2.4 \cdot 10^{-2} \text{ m}$, and a time step $\Delta t = 0.5 \text{ s}$. Because of the prescribed level η_0 of water in the canal, matrix T in the resulting piecewise linear system (2) is tridiagonal and satisfies T1. The simulation has been carried out for $T = 900 \text{ s}$. In almost all time steps, Algorithm 1 requires only 1 outer iteration, each requiring 1 inner iteration. The computed results shown in Figures 3 and 4 are in perfect agreement with the analytical solution reported in [12], thus, confirming the effectiveness of scheme (31). In particular, in Figure 3 the free-surface profiles are plotted in the solid line, whereas the piezometric head profiles are plotted in the dotted line at different times. Figure 4 shows the amount of recharging water $Q(t)$.

4.2. Free-surface dynamics in the proximity of a well. As a second test problem, consider the case in which the aquifer's bottom and ceiling have been assumed to be symmetrical, with respect to the reference vertical level, both described by a paraboloid of revolution given by

$$(32) \quad u(x, y) \equiv h(x, y) = h_0 \left(1 - \frac{x^2 + y^2}{L^2} \right),$$

where h_0 and L are given positive constants. It is also assumed that at the initial time $t_0 = 0$ all of the porous space is filled with water so that $\eta(x, y, 0) = u(x, y)$ and, consequently, $\Omega(0) = \{(x, y) : x^2 + y^2 < L^2\} \equiv \Omega_c(0)$. The resulting matrix T , which satisfies T2, is sparse, symmetric positive semidefinite, and of size n^ℓ , which, at the initial time t_0 , is $n^\ell \sim N^2 \frac{\pi}{4}$.

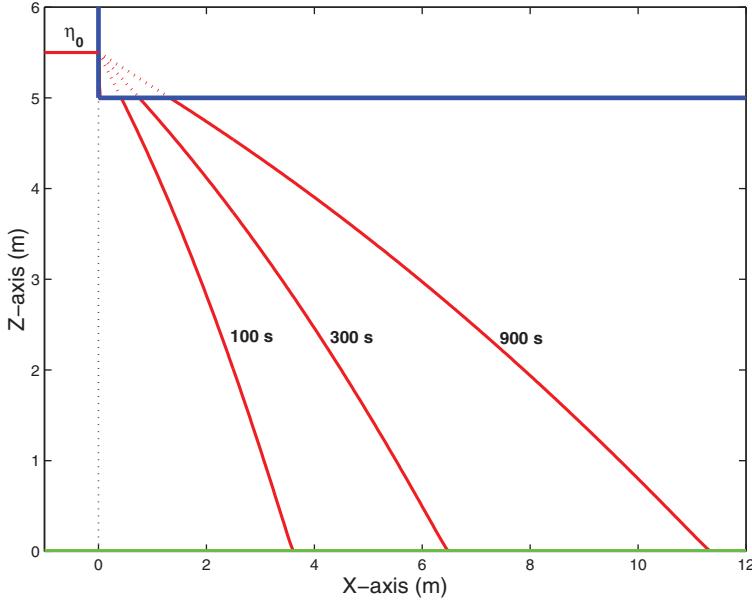


FIG. 3. Predicted free-surface and piezometric head profiles at different times.

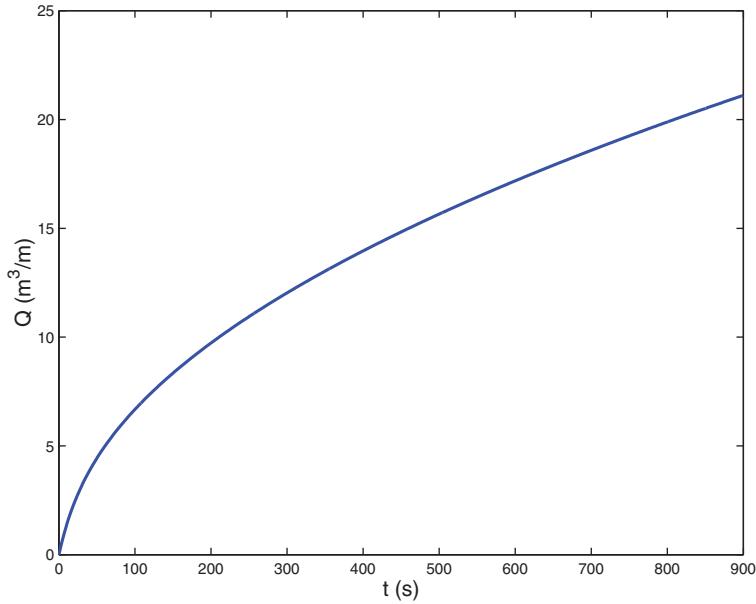


FIG. 4. Amount of recharging water.

A numerical simulation has been carried out for 14 days by using $\Delta x = \Delta y = 2L/(N+1)$ and a relatively large time step size $\Delta t = 1$ day. For the present simulations the chosen parameters are $\varepsilon = 0.4$, $\kappa = 1$ m/s, $h_0 = 10$ m, and $L = 10^3$ m. The flow is then driven by an idealized pointwise sink, representing a well located at the origin that pumps water out of the aquifer at a constant rate $q = 10$ m³/s. Thus, by setting

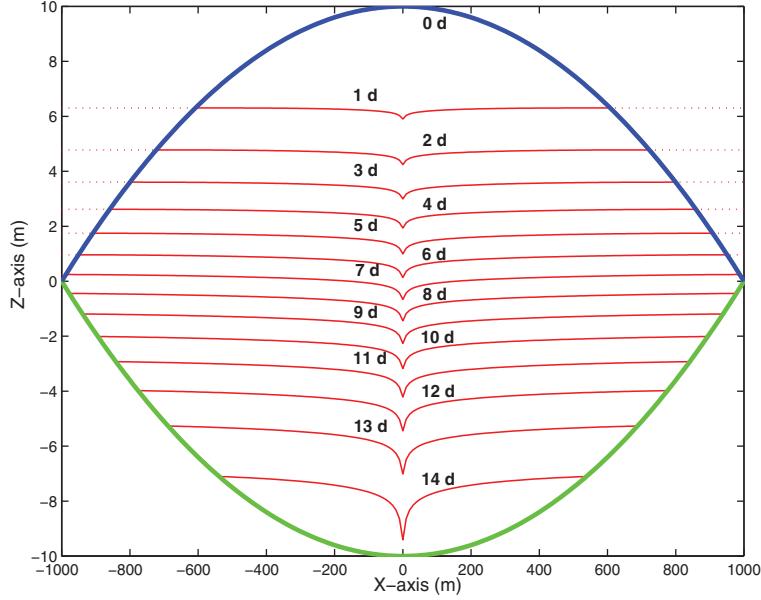
FIG. 5. Computed free surfaces at the cross section $y = 0$.

TABLE 1
Numerical results for the second test problem.

ℓ	$N = 50$				$N = 100$				$N = 200$			
	k_{out}	ν_{tot}	n_ℓ	V_ℓ	k_{out}	ν_{tot}	n_ℓ	V_ℓ	k_{out}	ν_{tot}	n_ℓ	V_ℓ
0	-	-	-	12564992	-	-	-	12566221	-	-	-	12566346
1	5	5	2085	11700992	5	5	8109	11702221	5	5	31965	11702346
2	3	3	2085	10836992	4	4	8109	10838221	4	4	31965	10838346
3	3	3	2085	9972992	3	3	8109	9974221	3	3	31965	9974346
4	3	3	2085	9108992	3	3	8109	9110221	3	3	31965	9110346
5	2	2	2085	8244992	3	3	8109	8246221	3	3	31965	8246346
6	3	3	2085	7380992	3	3	8109	7382221	3	3	31965	7382346
7	3	3	2085	6516992	3	3	8109	6518221	3	3	31965	6518346
8	2	5	2085	5652992	2	5	8109	5654221	2	6	31965	5654346
9	1	2	2025	4788992	1	3	7793	4790221	1	3	30597	4790346
10	1	3	1877	3924992	1	3	7177	3926221	1	3	28177	3926346
11	1	3	1693	3060992	1	3	6533	3062221	1	3	25621	3062346
12	1	3	1509	2196992	1	3	5797	2198221	1	3	22689	2198346
13	1	3	1297	1332992	1	3	4929	1334221	1	3	19349	1334346
14	1	4	1033	468992	1	4	3909	470221	1	4	15249	470346

$\varphi_{ij}^\ell = 0$, except $\varphi_{00}^\ell = -\frac{q}{\Delta x \Delta y}$, the new water volume at time $t_{\ell+1}$ is given by

$$(33) \quad V^{\ell+1} = \varepsilon \Delta x \Delta y \sum_{ij} H_{ij}^{\ell+1} = V^\ell - q \Delta t.$$

Figure 5 shows, in the solid line, the resulting free-surface elevation (and, where appropriate, the piezometric head in the dotted line) η_{i0}^ℓ at the cross section $y = 0$ for $\ell = 0, 1, \dots, 14$.

For specified grid resolutions corresponding to $N = 50, 100$, and 200 , Table 1 shows, for each time step, the number of outer (k_{out}) and total inner (ν_{tot}) iterations required, along with the size n^ℓ of the resulting system, and, finally, the computed

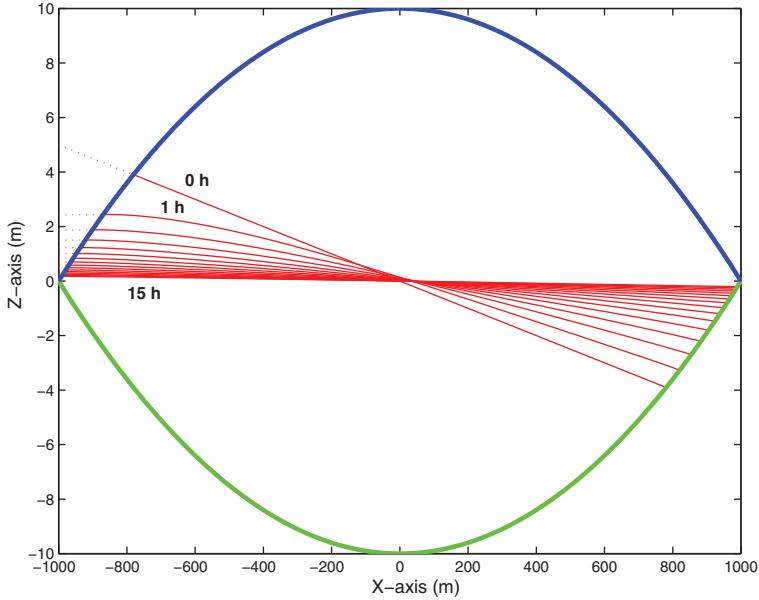


FIG. 6. *Time evolution to steady state at the cross section $y = 0$.*

water volume at each time step. As one expects from Figure 5, n^ℓ is constant for the first 8 time levels and then decreases in the subsequent ones. Moreover, the number of both inner and outer iterations turns out to be remarkably small and practically insensitive to grid resolution.

Table 1 also shows that, for all $\ell = 1, 2, \dots, 14$, the water volumes are strictly positive and linearly decreasing at a constant rate. In fact, the volume difference between two subsequent time levels (see (33)) is correctly given by $V^{\ell+1} - V^\ell = -q\Delta t = -864,000 \text{ m}^3$. Thus, any attempt to extend the simulation beyond day 14 would produce a *physically unrealistic* negative water volume $V^{15} < 0$. It can be shown that this implies that the right-hand side of (31) violates the first inequality in (10). Consequently, when $\ell = 15$, Corollary 2 applies, indicating that problem (31) does not have a solution. This is an interesting example demonstrating that the proposed algorithm does not permit artificial overdrainage.

It is also worth noting that, in the first 7 time levels, η_{ij} is everywhere no less than l_{ij} . Consistently with that observed in section 2.1, convergence is achieved in only 1 inner iteration per outer iteration. On the other hand, starting from time level 9, η_{ij} is everywhere no larger than u_{ij} and only 1 outer iteration is required for convergence.

4.3. Evolution to steady state. As a third and final test problem, another aquifer is considered, with the same functions $h(x, y)$ and $u(x, y)$ as defined in (32) and the initial water level and piezometric head $\eta(x, y, 0)$ being flat, at a slope, as indicated in Figure 6.

A numerical simulation has been carried out for 15 hours, by using a time step size $\Delta t = 0.1$ hour. The chosen parameters are the same as in the previous test problem and with $\varphi_{ij}^\ell \equiv 0$.

For specified $N = 50, 100$, and 200 , Table 2 shows, at selected time steps, the number of outer (k_{out}) and total inner (ν_{tot}) iterations required for convergence, along

TABLE 2
Numerical results for the third test problem.

ℓ	$N = 50, V = 6282496$			$N = 100, V = 6283110$			$N = 200, V = 6283173$		
	k_{out}	ν_{tot}	n_ℓ	k_{out}	ν_{tot}	n_ℓ	k_{out}	ν_{tot}	n_ℓ
10	2	3	1889	2	4	7277	2	4	28577
20	2	3	1926	2	3	7444	2	3	29247
30	2	3	1957	3	4	7582	2	3	29802
40	2	3	1989	2	3	7693	2	3	30228
50	2	3	2006	2	3	7774	2	3	30564
60	1	2	2022	2	3	7828	2	3	30823
70	2	2	2036	2	3	7889	2	3	31026
80	1	1	2049	2	3	7923	2	3	31205
90	2	2	2057	2	3	7972	2	3	31340
100	1	1	2063	1	2	7994	2	3	31459
110	2	2	2065	2	3	8010	2	3	31545
120	1	1	2069	1	2	8029	2	3	31614
130	1	1	2073	2	3	8041	2	3	31670
140	1	1	2079	1	1	8067	2	2	31738
150	1	1	2083	1	1	8075	2	3	31768

with the size n^ℓ of the resulting system. As expected, during the whole simulation the water volume is *exactly* conserved. Moreover, as shown in Figure 6, as time advances, the phreatic surface approaches its steady state configuration. Also in this case, the number of both inner and outer iterations is remarkably small and almost insensitive to grid resolution, thus, confirming the usefulness of the proposed algorithm for real world applications.

5. Conclusions. Two simple algorithms based on a nested iterative procedure, aimed to solve certain *piecewise linear systems* that arise from the numerical modeling of free-surface hydrodynamics, have been described and analyzed.

It is shown that, under rather general assumptions, the iterates are well defined and converge to the exact solution in a finite number of steps. Existence of the solution has been established under the same assumptions for which convergence is assured. Moreover, convergence to a solution is guaranteed also in the case where there is no uniqueness. In such a case, it has been shown how to retrieve all solutions of the problem.

Simple, and yet nontrivial, numerical tests have confirmed the efficiency, the robustness, and the usefulness of the proposed algorithms for real world applications to flows in confined, unconfined and mixed confined-unconfined aquifers.

REFERENCES

- [1] J. BEAR AND A. VERRUIJT, *Modeling Groundwater Flow and Pollution*, D. Reidel, Dordrecht, Holland, 1987.
- [2] S. BELLAVIA, M. MACCONI, AND B. MORINI, *An affine scaling trust-region approach to bound-constrained nonlinear systems*, Appl. Numer. Math., 44 (2003), pp. 257–280.
- [3] L. BRUGNANO AND V. CASULLI, *Iterative solution of piecewise linear systems*, SIAM J. Sci. Comput., 30 (2008), pp. 463–472.
- [4] V. CASULLI, *Semi-implicit finite difference methods for the two-dimensional shallow water equations*, J. Comput. Phys., 86 (1990), pp. 56–74.
- [5] V. CASULLI, *A high resolution wetting and drying algorithm for free-surface hydrodynamics*, Internat. J. Numer. Methods Fluids, to appear.
- [6] C.X. CHEN, L.T. HU, AND X.S. WANG, *Analysis of steady ground water flow toward wells in a confined-unconfined aquifer*, Ground Water, 44 (2006), pp. 609–612.

- [7] C.W. CRYER, *Successive overrelaxation methods for solving linear complementarity problems arising from free boundary value problems*, in Proceedings of the Free Boundary Problems, Vol. I, Pavia, 1979 INdAM "F. Severi", pp. 109–131.
- [8] C.W. CRYER, *The efficient solution of linear complementarity problems for tridiagonal Minkowski matrices*, ACM Trans. Math. Software, 9 (1983), pp. 199–214.
- [9] G.H. GOLUB AND C.F. VAN LOAN, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.
- [10] R.A. HORN AND C.R. JOHNSON, *Topics in Matrix Analysis*, Cambridge University Press, New York, 1991.
- [11] C. KANZOW AND A. KLUG, *An interior-point affine-scaling trust-region method for semismooth equations with box constraints*, Comput. Optim. Appl., 37 (2007), pp. 329–353.
- [12] L. LI, D.A. LOCKINGTON, D.A. BARRY, J.-Y. PARLANGE, AND P. PERROCHET, *Confined-unconfined flow in a horizontal aquifer*, J. Hydrology, 271 (2003), pp. 150–155.
- [13] Y. LIN AND C.W. CRYER, *An alternating direction implicit algorithm for the solution of linear complementarity problems arising from free boundary value problems*, Appl. Math. Optim., 13 (1985), pp. 1–17.
- [14] J.-S. PANG AND L. QI, *Nonsmooth equations: Motivation and algorithms*, SIAM J. Optim., 3 (1993), pp. 443–465.
- [15] Y. SAAD, *Iterative Methods for Sparse Linear Systems*, 2nd ed., SIAM, Philadelphia, 2003.