INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN FLUIDS, VOL. 20, 191–212 (1995)

A WEIGHTED LEAST SQUARES METHOD FOR FIRST-ORDER HYPERBOLIC SYSTEMS

D. G. ZEITOUN*, J. P. LAIBLE AND G. F. PINDER

College of Engineering and Mathematics, 101 Votey Building, University of Vermont, Burlington, VT 05403, U.S.A.

SUMMARY

The paper presents a generalization of the classical L^2-norm weighted least squares method for the numerical solution of a first-order hyperbolic system. This alternative least squares method consists of the minimization of the weighted sum of the L^2 residuals for each equation of the system. The order of accuracy of global conservation of each equation of the system is shown to be inversely proportional to the weight associated with the equation. The optimal relative weights between the equations are then determined in order to satisfy global conservation of the energy of the physical system.

As an application of the algorithm, the shallow water equations on an irregular domain are first discretized in time and then solved using Laplace modification and the proposed least squares method.

KEY WORDS Weighted least squares Hyperbolic system of equations Shallow water equations Newton–Raphson method

1. INTRODUCTION

While there are a number of different least squares formulations for solving boundary value problems, they may be generally classified into three groups.

In the first group the partial differential equation is transformed into an optimal control problem via the introduction of a state function, the solution of a given state equation. This equation is generally discretized using a Galerkin finite element approximation. The least squares solution algorithm employs a conjugate gradient method. Conservation of mass and satisfaction of boundary conditions are assured by requiring the state vector to belong to a suitable Sobolev space.[1,2]

In the second group the least squares approximation is combined with the Galerkin finite element method to obtain convergent finite element approximations. An example of this type of method is the Galerkin least squares method developed by Hughes *et al.*[3] In this approach the least squares form of the residual is added to the Galerkin method in order to stabilize the Galerkin method without degrading accuracy.

In the third group the mathematical form of the approximate solution (trial solution) is first chosen and then the norm of the corresponding residual is minimized by least squares.[4] This is the approach addressed in the present paper.

A collocation L^2 least squares formulation is proposed herein which takes into account in an optimal way the relative weights identified with the different equations of the system. While the methodology developed in this paper may be applied to various types of partial differential equations, herein the solution of a first-order hyperbolic system generally denoted as the shallow water equations is considered.

*Present address: TAHAL Consulting Eng. Ltd., P.O. Box 111700-6111, Tel-Aviv, Israel.

CCC 0271–2091/95/030191–22

Received 13 December 1993
Revised 19 May 1994

After a brief review in Section 2 of numerical methods for solving first- order systems of partial differential equations, basic definitions and mathematical properties of the weighted least squares method are discussed in Sections 3 and 4. The absence of global conservation properties of the classical least squares formulation leads to a new method proposed in Section 5 and applied to the solution of the shallow water equations in Section 6.

2. NUMERICAL METHODS FOR FIRST-ORDER SYSTEMS

Let Ω be a bounded domain in the (x, y) plane with boundary Γ.

The vector $\mathbf{n} = (n_x, n_y)$ is the outward unit normal to Γ.

The first-order system of coupled equations considered in this work may be written as

$$\mathbf{Lu} \equiv \mathbf{A}\frac{\partial \mathbf{u}}{\partial x} + \mathbf{D}\frac{\partial \mathbf{u}}{\partial y} + \mathbf{Cu} = \mathbf{f}, \tag{1}$$

with the boundary condition

$$\mathbf{Bu} \equiv (\mathbf{A}n_x + \mathbf{D}n_y - \mathbf{M})\mathbf{u} = \mathbf{g}. \tag{2}$$

$\mathbf{A}$, $\mathbf{D}$, $\mathbf{C}$ and $\mathbf{M}$ are $p \times p$ matrix-valued functions, the column vector forcing term is $\mathbf{f}^{\mathrm{T}} = (f_1, f_2, \ldots, f_p)$, $\mathbf{g}^{\mathrm{T}} = (g_1, g_2, \ldots, g_{pb})$ is the column vector of applied boundary conditions and the unknown column vector is $\mathbf{u}^{\mathrm{T}} = (u_1, u_2, \ldots, u_p)$ (T denotes transposition). $\mathbf{L}$ represents the linear differential operator defined in the domain Ω and $\mathbf{B}$ represents the boundary differential operator.

Sufficient conditions for problem (1), (2) to have a unique strong solution have been obtained for symmetric positive systems in the sense of Friedrichs.[5]

Numerical approximations of pure hyberbolic problems have been extensively studied. Different schemes based on finite difference and Galerkin finite element approximations have been proposed for the solution of linear and non-linear hyperbolic equations and first-order systems (see Reference 6 for a review).

A perceived disadvantage of the least squares method compared with the Galerkin method is the higher-order finite element continuity requirement. However, for basis functions defined on rectangular subspaces, such as in the case presented herein, these continuity conditions are easily accommodated and higher-order accuracy and function continuity are achieved. These higher-order continuity constraints may be circumvented by expressing higher-order equations as sets of lower-order equations, although additional field variables are thereby introduced.

The standard Galerkin method leads to 'central-type' discrete operators which exhibit oscillatory behaviour on practical meshes. A way to avoid this difficulty is to select a test function for the convective term that explicitly accommodates the directional property of the hyperbolic propagation (upwinding). Such methods are often denoted as Petrov–Galerkin methods. The main drawback of this approach is the absence of a generally applicable systematic procedure for the selection of the test function. Recently the Lax–Wendroff scheme has been used for developing the Taylor–Galerkin method. In addition, the method of characteristics has been used for developing the Galerkin method.

Finite element methods for solving first-order systems which are symmetric and positive in the sense of Friedrichs have been proposed by Lesaint and Raviart.[7] For symmetric systems arising out of the heat equation, Aziz and Liu[8] proposed a weighted least squares solution method. In their work a unique weight associated with the boundary conditions was determined in order to reduce the error estimates.

For mixed methods based on splines, the least squares theory for second-order elliptic systems developed in Reference 9 specifies the optimal weights in terms of the spline mesh size. Other first-order systems have also been solved numerically using a weighted least squares formulation where

optimal weights are found theoretically. For example, Wendland[10] formulates a least squares method for strong elliptic systems satisfying the Lopatinsky conditions for the boundary operator. For elliptic operators in the sense of Douglis and Nirenberg, wherein the operator satisfies the supplementary condition for the operator **L** and the complementary conditions for the boundary operator **B**, a weighted least squares method has been proposed by Aziz *et al.*[11]

A difficulty appearing in the implementation of the method, and one that is addressed in this paper, is the determination of the weights associated with the least squares method for a general first-order system. In most numerical implementations these weights must be specified by the analyst.[4] Without some rational criteria for selecting these weights, the efficiency of the method may be seriously affected even if other computational objectives are achieved.

The formulation proposed herein is similar to the least squares finite element method proposed by Jiang and Carey for linear and non-linear hyperbolic systems.[12–14] However, the method differs in as much as collocation points and weighting functions are introduced.

As noted earlier, the least squares method is validated on the two-dimensional shallow water equations. These equations are often used to obtain flow fields necessary for pollution problems and transport in shallow estuaries. A special advantage of using the collocation least squares method for solving the shallow water equations is the ability to solve the equations in an irregular domain with a completely orthogonal computational mesh.[15,16] This significantly enhances both the accuracy and computation time.

When the non-linear terms appearing in these equations are linearized, they reduce to a system of first-order equations. Classical collocation least squares has good noise control characteristics compared with the Galerkin method when solving the shallow water equations. However, one of the difficulties inhibiting its widespread use is that the accuracy of the solution depends upon the weights appearing in the formulation.[16]

3. A WEIGHTED LEAST SQUARES METHOD

A least squares formulation is proposed herein which takes into account the relative weights between the different first-order equations appearing in the system. The least squares formulation is presented in terms of the matrix of differential operators **L** and **B** defined in equations (1) and (2). These equations may be written as

$$\mathbf{Lu} = \mathbf{f} \quad \text{in } \Omega, \qquad \mathbf{Bu} = \mathbf{g} \quad \text{on } \Gamma. \tag{3}$$

In order to formulate the method proposed, let us now define the functional spaces and their associated norms. In the following we define the working space V_g as the p-product of the Sobolev space $H^l(\Omega)$:

$$V_g = \underbrace{H^1(\Omega) \times H^1(\Omega) \times \cdots \times H^1(\Omega)}_{p \text{ times}} = [H^1(\Omega)]^p,$$
$$H^1(\Omega) = \left\{ v \in L^2(\Omega); \frac{\partial v}{\partial x} \in L^2(\Omega); \frac{\partial v}{\partial y} \in L^2(\Omega) \right\}. \tag{4}$$

Notice that each element $\mathbf{u} \in V_g$ is a vector of p components u_i, $i = 1, 2, \ldots, p$.

If for any $\mathbf{u} \in V_g$ and $\mathbf{v} \in V_g$, $\mathbf{u} \cdot \mathbf{v}$ denotes the classical scalar product on $\mathbb{R}^p$, an L^2 scalar product may be defined on the domain Ω and the boundary Γ as

$$(\mathbf{u}, \mathbf{v}) = \int_\Omega \mathbf{u} \cdot \mathbf{v} d\Omega, \qquad [\mathbf{u}, \mathbf{v}] = \int_\Gamma \mathbf{u} \cdot \mathbf{v} d\Gamma. \tag{5}$$

In the following, $\|\cdot\|_\Omega$ denotes the norm associated with (., .) and $\|\cdot\|_\Gamma$ denotes the norm associated with [., .].

The L^2 weighted least squares formulation proposed in this work may be separated into two groups.

Group 1. L^2 continuous least squares

This formulation corresponds to a minimization problem with the choice of an L^2 (Ω)-norm for the interior domain and an L^2 (Γ)-type norm for the boundary conditions:

$$\underset{(\mathbf{v}\in V_g)}{\text{Min}}\, T(\mathbf{v}) = (\mathbf{Lv}-\mathbf{f},\ \mathbf{W}(\mathbf{Lv}-\mathbf{f})) + [\mathbf{Bv}-\mathbf{g},\ \mathbf{P}(\mathbf{Bv}-\mathbf{g})] \iff \underset{(\mathbf{v}\in V_g)}{\text{Min}} \sum_{i=1}^{p} w_i \int_\Omega \left(\sum_{j=1}^{p} L_{ij}v_j - f_i\right)^2 d\Omega + \sum_{i=1}^{p} p_i \int_\Gamma \left(\sum_{j=1}^{pb} B_{ij}v_j - g_i\right)^2 d\Gamma, \quad (6)$$

where **W** and **P** represent the diagonal matrices of positive weights associated respectively with each equation of the system (w_i) and each boundary condition (p_i).

In practice, problem (6) is never solved on a functional space of infinite dimension. A finite-dimensional approximation of problem (6) is generally solved by restricting the above functional to be on a finite-dimensional subspace of V_g. One such approach is based on finite element approximations.

With this approach the discretization procedure may be formulated as follows.

The closed domain Ω is approximated by a polygonal domain Ω_h with a standard triangulation $\mathcal{T}_h$ of Ω_h, i.e. $\mathcal{T}_h$ is a set of finite elements E.[17]

The Sobolev space H^1 (Ω) is then approximated on $\mathcal{T}_h$ by

$$H_h^1(\Omega_h) = \{v_h | v_h \in C^0(\bar{\Omega}_h);\ v_{h|E} \in P_1,\ \forall E \in \mathcal{T}_h\}, \quad (7)$$

where P_1 is the space of polynomials in two variables of degree less than or equal to one. In the treatment of parabolic problems written in terms of second-order operators, the least squares procedure requires a space of polynomials which are at least P_2. In this case the Hermite interpolation functions are generally used for the interpolation over a single element.[6,17]

The space V_h, the approximation of V_g, is then defned as the p-product of $H_h^1(\Omega_h)$. Each component v_{hj} of the vector function $\mathbf{v}_h \in V_h$ may be written in terms of the global interpolation function defined as

$$\mathbf{\Phi}^j(\mathbf{x}) = \begin{pmatrix} \phi_1^j(\mathbf{x}) \\ \vdots \\ \phi_{N_p^j}^j(\mathbf{x}) \end{pmatrix}, \quad (8)$$

$$\phi_i^j(\mathbf{x}) \in C^0(\bar{\Omega}_h), \quad (9)$$

where N_p^j is the total number of unknowns for the function $v_j(\mathbf{x})$.

Formally, any function $v_{hj} \in H_h^1(\Omega_h)$ may be written in terms of the vector of global interpolation functions and the vector of unknown parameters, $\mathbf{a}^j$, as

$$v_{hj}(\mathbf{x}) = \sum_{E\in\mathcal{T}_h} \sum_{i=1}^{N_T} \phi_{i|E}^j(\mathbf{x}) a_i^j = \lfloor \mathbf{\Phi}^j(\mathbf{x}) \rfloor \mathbf{a}^j, \quad (10)$$

where N_T is the number of unknowns on the element E. The independent space variable of the domain is represented by $\mathbf{x}$. The L^2 approximate continuous least squares formulation consists of the minimization of a functional $I_c(\mathbf{a}, \mathbf{W}, \mathbf{P})$ containing a weighted sum of the interior and boundary residuals with respect to the unknown vectors $\mathbf{a}^j$.

The interior residual R_{Li} corresponding to the ith equation of the system and the boundary residuals R_{Bi} corresponding to the ith boundary condition are defined by

$$\begin{aligned} R_{Li}(\mathbf{a}, \mathbf{x}) &= \sum_{j=1}^{p} L_{ij} v_{hj}(\mathbf{x}) - f_i, \quad \mathbf{x} \text{ in } \Omega, \\ R_{Bi}(\mathbf{a}, \mathbf{x}) &= \sum_{j=1}^{pb} B_{ij} v_{hj}(\mathbf{x}) - g_i, \quad \mathbf{x} \text{ on } \Gamma. \end{aligned} \tag{11}$$

Consider now the global unknown vector $\mathbf{a}$ formed by the set of all the vectors $\mathbf{a}^j$, $j = 1, \ldots, p$. The objective function $I_c(\mathbf{a}, \mathbf{W}, \mathbf{P})$ to minimize may be defined as

$$I_c(\mathbf{a}, \mathbf{W}, \mathbf{P}) = \sum_{i=1}^{p} w_i \int_{\Omega} [R_{Li}(\mathbf{a}, \mathbf{x})]^2 d\Omega + \sum_{i=1}^{pb} p_i \int_{\Gamma} [R_{Bi}(\mathbf{a}, \mathbf{x})]^2 d\Gamma. \tag{12}$$

The approximate L^2 continuous least squares formulation corresponding to equation (6) may be defined as

$$\underset{(v_h \in V_h)}{\text{Min}} I_c(\mathbf{a}, \mathbf{W}, \mathbf{P}) . \tag{13}$$

In this general framework an error analysis of the least squares approximation has been conducted for particular problems in fluid dynamics.[1,4,5,18] The choice of a suitable norm $\| \cdot \|_\Omega$ on V_g depends on the type of differential operator. For example, the choice of an H^1-norm instead of the classical L^2-norm stabilized the least squares solution of non-linear hyperbolic problems which generate shocks.[13]

Group 2. L^2 collocation (or discrete) least squares

The discrete formulation corresponding to the continuous least squares method, also called collocation least squares, consists of selecting a series of collocation points inside the domain and on the boundary and minimizing the function

$$I_d(\mathbf{a}, \mathbf{W}, \mathbf{P}) = \sum_{i=1}^{p} w_i \sum_{l=1}^{k} c_l [R_{Li}(\mathbf{a}, \mathbf{x}_l)]^2 d\Omega + \sum_{i=1}^{pb} p_i \sum_{l=k+1}^{m} c_l [R_{Bi}(\mathbf{a}, \mathbf{x}_l)]^2 d\Gamma. \tag{14}$$

The points $\mathbf{x}_l$ *for* $l = 1, \ldots, k$ correspond to the interior points and for $l = k+1, \ldots, m$ to the boundary points of Φ.[4] The weights c_l are associated with the collocation points. One of the difficulties appearing in the implementation of the method is the determination of the weights w_i, $i = 1, \ldots, p$, the weights p_i, $i = 1, \ldots, p$, and the weights c_l, $l = 1, \ldots, m$.

The addition of the integral of the boundary residual may result in a solution vector with large interior residuals. This scaling problem may degrade the satisfaction of the governing balance laws. This is particularly true when the weights appearing in the matrix $\mathbf{P}$ are very large.

Basically, the three types of weights introduced (w_i, p_i and c_l) are each of a different nature. The first and second types of weights, w_i, $i = 1, \ldots, p$, and p_i, $i = 1, \ldots, p_b$, determine the relative contributions of the interior and boundary residuals. When a system of equations is solved using an L^2 least squares method, these weights will balance the different contributions of each equation and each boundary condition. These weights are generally specified by the analyst.

The third type of weights, c_l, $i = 1, \ldots, m$, represent the distribution of weights among the collocation points. These weights may be determined by using optimal criteria for numerical integration, such as Gaussian quadrature. For a collocation least squares method using finite element discretization, the use of Gaussian points and their weights leads to an optimal truncation error. More sophisticated criteria have been proposed by the authors.[19]

In the present work we concentrate on the first and second types of weights and a general methodology is developed for computation of the optimal values of these weights.

4. PROPERTIES OF THE L^2 LEAST SQUARES

In the least squares literature two different methods have been proposed for computing the numerical solution of the minimization problem (problem (13) here). The first consists of using an unconstrained optimization algorithm for the quadratic functional I_c or I_d. Classical methods such as conjugate gradient or quasi-Newton generally require the computation of both a functional and its Jacobian. The second approach consists of solving the normal equation coresponding to a first-order necessary condition for a minimum of the functional.

In the present work we deduce the normal equation from the multidimensional minimization problem (equation (6)) rather than from the finite-dimensional problem as is usually done. Then we discuss the basic properties of the least squares method.

The following theorem gives a necessary condition for a minimum for the least squares and approximate least squares methods.

Theorem 1

If $\tilde{\mathbf{u}}$, resp. $\mathbf{u}_h$, is the vector solution of the minimization problem (6), resp. (13), then for all $\mathbf{v} \in V_g$, resp. $\mathbf{v}_h \in V_h$.

$$(\mathbf{L}\tilde{\mathbf{u}} - \mathbf{f},\ \mathbf{W}(\mathbf{L}\mathbf{v})) + [\mathbf{B}\tilde{\mathbf{u}} - \mathbf{g},\ \mathbf{P}(\mathbf{B}\mathbf{v})] = 0, \tag{15}$$

$$(\mathbf{L}\mathbf{u}_h - \mathbf{f},\ \mathbf{W}(\mathbf{L}\mathbf{u}_h)) + [\mathbf{B}\mathbf{u}_h - \mathbf{g},\ \mathbf{P}(\mathbf{B}\mathbf{v}_h)] = 0. \tag{16}$$

Proof. Consider the least squares functional $J(\mathbf{v}) = (\mathbf{L}\mathbf{v} - \mathbf{f}, \mathbf{W}(\mathbf{L}\mathbf{v} - \mathbf{f})) + [\mathbf{B}\mathbf{v} - \mathbf{g}, \mathbf{P}(\mathbf{B}\mathbf{v} - \mathbf{g})]$. If $\tilde{\mathbf{u}}$ is a solution of equation (15), then $\forall \mathbf{v} \in V_g$

$$\lim_{t \to 0} \frac{J(\tilde{\mathbf{u}} + \mathbf{t}\mathbf{v}) - J(\tilde{\mathbf{u}})}{\mathbf{t}} \geq 0. \tag{17}$$

The Taylor expansion of $J(\tilde{\mathbf{u}} + t\mathbf{v})$ may be written as

$$J(\tilde{\mathbf{u}} + \mathbf{t}\mathbf{v}) = \mathbf{J}(\tilde{\mathbf{u}}) + \mathbf{t}\{(\mathbf{L}\tilde{\mathbf{u}} - \mathbf{f},\ \mathbf{W}(\mathbf{L}\mathbf{v})) + [\mathbf{B}\tilde{\mathbf{u}} - \mathbf{g},\ \mathbf{P}(\mathbf{B}\mathbf{v})]\} + O(t^2).$$

Then for all $\mathbf{v} \in V_g$ and for $t > 0$ the above limit leads to

$$(\mathbf{L}\tilde{\mathbf{u}} - \mathbf{f},\ \mathbf{W}(\mathbf{L}\mathbf{v})) + [\mathbf{B}\tilde{\mathbf{u}} - \mathbf{g},\ \mathbf{P}(\mathbf{B}\mathbf{v})] \geq 0. \tag{18}$$

Taking now $-\mathbf{v}$ instead of $\mathbf{v}$ in the last inequality, we see that we obtain (15).

The same proof may be developed for the approximate least squares problem (equation (16)). □

Equation (16) may be written as a linear system of algebraic equations with unknown vector **a**. The solution vector $\mathbf{u}_h$ and the weighting function vector $\mathbf{v}_h$ may be expressed in terms of their components as

$$u_{hj}(\mathbf{x}) = \lfloor \Phi^j(\mathbf{x}) \rfloor \mathbf{a}^j, \qquad v_{hj}(\mathbf{x}) = \lfloor \Phi^j(\mathbf{x}) \rfloor \mathbf{d}^j. \tag{19}$$

Equation (16) may be written in terms of the components of the differential operators **L** and **B** as

$$\sum_{i=1}^{p} w_i \int_\Omega \left(\sum_{j=1}^{p} L_{ij}(\lfloor \Phi^j(\mathbf{x}) \rfloor)\mathbf{a}^j - f_i \right) \left(\sum_{l=1}^{p} L_{il}(\lfloor \Phi^l(\mathbf{x}) \rfloor \mathbf{d}^l \right) d\Omega + \sum_{i=1}^{pb} p_i \int_\Gamma \left(\sum_{j=1}^{pb} B_{ij}(\lfloor \Phi^j(\mathbf{x}) \rfloor)\mathbf{a}^j - g_i \right) \left(\sum_{l=1}^{pb} B_{il}(\lfloor \Phi^l(\mathbf{x}) \rfloor)\mathbf{d}^l \right) d\Gamma = 0. \tag{20}$$

Let us assume that for any $j = 1, \ldots, p$ the dimension of the column vector $\mathbf{a}^j$ is independent of j and equals D. Introduce now the $N^D \times pN^D$ matrices $\mathbf{I}^j$ and the two multipliers $\mathbf{L}_i(\mathbf{x})$ and $\mathbf{B}i(\mathbf{x})$ defined as

$$\mathbf{a}^j = \mathbf{I}^j \mathbf{a}, \qquad \mathbf{L}_i(\mathbf{x}) = \sum_{j=1}^{p} L_{ij}(\lfloor \Phi^j(\mathbf{x}) \rfloor)\mathbf{I}^j, \qquad \mathbf{b}_i(\mathbf{x}) = \sum_{j=1}^{pb} B_{ij}(\lfloor \Phi^j(\mathbf{x}) \rfloor)\mathbf{I}^j. \tag{21}$$

With this notation equation (16) may be written in matrix form as

$$\left[\left(\sum_{i=1}^{p} w_i \mathbf{A}_i \right) \mathbf{a} - \sum_{i=1}^{p} w_i \mathbf{f}_i \right] \cdot \mathbf{d} + \left[\left(\sum_{i=1}^{pb} p_i \mathbf{B}_i \right) \mathbf{a} - \sum_{i=1}^{pb} p_i \mathbf{g}_i \right] \cdot \mathbf{d} = 0, \tag{22}$$

where the matrices $\mathbf{A}_i$ and $\mathbf{B}_i$ and the vectors $\mathbf{f}_i$ and $\mathbf{g}_i$ are defined for continuous least squares as

$$\mathbf{A}_i = \int_\Omega \mathbf{L}_i^L(\mathbf{x})\mathbf{L}_i(\mathbf{x})d\mathbf{x}, \qquad \mathbf{B}_i = \int_\Gamma \mathbf{b}_i^{\mathrm{T}}(\mathbf{x})\mathbf{b}_i(\mathbf{x})d\mathbf{x}, \qquad \mathbf{f}_i = \int_\Omega \mathbf{L}_i^{\mathrm{T}}(\mathbf{x})\mathbf{f}_i d\mathbf{x}, \qquad \mathbf{g}_i = \int_\Gamma \mathbf{b}_i^{\mathrm{T}}(\mathbf{x})\mathbf{g}_i d\mathbf{x} \tag{23}$$

and for collocation least squares as

$$\mathbf{A}_i = \sum_{j=1}^{k} \mathbf{v}_j \mathbf{L}_i^T(\mathbf{x})\mathbf{L}_i(\mathbf{x}_j), \qquad \mathbf{B}_i = \sum_{j=k+1}^{m} \mathbf{b}_i^{\mathrm{T}}(\mathbf{X}_j)\mathbf{b}_i(\mathbf{x}_j), \qquad \mathbf{f}_i = \sum_{j=1}^{k} \mathbf{L}_i^{\mathrm{T}}(\mathbf{x}_j)\mathbf{f}_i, \qquad \mathbf{g}_i = \sum_{j=k+1}^{m} \mathbf{b}_i^T(\mathbf{x}_i)\mathbf{g}_i. \tag{24}$$

Equation (22) is valid for any vector **d** and may therefore be written in the classical form of the normal equation, i.e.

$$\left(\sum_{i=1}^{p} w_i \mathbf{A}_i + \sum_{i=1}^{pb} p_i \mathbf{B}_i \right) \mathbf{a} = \sum_{i=1}^{p} w_i \mathbf{f}_i + \sum_{i=1}^{pb} p_i \mathbf{g}_i. \tag{25}$$

Without loss of generality, equation (25) may be scaled such that the first weights w_1 will be equal t one. Defining the new weights $\epsilon_i = 1/w_i$, $i = 2, \ldots, p$, and $v_j = 1/w_j$, $j = 1, \ldots, p_b$, equation (25) may now be expressed as

$$\left(\mathbf{A}_1 + \sum_{i=2}^{p} \frac{1}{\epsilon_i} \mathbf{A}_i + \sum_{i=1}^{pb} \frac{1}{v_i} \mathbf{B}_i \right) \mathbf{a} = \mathbf{f}_1 + \sum_{i=1}^{p} \frac{1}{\epsilon_i} \mathbf{f}_i + \sum_{i=1}^{pb} \frac{1}{v_i} \mathbf{g}_i. \tag{26}$$

The choice of the optimal relative weightings depends on the criterion of optimality defined. For the numerical least squares method presented in this contribution, these criteria may be separated into four categories:

(a) accuracy of the numerical scheme
(b) numerical stability of the linear system resulting from the normal equation (25)
(c) satisfaction of the boundary conditions
(d) conservation of the mass balance.

The dependence of the weighting on the overall accuracy of the numerical scheme is still a subject of research. Error estimates using different norms on suitable Sobolev spaces have been derived by several authors for the penalty method (see e.g. Reference 9). Compared with our formulation, the classical penalty method corresponds to a system with two equations and a single weight. The optimal weight is chosen for the reduction of the error estimation. This weight depends on the type of partial differential equation, the discretization parameter and the type of boundary conditions. Such analysis shows that a better theoretical error estimate is obtained when the weight is reduced.

It is well known that for a large weight the condition number of the normal equation resulting from the penalty least squares method depends linearly on this weight.[1] According to this criterion, one has to increase the weight for a good stability of the normal equation.

Thus the weighting strategy for criterion (a) is the opposite of that for criterion (b).

Among the basic properties we may require from a numerical method, global conservation of the system is of primary importance.[3] This property depends on the physical problem to be solved and on the type of differential operator. It may be stated that the flux, the energy or the forces will be conserved globally inside Ω. For the first-order hyperbolic system considered here, the global conservation property may be written for each equation of the system as follows:

$$\text{for } i = 1, \ldots, p, \quad \int_\Omega \left(\sum_{j=1}^{p} L_{ij} v_j - f_i \right) \mathrm{d}\Omega = 0. \tag{27}$$

The global conservation property is not always satisfied by the numerical solution of the weighted least squares formulation presented herein. On the basis of Lagrangian multiplier functions, the order of accuracy of the global conservation property is analysed. The following theorem presents an error estimate for the global conservation property for each equation of the first-order system.

Theorem 2

If $\mathbf{a}$ is the vector of dimension pN^D which is the solution of equation (26), then we have the following.

(i) For the continuous least squares method the following properties hold:

$$\begin{aligned} &\text{for } i = 1, \ldots, p, \quad \int_\Omega \left(\sum_{j=1}^{p} L_{ij} u_{hj} - f_i \right) \mathrm{d}\Omega = H_i(\mathbf{x}) \epsilon_i + O(\epsilon_i^2), \\ &\text{for } i = 1, \ldots, pb, \quad \sum_{j=1}^{pb} B_{ij} u_{hj} - g_i = D_i v_i + O(v_i^2). \end{aligned} \tag{28}$$

(ii) For the collocation least squares method the following properties hold:

$$\begin{aligned} &\text{for } i = 1, \ldots, p, \quad \sum_{l=1}^{k} v_l \left(\sum_{j=1}^{p} L_{ij} u_{hj}(\mathbf{x}_l) - f_i \right) d\Omega = H_i(\mathbf{x}) \epsilon_i + O(\epsilon_i^2), \\ &\text{for } i = 1, \ldots, pb \quad \text{for } l = k+1, \ldots, m, \quad \sum_{j=1}^{pb} B_{ij} u_{hj}(\mathbf{x}_l) - g_i = D_i(\mathbf{x}) v_i + O(v_i^2). \end{aligned} \tag{29}$$

Here the function $H_i(\mathbf{x})$ does not depend on ϵ_i and D_i does not depend on v_i.

Proof. The proof is developed in the Appendix. □

For first-order systems a rigorous study of the dependence of the weights on the accuracy of our least squares formulation is outside the scope of this contribution. However, Theorem 2 points out the difficulty in choosing the relative weights appearing in the least squares functional between the different equations.

In terms of global mass balance conservation, the weighting strategy is to increase all the weight appearing in equations (28) and (29). However, in terms of conditioning of the matrix, the weighting strategy is to reduce these weights. For a given equation (i) the choice of a small value for one of the ϵ_i, $i = 2, \ldots, p$, will improve the global conservation property for this equation. A good strategy should be to choose large weights for each equation, but unfortunately this will cause ill conditioning in the linear system of equations (26).

So far the goal of Theorem 2 and the above discussion was to justify the need for the definition of a strategy of optimal weighting.

In the following a criterion of optimality based on energy balance considerations is introduced for the computation of these weights.

The optimal weights are those for which the corresponding least squares solution respects more accurately the mechanical energy balance of the system. This balance equation corresponds to the natural proportions between the different equations.

For the first-order hyperbolic system considered here, one may require a conservation property of the type

$$\sum_{i=1}^{p} \int_{\Omega} \left(\sum_{j=1}^{p} L_{ij} v_j - f_i \right) v_i = 0, \tag{30}$$

which may be written in a functional form as

$$\Phi_a(\mathbf{v}) = 0. \tag{31}$$

The property of global conservation is implicitly verified in the Galerkin method but is not respected by the formulation (6).

The basic idea of the new least squares formulation proposed is to determine the optimal set of weights w_i, $i = 1, \ldots, p$, and p_i, $i = 1, \ldots, p$, in order that the least squares solution will respect approximately the global conservation law $\Psi_a(\mathbf{v}) = 0$. In the case where $\Psi_a(v)$ is a positive function, the determination of the optimal set of weights may be achieved by minimizing the function $\Psi_a(\mathbf{v})$ with respect to the weights.

5. OPTIMIZATION ALGORITHM

5.1. General methodology

The basic steps of the methodology proposed are as follows.
Consider after k iterations the set of known weights $\mathbf{W}^k$ and $\mathbf{P}^k$.

Step 1. Determine $\mathbf{u}_h(\mathbf{x}, \mathbf{W}^k, \mathbf{P}^k) = \text{ArgMin } J(\mathbf{v})$, the solution of the minimization problem (6).

Step 2. For a positive function $\Psi_a(\mathbf{v})$, determine the new matrices with positive coefficients, $\mathbf{W}^{(k+1)}$ and $\mathbf{P}^{(k+1)}$, which are obtained via solution of the minimization problem with respect to the weights:

$$\text{Min}\Psi_a(\mathbf{u}_h, \mathbf{W}^{(k+1)}, \mathbf{P}^{(k+1)}) \quad (w_i > 0;\ p_i > 0;\ i = 1, \ldots, p).$$

Step 3

If $|\Psi_a(\mathbf{u}_h, \mathbf{W}^{(k+1)}, \mathbf{P}^{(k+1)})| < \epsilon$ then END
If not
Set $\mathbf{W}^k = \mathbf{W}^{(k+1)}$ and $\mathbf{P}^k = \mathbf{P}^{(k+1)}$
GO TO Step 1.

In the next section the algorithm developed for the solution of the shallow water equations is presented.

5.2. Methods and procedures

Here we describe the process for determining the weights used in the present least squares method. An associate norm derived from an auxiliary finite element formulation is first computed. Then the weights are determined in order to force the auxiliary norm to a minimum. In the following we will need to distinguish between the discretized form of the differential equations as derived from the least squares method and the discretized form of the differential equations as derived by some other finite element procedure. We will denote the discretized least squares formulation equation (25) as

$$\mathbf{Ax} - \mathbf{f} = \mathbf{0}, \tag{32}$$

where

$$\mathbf{A} = \sum_{i=1}^{p}(w_i\mathbf{A}_i + p_i\mathbf{B}_i), \qquad \mathbf{x} = \mathbf{a}, \qquad \mathbf{f} = \sum_{i=1}^{p}(w_i\mathbf{f}_i + p_i\mathbf{g}_i), \tag{33}$$

and the discretized auxiliary formulation as

$$\mathbf{A}_a\mathbf{x}_a - \mathbf{f}_a = \mathbf{0}. \tag{34}$$

One possible auxiliary norm may be defined as

$$\Phi_a = [\mathbf{A}_a\mathbf{x}(\mathbf{S}) - \mathbf{f}_a]^T[\mathbf{A}_a\mathbf{x}(\mathbf{S}) - \mathbf{f}_a], \tag{35}$$

where $\mathbf{x}(\mathbf{S})$ represents the least squares solution for a given set of weights $\mathbf{S} = (s_1, s_2, \ldots, s_{2n}) = (w_1, w_2, \ldots, w_n, p_1, \ldots, p_n)$ and T denotes the transposed vector. Notice that $\mathbf{x}(\mathbf{S})$ is not the solution of the auxiliary formulation. Since $\mathbf{A}$ and $\mathbf{f}$ are dependent on the weights used in the least squares formulation, $\mathbf{x}$ is also a function of the weights. Our minimization problem may thus be

stated as

$$\underset{s_1,s_2,\cdots,s_{2n}}{\text{Min}} \Phi_a(\mathbf{x}(\mathbf{S}))\text{subject to } \mathbf{A}(\mathbf{S})\mathbf{x}(\mathbf{S}) - \mathbf{f}(\mathbf{S}) = \mathbf{0}, \quad s_i > 0. \tag{36}$$

Here we detail a procedure to solve this problem and contrast this approach with the well-known Newton–Raphson and quasi-Newton methods. Consider a first-order Taylor expansion of $\mathbf{x}(\mathbf{S})$ around $\mathbf{S}$:

$$\Phi_a = [\mathbf{A}_a(\mathbf{x} + \nabla_s\mathbf{x} \cdot (\Delta\mathbf{S})) - \mathbf{f}_a] \cdot [\mathbf{A}_a(\mathbf{x} + \nabla_s\mathbf{x} \cdot (\Delta\mathbf{S})) - \mathbf{f}_a] + O(\|\Delta\mathbf{S}\|^2), \tag{37}$$

where $\Gamma_s\mathbf{x}$ represents the gradient vector of the vector $\mathbf{x}$ with respect to the weights (also called the sensitivity matrix). Applying the minimization criterion $\partial\Psi_a/\partial\Delta\mathbf{S} = \mathbf{0}$ as in a least squares problem, we obtain

$$[\mathbf{G}^T\mathbf{A}_a^T\mathbf{A}_a\mathbf{G}](\Delta\mathbf{S}) = -[\mathbf{G}^T\mathbf{A}_a^T]\{\mathbf{A}_a\mathbf{x} - \mathbf{f}_a\}, \tag{38}$$

where $\mathbf{G} = \Gamma_s\mathbf{x}$ is the sensitivity matrix. This matrix is obtained by differentiation with respect to each weight s_i of the least squares equation:

$$\frac{\partial}{\partial s_i}[\mathbf{A}\mathbf{x} - \mathbf{f}]. \tag{39}$$

Expanding this last equation and solving for $\partial x/\partial s_i$ yields

$$\mathbf{g}_i = \frac{\partial\mathbf{x}}{\partial s_i} = \mathbf{A}^{-1}\left\{\frac{\partial\mathbf{f}}{\partial s_i} - \left[\frac{\partial\mathbf{A}}{\partial s_i}\right]\mathbf{x}\right\}. \tag{40}$$

The sensitivity matrix $\mathbf{G}$ is thus defined as $\mathbf{G} = [\mathbf{g}_1|\mathbf{g}_2|\mathbf{g}_3|\cdots|\mathbf{g}_n]$, where n is the number of weights. The steps of the procedure are as follows.

1. Assume a set of weights $\mathbf{S}$.
2. Solve $\mathbf{A}\mathbf{x}-\mathbf{f} = \mathbf{0}$.
3. Employ equation (40) to obtain $\mathbf{G}$.
4. Form and solve equation (38) to obtain $\Delta\mathbf{S}$.
5. If $\|\Delta\mathbf{S}\| < \epsilon_{\text{tol}}$, end; else $\mathbf{S}^{\text{new}} = \mathbf{S}^{\text{old}} + \alpha\{\Delta\mathbf{S}\}$ and go to step 2.

Alternatively one may expand Ψ_a around the vector $\mathbf{S}_0$ by Taylor's theorem to obtain

$$\Phi_a = \Phi_a(\mathbf{S}_0) + \nabla_s(\Phi_a) \cdot \{\Delta\mathbf{S}\} + \tfrac{1}{2}\{\Delta\mathbf{S}\} \cdot \mathbf{H}_s\{\Delta\mathbf{S}\} + O(\|\Delta\mathbf{S}\|^2), \tag{41}$$

where $\mathbf{H}_s$ is the Hessian matrix of Ψ_a with respect to the weights and $\Delta\mathbf{S} = \mathbf{S} - \mathbf{S}_0$. The minimum of Ψ_a is obtained from $\partial\Psi_a/\partial\Delta\mathbf{S} = \mathbf{0}$. This yields

$$\mathbf{H}_s\{\Delta\mathbf{S}\} = -\nabla_s(\Phi_a) = \mathbf{G}^T\mathbf{A}_a^T[\mathbf{A}_a\mathbf{x} - \mathbf{f}_a] \tag{42}$$

This last equation is the basis for the second-order Newton–Raphson methods. It can be shown that the exact expression for the second-order derivatives is

$$\mathbf{H}_w = [\mathbf{G}^T\mathbf{A}_a^T\mathbf{A}_a\mathbf{G}] + \left[\left[\frac{\partial\mathbf{G}}{\partial\mathbf{S}}\right]\right][\mathbf{A}_a^T(\mathbf{A}_a\mathbf{x} - \mathbf{f}_a)], \tag{43}$$

where $[[\partial\mathbf{G}/\partial\mathbf{S}]]$ is a third-order tensor. For the problem under consideration this tensor can be determined analytically but not without considerable computational expense. To improve on the inverse of the Hessian matrix, one can employ the BFGS methods readily available in the IMSL Fortran Library.

6. HYPERBOLIC SHALLOW WATER EQUATIONS

Here we discuss a specific case of the general methodology defined in Section 4. The specific equations used are the shallow water equations, which may be written in scalar form as

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} + \frac{C_b}{H}u - fv + g\frac{\partial \zeta}{\partial x} - \frac{\tau_x}{H} = 0, \tag{44}$$

$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} + \frac{C_b}{H}v + fu + g\frac{\partial \zeta}{\partial y} - \frac{\tau_y}{H} = 0, \tag{45}$$

$$\frac{\partial \zeta}{\partial t} + \frac{\partial (Hu)}{\partial x} + \frac{\partial (Hv)}{\partial y} - r = 0, \tag{46}$$

$$\zeta - \zeta_p = 0, \tag{47}$$

$$l_x u + l_y v = v_n, \tag{48}$$

$$H = h + \zeta. \tag{49}$$

The first two equations are the x- and y-momentum equations respectively. The third equation is the fluid continuity equation. These equations are valid over the region Ω. The fourth equation is a prescribed surface elevation to be enforced on a portion Φ_1 of the boundary and the fifth equation is a prescribed normal flow to be applied on a portion Φ_2 of the boundary. The last equation defines the total depth. For problems with moving boundaries this equation becomes a third-type boundary condition. Here we will restrict our attention to problems of constant geometry. In these equations x and y are Cartesian co-ordinates, t is time, g is gravity, f is the Coriolis parameter, C_b is the bottom friction parameter, u and v are x and y vertically integrated velocities respectively, ζ is the surface elevation, h is the water depth from mean sea level, ζ_p is the prescribed surface elevation, v_n is the prescribed normal velocity, l_x and l_y are the direction cosines of the outwardly directed unit normal on the boundary, τ_x and τ_y represent the x- and y-components of the wind shear stress respectively and r is the fluid source. The bottom friction parameter may take on various forms and is generally dependent on u and v.

These equations in their full non-linear form have been solved by the least squares method on irregular domains developed by Laible and Pinder.[15] Since we wish to use a test problem that has an analytic solution, we will focus on the linearized form of these equations.

Starting with the scalar equations, we now seek to bring the shallow water equations into the form of equations (1) and (2). First we note that the continuity equations can be expanded in terms of h and ξ. Inserting the equality $H = h + \xi$ into the continuity equation, the following continuity equation may be obtained:

$$\frac{\partial \xi}{\partial t} + u\frac{\partial \xi}{\partial x} + v\frac{\partial \xi}{\partial y} + \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right)\xi + h\frac{\partial u}{\partial x} + h\frac{\partial v}{\partial y} + \frac{\partial h}{\partial x}u + \frac{\partial h}{\partial y}v - \gamma = 0. \tag{50}$$

This equation is now in terms of the first-order derivatives of the unknowns u, v and ξ.

Equations (44)–(46) are non-linear. We now seek to obtain a linearized form. Here we introduce a modification (similar to the idea of Laplace modification used to solve problems with time-dependent coefficients). This modification is applied to all the non-linear terms of equations (44)–(46). Suppose we select some time-invariant characteristic values for u, v and h denoted as U, V and Hi. These values must generally be greater than any anticipated values of u, v and h respectively. The modification is accomplished by writing an identity equation for each of equations (44)–(46) containing the non-linear

terms. To illustrate the process, we will consider equation (44) and deduce the remaining 'Laplace-modified' equations. The identity for equation (44) may be written as

$$U\frac{\partial u}{\partial x}+V\frac{\partial u}{\partial y}+\frac{C_b}{Hi}u=U\frac{\partial u}{\partial x}+V\frac{\partial u}{\partial y}+\frac{C_b}{Hi}u. \tag{51}$$

The wind stress term τ_x/H and C_b are taken here to be independent of u, v and ξ, although they could also be similarly treated.

If we now subtract this identity from equation (44), one obtains

$$\frac{\partial u}{\partial t}+U\frac{\partial u}{\partial x}+V\frac{\partial u}{\partial y}+\frac{C_b}{H}u-fv+g\frac{\partial \zeta}{\partial x} \tag{52}$$

$$=(U-u)\frac{\partial u}{\partial x}+(V-v)\frac{\partial u}{\partial y}+\left(\frac{C_b}{Hi}-\frac{C_b}{H}\right)u+\frac{\tau_x}{H}. \tag{53}$$

This process itself does not introduce any numerical approximations. In the numerical formulation, however, the left-hand side now contains a linear operator with constant coefficients. In the solution process values of u, v and H are assumed known (from a previous time step or by extrapolation from previous values). Therefore the right-hand side will contain known values. After solution for u, v and ξ the right-hand side is updated and the system solved again. Iteration within a time step is carried out until some norm of the variation of u, v and ξ meets a convergence tolerance. In the following we will denote the assumed or extrapolated values for u, v and ξ on the right-hand side as u^*, v^* and ξ^* (also $H^*=h+\xi^*$).

Applying this modification to equations (45) and (46), we finally have

$$\frac{\partial u}{\partial t}+U\frac{\partial u}{\partial x}+V\frac{\partial u}{\partial y}+\frac{C_b}{H}u-fv+g\frac{\partial \zeta}{\partial x}=\hat{F}_u, \tag{54}$$

$$\frac{\partial v}{\partial t}+U\frac{\partial v}{\partial x}+v\frac{\partial v}{\partial y}+\frac{C_b}{H}v+fu+g\frac{\partial \zeta}{\partial y}=\hat{F}_v, \tag{55}$$

$$\frac{\partial \xi}{\partial t}+U\frac{\partial \xi}{\partial x}+v\frac{\partial \xi}{\partial y}+\left(\frac{\partial U}{\partial x}+\frac{\partial V}{\partial y}\right)\xi+h\frac{\partial u}{\partial x}+h\frac{\partial v}{\partial y}+\frac{\partial h}{\partial x}u+\frac{\partial h}{\partial y}v=\hat{F}_\xi, \tag{56}$$

where

$$\hat{F}_u=(U-u^*)\frac{\partial u^*}{\partial x}+(V-v^*)\frac{\partial u^*}{\partial y}+\left(\frac{C_b}{H}-\frac{C_b^*}{H^*}\right)u^*+\frac{\tau_x}{H^*} \tag{57}$$

$$\hat{F}_v=(U-u^*)\frac{\partial v^*}{\partial x}+(V-v^*)\frac{\partial v^*}{\partial y}+\left(\frac{C_b}{H}-\frac{C_b^*}{H^*}\right)v^*+\frac{\tau_x}{H^*} \tag{58}$$

$$\hat{F}_\xi=(U-u^*)\frac{\partial \xi^*}{\partial x}+(V-v^*)\frac{\partial \xi^*}{\partial y}+\left[\left(\frac{\partial U}{\partial x}+\frac{\partial V}{\partial y}\right)-\left(\frac{\partial u^*}{\partial x}+\frac{\partial v^*}{\partial y}\right)\right]\xi^*+\gamma. \tag{59}$$

We now introduce the time discretization scheme. The local time derivative term is approximated by a backward difference, e.g. $\partial u/\partial t\approx(u^+-u^-)/\Delta t$. The variables u, v, ξ, u^*, v^+ and ξ^* are evaluated at an intermediate time between t and $t+\Delta t$ by the trapezoidal rule, e.g. $u=\alpha u^+ +(1-\alpha)u^-$. Here u^+ is

at time $t + \Delta t$ and u^- is at time t. With these approximations we may finally collect all terms that multiply u, $\partial u/\partial x$ and $\partial u/\partial y$ as in the standard form of equation (1). The vector $\mathbf{u}$ is now taken as

$$\mathbf{u} = \begin{Bmatrix} u \\ v \\ \xi \end{Bmatrix}. \tag{60}$$

After some rearrangement we find

$$\mathbf{A}_\alpha \frac{\partial \mathbf{u}^+}{\partial x} + \mathbf{D}_\alpha \frac{\partial \mathbf{u}^+}{\partial y} + \mathbf{C}_\alpha \mathbf{u}^+ = \mathbf{f}, \tag{61}$$

where

$$\mathbf{A}_\alpha = \begin{bmatrix} \alpha U & 0 & \alpha g \\ 0 & \alpha U & 0 \\ \alpha h & 0 & \alpha U \end{bmatrix}, \qquad \mathbf{D}_\alpha = \begin{bmatrix} \alpha V & 0 & 0 \\ 0 & \alpha V & \alpha g \\ 0 & \alpha h & \alpha V \end{bmatrix},$$

$$\mathbf{C}_\alpha = \begin{bmatrix} \dfrac{1}{\Delta t} + \dfrac{C_b}{H}\alpha & -\alpha f & 0 \\ \alpha f & \dfrac{1}{\Delta t} + \dfrac{C_b}{H}\alpha & 0 \\ \alpha \dfrac{\partial h}{\partial x} & \alpha \dfrac{\partial h}{\partial y} & \dfrac{1}{\Delta t} + \alpha\left(\dfrac{\partial U}{\partial x} + \dfrac{\partial V}{\partial y}\right) \end{bmatrix},$$

$$\mathbf{F} = \begin{Bmatrix} \hat{F}_u \\ \hat{F}_v \\ \hat{F}_\xi \end{Bmatrix} + \mathbf{A}_\beta \frac{\partial \mathbf{u}^-}{\partial x} + \mathbf{D}_\beta \frac{\partial \mathbf{u}^-}{\partial y} + \mathbf{C}_\beta \mathbf{u}^-. \tag{62}$$

The matrices $[\mathbf{A}]_\beta$, $[\mathbf{D}]_\beta$ and $[\mathbf{C}]_\beta$ are identical to $[\mathbf{A}]_\alpha$, $[\mathbf{B}]_\alpha$ and $[\mathbf{C}]_\alpha$ except that α is replaced by $\beta = (\alpha - 1)$.

The boundary conditions are also written in the standard form (equation (2)) as

$$\begin{bmatrix} 0 & 0 & 1 \\ n_x & n_y & 0 \end{bmatrix} \begin{Bmatrix} u^+ \\ v^+ \\ \xi^+ \end{Bmatrix} = \begin{Bmatrix} \xi \\ v_p \end{Bmatrix}.$$

Residuals

It is now possible to express the residuals of the differential equations and boundary conditions in terms of the standard matrices [A], [D], [C] and [M]. Here we use the matrix notation dropping the subscript α. To develop the residuals due to time and space discretization, we now introduce the spatial approximations. We introduce basis function expansions of our unknowns with local support over rectangular finite elements:

$$\{\mathbf{u}\} \approx [\Phi]\{\bar{\mathbf{u}}\}, \tag{63}$$

where

$$[\Phi] = \begin{bmatrix} \lfloor \phi_i \rfloor & 0 & 0 \\ 0 & \lfloor \phi_i \rfloor & 0 \\ 0 & 0 & \lfloor \phi_i \rfloor \end{bmatrix}.$$

The vector $\lfloor \phi_i \rfloor$ contains the four basic functions for the bilinear rectanguular element. The matrix $[\epsilon]$ is thus a 3×12 matrix and $\{\bar{\mathbf{u}}\}$ is a 12×1 vector of nodal values on an elemental region.

Substitution of (63) into equation (61) yields the residuals over the domain Ω,

$$\epsilon_\Omega = \begin{Bmatrix} \epsilon_u \\ \epsilon_v \\ \epsilon_\xi \end{Bmatrix} = \left[[\mathbf{A}]_\alpha \frac{\partial \Phi}{\partial x} + [\mathbf{D}]_\alpha \frac{\partial \Phi}{\partial y} + [\mathbf{C}]_\alpha \Phi \right] \{\bar{\mathbf{u}}\} - \mathbf{f}\},$$

and over the boundary Γ,

$$\epsilon_\Gamma = \begin{Bmatrix} \epsilon_{b\xi} \\ \epsilon_{bu} \end{Bmatrix} = [\mathbf{M}][\Phi]\{\bar{\mathbf{u}}\}^+ - \{\mathbf{g}\}.$$

Collectively we may write the residuals as

$$\epsilon_\Omega = \begin{Bmatrix} \epsilon_u \\ \epsilon_v \\ \epsilon_\xi \\ \epsilon_{b\xi} \\ \epsilon_{bu} \end{Bmatrix} = \begin{bmatrix} [A]_\alpha \dfrac{\partial \Phi}{\partial x} + [D]_\alpha \dfrac{\partial \Phi}{\partial y} + [C]_\alpha \Phi \\ [M][\Phi] \end{bmatrix} \{\bar{\mathbf{u}}\}^+ - \begin{Bmatrix} f \\ g \end{Bmatrix}$$

or equivalently

$$\{\epsilon\} = [\mathbf{LB}^+]\{\bar{\mathbf{u}}\}^+ - [\mathbf{LB}^-]\{\bar{\mathbf{u}}\}^- - \begin{Bmatrix} \hat{f} \\ \hat{g} \end{Bmatrix}, \tag{64}$$

where $\{\bar{\mathbf{u}}\}^+$ and $\{\bar{\mathbf{u}}\}^-$ are the vectors of nodal values of the variables at two different time levels and $[\mathbf{LB}^+]$ and $[\mathbf{LB}^-]$ are the numerical counterparts of the differential operator matrix $[\mathbf{LB}]$.

The total squared weighted residual is given by

$$E^2 = \sum_{\Omega,\Gamma} \{\epsilon\}^{\mathrm{T}} [\mathbf{S}]\{\epsilon\}. \tag{65}$$

Substitution of equation (64) into equation (65) and application of the minimization criterion $\partial E^2 / \partial \{\mathbf{u}\}^+ = \mathbf{0}$ leads to

$$[\mathbf{A}]^+ \{\mathbf{u}\}^+ - \{\mathbf{F}\} = \mathbf{0}, \tag{66}$$

where

$$[\mathbf{A}]^+ = \sum_{\Omega,\Gamma} [\mathbf{LB}^+]^{\mathrm{T}} [\mathbf{S}][\mathbf{LB}^+], \tag{67}$$

$$[\mathbf{A}]^- = \sum_{\Omega,\Gamma} [\mathbf{LB}^+]^{\mathrm{T}} [\mathbf{S}][\mathbf{LB}^-], \tag{68}$$

$$\{\mathbf{P}\} = \sum_{\Omega,\Gamma} [\mathbf{LB}^+][\mathbf{S}] \begin{Bmatrix} \hat{f} \\ \hat{g} \end{Bmatrix}, \qquad \{\mathbf{F}\} = [\mathbf{A}]\{\mathbf{u}\}^- + \{\mathbf{P}\}. \tag{69}$$

6.1. Test problem description

The test problem is shown in Figure 1. The quarter-ring domain has a quadratically varying bathymetry defined by $h = H_0 r^2$, $H_0 = h_1 / r_1^2$, $h_1 = 16$ m. The surface elevation is prescribed at the outer boundary ($r = r^2$) and all other boundaries have a zero-normal-flow condition. The steady state numerical and analytical solutions were obtained for an x-directed constant wind loading. The velocity

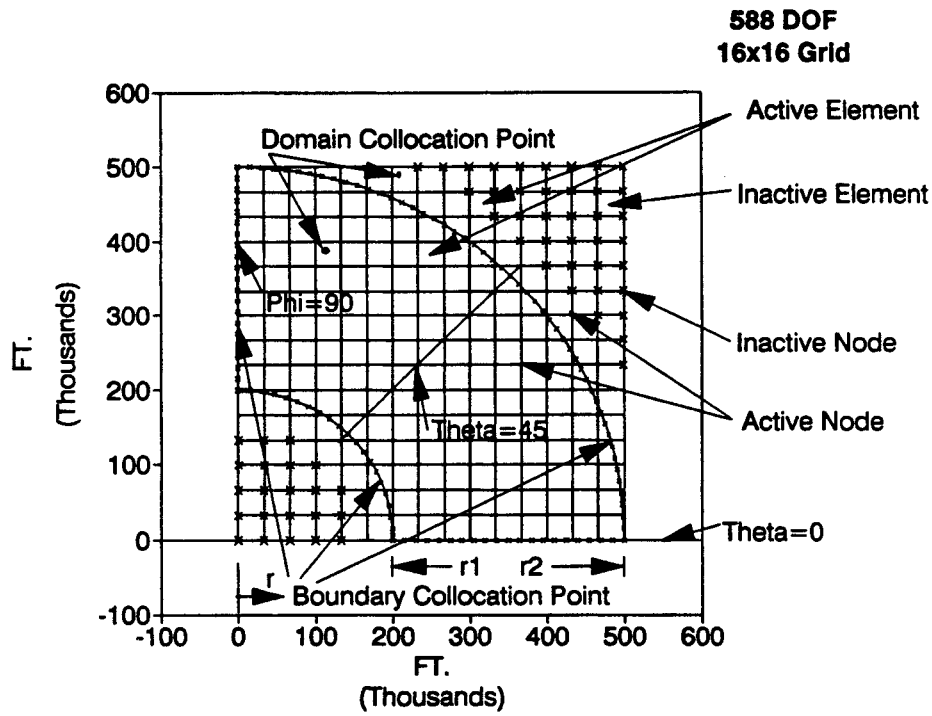

Figure 1. Test grid

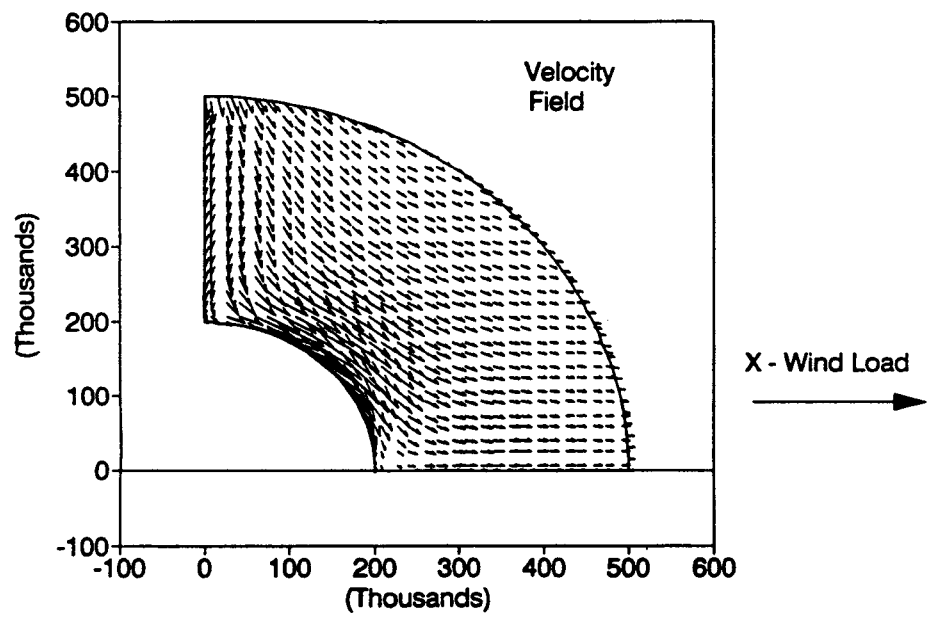

Figure 2. Wind circulation, numerical solution; quadratic bathymetry $h = H_0 r^2$

Table 1. Norms

Iteration	Residual	Galerkin	Exact	Gradient
0	$6{\cdot}515 \times 10^{-13}$	$7{\cdot}225 \times 10^{4}$	$4{\cdot}093 \times 10^{-6}$	$1{\cdot}24 \times 10^{7}$
1	$1{\cdot}554 \times 10^{-12}$	$5{\cdot}548 \times 10^{4}$	$2{\cdot}760 \times 10^{-6}$	$1{\cdot}43 \times 10^{5}$
2	$3{\cdot}586 \times 10^{-12}$	$4{\cdot}233 \times 10^{4}$	$1{\cdot}646 \times 10^{-6}$	$2{\cdot}95 \times 10^{4}$
3	$6{\cdot}666 \times 10^{-12}$	$3{\cdot}315 \times 10^{4}$	$9{\cdot}745 \times 10^{-7}$	$1{\cdot}01 \times 10^{4}$
4	$1{\cdot}254 \times 10^{-11}$	$2{\cdot}624 \times 10^{4}$	$5{\cdot}950 \times 10^{-7}$	$2{\cdot}92 \times 10^{3}$
5	$2{\cdot}578 \times 10^{-11}$	$2{\cdot}027 \times 10^{4}$	$3{\cdot}708 \times 10^{-7}$	$6{\cdot}97 \times 10^{2}$
6	$4{\cdot}691 \times 10^{-11}$	$1{\cdot}582 \times 10^{4}$	$2{\cdot}558 \times 10^{-7}$	$1{\cdot}85 \times 10^{2}$
7	$7{\cdot}186 \times 10^{-11}$	$1{\cdot}291 \times 10^{4}$	$2{\cdot}072 \times 10^{-7}$	$5{\cdot}56 \times 10^{1}$
8	$9{\cdot}896 \times 10^{-11}$	$1{\cdot}102 \times 10^{4}$	$1{\cdot}892 \times 10^{-7}$	$1{\cdot}72 \times 10^{1}$
9	$1{\cdot}279 \times 10^{-10}$	$9{\cdot}778 \times 10^{3}$	$1{\cdot}821 \times 10^{-7}$	$5{\cdot}17 \times 10^{0}$

Table 2. Weights

Iteration	W_1	W_2	W_3	W_4	W_5
0	1·0	1·0	1·0	1·0	1
1	16·4	16·4	16·5	15·1	1
2	41·7	42·3	43·6	19·6	1
3	86·4	88·8	95·4	23·7	1
4	179·0	184·9	208·3	34·1	1
5	382·6	393·8	465·7	54·8	1
6	743·4	753·7	945·1	80·6	1
7	1307·1	1296·7	1744·2	107·5	1
8	2207·4	2141·2	3100·7	140·4	1
9	3722·5	3545·6	5502·8	187·9	1

solution is shown in Figure 2. A complete description of this test problem is given in Reference 16 as well as the analytic solution adapted from Reference 20.

There are five weights in this problem. W_1 and W_2 are the weights for the x- and y-momentum equations respectively. W_3 is associated with the continuity equation. W_4 and W_5 are the weights associated with the prescribed ξ (on $r = r_2$) and the prescribed zero normal flow on the remaining boundaries. All weights were set initially to 1·0. The weight W_5 (normal flow condition) was fixed at 1·0 and the optimizatiorn process was carried out to determine W_1–W_4 that minimize the auxiliary function Ψ_a. Tables I and II and Figure 3 summarize the results. In Table I the numerical values of various norms are listed for each step of the optimization. The residual norm is defined as the L^2-norm of the vector $\{\epsilon\}$ evaluated at each of the active domain collocation points and at the boundary collocation points. The Galerkin norm is actually the value of Ψ_a as defined by equation (35). The values in the exact column are the L^2-norm of the total error. This is simply the sum of the squares of the difference between the numerical and analytical solutions evaluated at the same points used for the residual norm. The gradient norm is the L^2-norm of $\{\partial\Psi_a/\partial s_i\}$ as defined by the left-hand side of equation (42). The four norms are also plotted in Figures 3(a) and 3(b). Owing to their varying magnitudes, the values are normalized with respect to the values in the first row of Table I. The corresponding weights at each iteration are listed in Table II and plotted in Figure 3(c). For this test problem the iteration procedure was carried out until the gradient norm was less than 10^{-6} of the initial gradient norm.

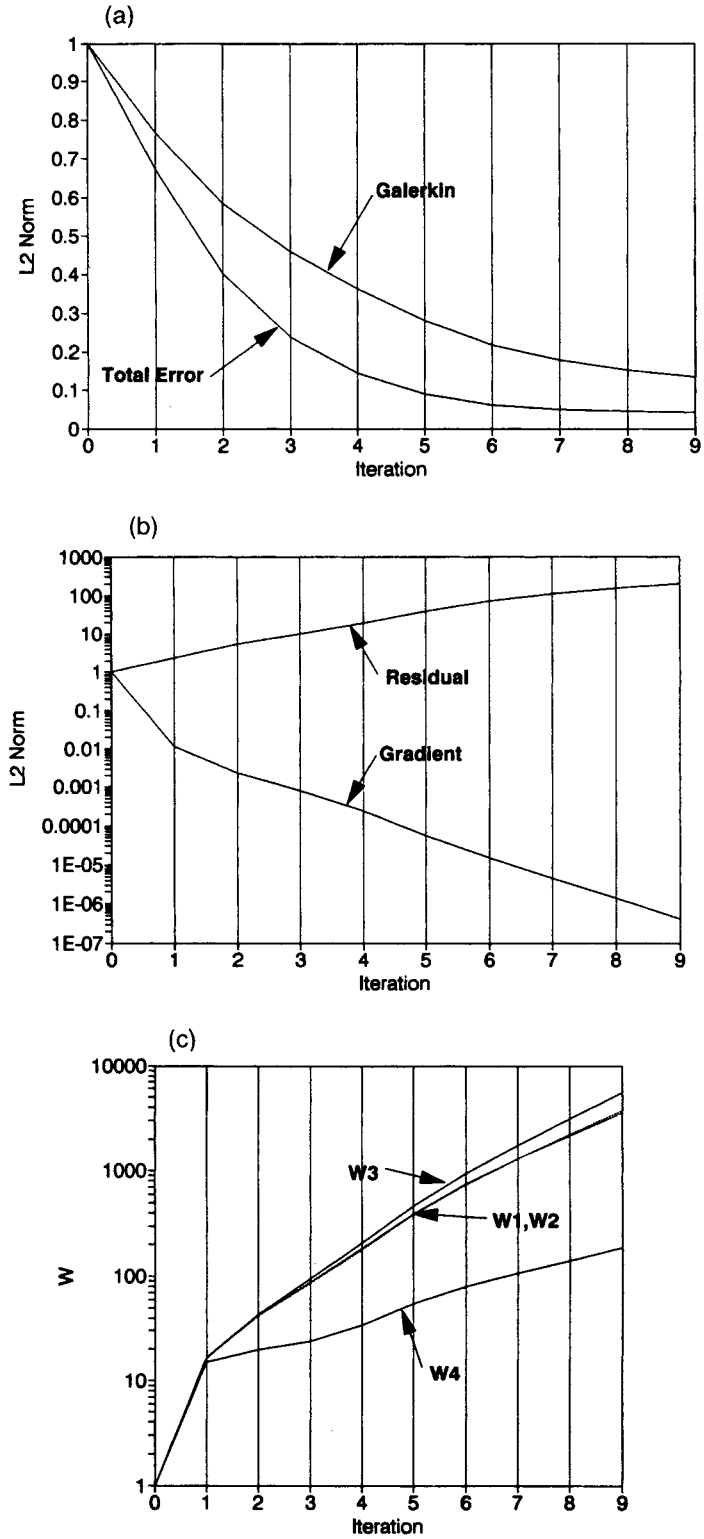

Figure 3. (a) Galerkin norm and total error norm. (b) Least squares residual norm and gradient norm. (c) Weights

6.2. Discussion

The significance of these results is best illustrated by Figure 3(a). We see that weights that reduce the Galerkin norm also reduce the exact (total error) norm. It is thus apparent that in the absence of any knowledge of an analytic solution we can rely on the Galerkin norm as a guide to decreasing the total error of the solution. The importance of this result is further appreciated by considering Figure 3(b), where it can be seen that the L^2 residual norm of the least squares method is actually increasing as the solution becomes more accurate. Figure 3(c) reveals that although the total error can be reduced by minimizing the Galerkin norm, the weights continue to increase ($\{\Delta s_i\}$ becomes constant) while the gradient $\{\partial \Phi_a / \partial s_i\}$ approaches zero. This implies that the second derivatives are zero and that we are approaching a plateau of neutral stability. The selection of weights larger than those that mark the onset of this plateau has a negligible affect on the numerical solution. At the extreme condition of very large weights, however, it is conceivable that ill conditioning may cause the solution to begin to degenerate.

7. CONCLUSIONS

An alternative least squares method for the solution of a first-order system of partial differential equations has been presented. The new feature of the method is a systematic methodology for the determination of the optimal weights appearing in the weighted least squares method.

The results obtained for the solution of the two-dimensional shallow water equations show a superiority of the new approach when compared with the classical least squares and Galerkin methods.

APPENDIX: PROOF OF THEOREM 2

Symbols and notations introduced in Section 4 are used in this appendix.

The proof consists of a comparison between two minimization problems. The first optimization problem does not involve weights and may be formulated as follows. Find the vector solution $\mathbf{a}_p$ of dimension pN^D of the minimization problem

$$\begin{gathered} \underset{\mathbf{a}}{\text{Min}} \int_\Omega [R_L(\mathbf{a}, \mathbf{x})]^2 d\Omega \\ \text{subject to } \mathbf{L}_i(\mathbf{x})\mathbf{a} - f_i = 0, \quad i = 2, \ldots, P, \\ \text{subject to } \mathbf{B}_l(\mathbf{x})\mathbf{a} - g_l = 0, \quad l = 1, \ldots, pb \\ (\text{problem } Q_h), \end{gathered} \tag{70}$$

where $R_L(\mathbf{a},\mathbf{x})$ is defined as the residual of the first equation:

$$R_L(\mathbf{a}, \mathbf{x}) = \mathbf{L}_1(\mathbf{x})\mathbf{a} - f_1. \tag{71}$$

The second minimization problems corresponds to the penalty formulation of problem Q_h and may be formulated as

$$\underset{\mathbf{a}}{\text{Min}} \int_\Omega [R_L(\mathbf{a}, \mathbf{x})]^2 d\Omega + \sum_{i=2}^{P} \frac{1}{\epsilon_i} \int_\Omega (\mathbf{L}_i(\mathbf{x})\mathbf{a} - f_i)^2 d\mu + \sum_{l=1}^{pb} \frac{1}{\nu_l} \int_\Gamma (\mathbf{B}_l(\mathbf{x})\mathbf{a} - g_l)^2 d\mu \tag{72}$$

$$(\text{problem } R_h).$$

For the continuous least squares formulation μ represents the measure associated with the volume integral on Ω.

The proof of Theorem 2 requires the introduction of the Lagrangian multipliers p_{ϵ_i}, p_{ν_j}, p_j and finally pb_l. This is the purpose of Lemmas 1 and 2; then Theorem 2 will result from the relationship between these Lagrangian multipliers.

Lemma 1

Problem Q_h admits a unique solution $\mathbf{a}_p$ characerized by the existence of Lagrangian multipliers p_j, $j = 2, \ldots, p$, and $pb_l = 1, \ldots, pb$, such that

$$\begin{aligned} &\mathbf{A}_1\mathbf{a}_p + \sum_{j=2}^{p}\frac{1}{2}\int_\Omega \mathbf{L}_j^{\mathrm{T}}(\mathbf{x})p_j(\mathbf{x})\mathrm{d}\mu + \sum_{j=1}^{pb}\frac{1}{2}\int_\Omega \mathbf{B}_l^{\mathrm{T}}(\mathbf{x})pb_l(\mathbf{x})\mathrm{d}\mu = \mathbf{f}_1, \\ &\mathbf{L}_i(\mathbf{x})\mathbf{a} - f_i = 0, \quad i = 2,\cdots,p, \qquad \mathbf{B}_l(\mathbf{x})\mathbf{a} - g_l = 0, \quad l = 1,\cdots,pb. \end{aligned} \tag{73}$$

Proof. The interior residual appearing in problem Q_h may be written in terms of the scalar product on $\mathbb{R}^{pD}$ as

$$[R_L(\mathbf{a},\ \mathbf{x})]^2 = (\mathbf{L}_1^{\mathrm{T}}(\mathbf{x})\mathbf{L}_1(\mathbf{x})\mathbf{a},\ \mathbf{a}) - 2(\mathbf{L}_1^{\mathrm{T}}(\mathbf{x})f_1(\mathbf{x}),\ \mathbf{a}) + f_1^2(\mathbf{x}). \tag{74}$$

The last term $f_1^2(\mathbf{x})$ does not depend on $\mathbf{a}$, so problem Q_h is equivalent to the minimization problem

$$\begin{aligned} &\operatorname*{Min}_{\mathbf{a}} \int_\Omega (\mathbf{L}_1^{\mathrm{T}}(\mathbf{x})\mathbf{L}_1(\mathbf{x})\mathbf{a},\ \mathbf{a})\mathrm{d}\mathbf{x} - 2\int_\Omega (\mathbf{L}_1^{\mathrm{T}}\mathbf{f}_1,\ \mathbf{a})\mathrm{d}\mathbf{x} \\ &\text{subject to } \mathbf{L}_i(\mathbf{x})\mathbf{a} - f_i = 0, \quad \mathbf{x}\in\Omega, \quad i = 2,\ldots,p,. \\ &\text{subject to } \mathbf{B}_l(\mathbf{x})\mathbf{a} - g_l = 0, \quad \mathbf{x}\in\Gamma, \quad l = 1,\ldots,pb, \\ &\qquad (\text{problem } Q_h'). \end{aligned} \tag{75}$$

The quadratic minimization problem Q_h' admits a unique solution α_p characterized by the existence of multipliers $p_j(\mathbf{x})$, $j = 2, \ldots, p$, and pb_l, $l = 1, \ldots, pb$, such that equation (73) is satisfied.[1] □

Lemma 2

Problem R_h admits a unique solution $\bar{\mathbf{a}}_\epsilon$ characterized by the existence of the Lagrangian multipliers p_{ϵ_j} such that

$$\begin{aligned} &\mathbf{A}_1\bar{\mathbf{a}}_\epsilon + \sum_{j=2}^{p}\frac{1}{2}\int_\Omega \mathbf{L}_j(\mathbf{x})\mathbf{p}(\mathbf{x})_{\epsilon_j}\mathrm{d}\mu + \sum_{j=1}^{pb}\frac{1}{2}\int_\Gamma \mathbf{B}_l(\mathbf{x})\mathbf{pb}(\mathbf{x})_{\nu_l}\mathrm{d}\mu = \mathbf{f}_1, \\ &\mathbf{p}(\mathbf{x})_{\epsilon_i} = \frac{2}{\epsilon_i}(\mathbf{L}_i(\mathbf{x})\mathbf{a} - f_i), \quad i = 2,\ldots,p, \qquad \mathbf{pb}(\mathbf{x})_{\nu_l} = \frac{2}{\nu_l}(\mathbf{B}_l(\mathbf{x})\mathbf{a} - g_l), \quad l = 11,\ldots,pb. \end{aligned} \tag{76}$$

Proof. Define the objective function of problem R_h as

$$\begin{aligned} f(\mathbf{a},\ \epsilon_1,\ldots,\epsilon_p,\ \nu_1,\ldots,\nu_{pb}) = &\int_\Omega [R_L(\mathbf{a},\ \mathbf{x})]^2\mathrm{d}\Omega + \sum_{i=2}^{p}\frac{1}{\epsilon_i}\int_\Omega (\mathbf{L}_i(\mathbf{x})\mathbf{a} - f_i)^2\mathrm{d}\mu \\ &+ \sum_{l=1}^{pb}\frac{1}{\nu_l}\int_\Gamma (\mathbf{B}_l(\mathbf{x})\mathbf{a} - g_l)^2\mathrm{d}\mu. \end{aligned} \tag{77}$$

The function $\mathbf{a} \rightarrow f(\mathbf{a}, \epsilon_1, \epsilon_2, \ldots, \epsilon_p, \nu_1, \ldots, \nu_{pb})$ defined from $\mathbb{R}^{(p(D+1)+pb)} \rightarrow \mathbb{R}$ is a differentiable function, so the first-order condition for a minimum corresponds to a null differential at the solution $\bar{\mathbf{a}}$.

After differentiation one can easily get

$$\mathbf{A}_1 \bar{\mathbf{a}}_\epsilon + \sum_{j=2}^{p} \frac{1}{\epsilon_j} \int_\Omega \mathbf{L}_i^{\mathrm{T}}(\mathbf{x})(\mathbf{L}_i([\mathbf{x})\bar{\mathbf{a}}_\epsilon - \mathrm{f_i})\mathrm{d}\mu + \sum_{l=1}^{pb} \frac{1}{\nu_l} \int_\Gamma \mathbf{B}_1^{\mathrm{T}}(\mathbf{x})(\mathbf{B}_1(\mathbf{x})\mathbf{a}_\epsilon - \mathbf{g}_1)\mathrm{d}\Gamma = \mathbf{f}_1. \tag{78}$$

By defining p_{ϵ_j} as in (76), equations (76) may be easily found from equation (78). □

By virtue of the implicit function theorem, it may be shown that the vectors $p_\epsilon(\mathbf{x}, c)$ and the $\bar{\mathbf{a}}_\epsilon$ are analytic functions of the variables ϵ_j, $j = 2, \ldots, p$. This last property permits one to prove the following lemma.

Lemma 3

Let $\bar{\mathbf{a}}_\epsilon$ and p_{ϵ_j} be the solution vectors associated with equations (76) and let $\mathbf{a}_p$ and p_j be the solution of equations (73). Then, if p_{ϵ_j} depends on ϵ_j only ($j = 2, \ldots, p$) and pb_{ν_j} depends on ν_j only ($j = 1, \ldots, pb$),

$$\begin{aligned} p_{\epsilon_j}(\mathbf{x}) - p_j(\mathbf{x}) &= H_j(\mathbf{x})\epsilon_j + O(\epsilon_j^2), \quad j = 2, \ldots, p, \\ pb_{\nu_l}(\mathbf{x}) - pb_l(\mathbf{x}) &= D_l(\mathbf{x})\nu_l + O(\nu_l^2), \quad l = 1, \ldots, pb, \end{aligned} \tag{79}$$

where $D_l(\mathbf{x})$ and $H_j(\mathbf{x})$ are functions independent of the penalty parameters ν_l and ϵ_j respectively.

Proof

$$\begin{aligned} p_{\epsilon_j}(\mathbf{x}) &= p_{0j}(\mathbf{x}) + \epsilon_j p_{1j}(\mathbf{x}) + O(\epsilon_j^2), \quad j = 2, \ldots, p, \\ pb_{\nu_l}(\mathbf{x}) &= pb_{0l}(\mathbf{x}) + \nu_l pb_{1l}(\mathbf{x}) + O(\nu_j^2), \quad l = 1, \ldots, pb. \end{aligned} \tag{80}$$

By virtue of the uniqueness of the solution of problem Q_h, one can easily see that $p_{0j}(\mathbf{x})$ and $pb_{\nu_l}(\mathbf{x})$ verify equations (73) and therefore $p_{0j}(\mathbf{x}) = p_j(\mathbf{x})$ and $pb_{0l}(\mathbf{x}) = pb_l(\mathbf{x})$. Define now $p_{1j}(\mathbf{x}) = H_j(\mathbf{x})$ and $pb_{1l}(\mathbf{x}) = D_l(\mathbf{x})$. Lemma 3 is proved. □

The last lemma permits us to express the Lagrangian multiplier vectors $p_{\epsilon_j}(\mathbf{x}), j = 2, \ldots, p$, in terms of the penalty weights ϵ_j and to express $pb_{\nu_j}(\mathbf{x}), j = 1, \ldots, pb$, in terms of the penalty weights ν_j:

$$\begin{aligned} p_{\epsilon_j}(\mathbf{x}) &= p_j(\mathbf{x}) + \epsilon_j H_j(\mathbf{x}) + O(\epsilon_j^2), \quad j = 1, \cdots, p, \\ pb_{\nu_l}(\mathbf{x}) &= pb_l(\mathbf{x}) + \nu_l D_l(x) + O(\nu_j^2), \quad l = 1, \cdots, pb, . \end{aligned} \tag{81}$$

By definition of $p_{\epsilon_j}(\mathbf{x})$ and $pb_{\nu_l}(\mathbf{x})$ (equations (76)) and after integration over Ω one can get

$$\begin{aligned} \int_\Omega (\mathbf{L}_j(\mathbf{x})\bar{\mathbf{a}}_\epsilon - f_j)\mathrm{d}\Omega &= \frac{\epsilon_j}{2} \int_\Omega p_j(\mathbf{x})\mathrm{d}\Omega + O(\epsilon_j^2), \quad j = 2, \ldots, p, \\ \int_\Omega (\mathbf{B}_l(\mathbf{x})\bar{\mathbf{a}}_\epsilon - f_l)\mathrm{d}\Omega &= \frac{\nu_l}{2} \int_\Omega pb_l(\mathbf{x})\mathrm{d}\Omega + O(\nu_l^2), \quad j = 1, \ldots, p. \end{aligned} \tag{82}$$

The last equation finishes the proof of Theorem 2 for the continuous case. For the collocation case the proof is similar, only the measure μ has to be changed in terms of the set of collocation points $\mathbf{x}_i$,

$i = 1, \ldots, k$, for the interior collocation points and $\mathbf{x}_i$, $i = k+1, \ldots, m$, for the boundary collocation points:

$$\int_\Omega f(\mathbf{x})\mathrm{d}\mu = \sum_{i=1}^{k} v_i f(\mathbf{x}_i), \qquad \int_\Gamma f(\mathbf{x})\mathrm{d}\mu = \sum_{i=k+1}^{m} f(\mathbf{x}_i). \tag{83}$$

REFERENCES

1. R. Glowinski, *Numerical Methods for Nonlinear Variational Problems,* Springer, Berlin, 1986.
2. J. Bensabat and D. G. Zeitoun, 'A least-squares formulation for the solution of transport problems', *Int. j. numer. methods fluids,* **10**, 623–636.
3. T. J. R. Hughes, L. P. Franca and G. M. Hulbert, 'A new finite element formulation for computational fluid dynamics: the Galerkin/least squares method for advective–diffusive equations', *Comput. Methods Appl. Mech. Eng.* **73**, 173–189 (1989).
4. E. D. Eason, 'A review of least-squares methods for solving partial differential equations', *Int. j. numer. methods eng.,* **10**, 1021–1046 (1976).
5. K. O. Friedrichs, 'Symmetric positive differential equations', *Commun. Pure Appl. Math.,* **11**, 333–418 (1958).
6. L. Lapidus and G. F. Pinder, *Numerical Solution of Partial Differential Equations in Science and Engineering,* Wiley, New York, 1982.
7. P. Lesaint and P. A. Raviart, 'Finite element collocation methods for first order systems', *Math. Comput.,* **33**, 891–918 (1979).
8. A. K. Aziz and J. L. Liu, 'A weighted least squares method for the backward–forward heat equation', *SIAM J. Numer. Anal.,* **28**, 156–167 (1991).
9. J. H. Bramble and A. H. Shatz, 'Least squares methods for 2*m*th order elliptic boundary value problems', *Math. Comput.,* **25**, 1–32 (1971).
10. W. L. Wendland, *Elliptic Systems in the Plane,* Pitan, London, 1979.
11. A. K. Aziz, R. B. Kellogg and A. B. Stephens, 'Least squares methods for elliptic systems', *Math. Comput,* **44**, 53_70 (1985).
12. G. F. Carey and B. N. Jiang, 'Least-squares finite elements for first order hyperbolic systems', *Int. j. numer. methods eng.,* **26**, 81–93 (1988).
13. B. N. Jiang and G. F. Carey, 'A stable least-squares finite element method for non-linear hyperbolic problems', *Int. j. numer. methods fluids,* **8**, 933–942 (1988).
14. B. N. Jiang and G. F. Carey, 'Least-squares finite element methods for compressible Euler equations', *Int. j. numer. methods fluids,* **10**, 557–568 (1990).
15. J. P. Laible and G. F. Pinder, 'Least squares collocation solution of differential equations on irregularly shaped domains using orthogonal meshes', *Numer. Methods PDEs,* **5**, (1989).
16. J. P. Laible and G. F. Pinder, 'Solution of the shallow water equations by least-squares collocation', *Adv. Water Resources Res.,* submitted.
17. P. A. Raviart and J. M. Thomas, *Introduction a l'Analyse Numerique des Equations aux Derivees Partielles,* Masson, Paris, 1988.
18. T.-F. Chen, 'On least-squares approximations to compressible flow problems', *Numer. Methods PDEs,* **2**, 207–228 (1986).
19. D. G. Zeitoun, J. P. Laible and G. F. Pinder, 'An iterative least- squares solution for boundary value problems', *Int. j. numer. methods eng.,* submitted.
20. D. R. Lynch and W. G. Gray, 'Analytic solutions for computer flow model testing', *J. Hydraul. Div., ASCE,* **104**, 1409–1428 (1978).