INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN FLUIDS, VOL. 23, 29–46 (1996)

AN ENTROPY VARIABLE FORMULATION AND APPLICATIONS FOR THE TWO-DIMENSIONAL SHALLOW WATER EQUATIONS

S. W. BOVA and G. F. CAREY

The Computational Fluid Dynamics Laboratory, The University of Texas at Austin, Mail Code C0600, Austin, TX 78712, U.S.A.

SUMMARY

A new symmetric formulation of the two-dimensional shallow water equations and a streamline upwind Petrov–Galerkin (SUPG) scheme are developed and tested. The symmetric formulation is constructed by means of a transformation of dependent variables derived from the relation for the total energy of the water column. This symmetric form is well suited to the SUPG approach as seen in analogous treatments of gas dynamics problems based on entropy variables. Particulars related to the construction of the upwind test functions and an appropriate discontinuity-capturing operator are included. A formal extension to the viscous, dissipative problem and a stability analysis are also presented. Numerical results for shallow water flow in a channel with (a) a step transition, (b) a curved wall transition and (c) a straight wall transition are compared with experimental and other computational results from the literature.

KEY WORDS: shallow water equations; entropy variables; streamline upwind Petrov–Galerkin; symmetric formulations; finite elements

1. INTRODUCTION

The shallow water equations for free surface flows may be introduced when the free surface wavelength is very long compared with the depth. Finite element models based on various forms of the shallow water equations have been successfully applied in coastal and estuarine hydrodynamics, in part because of the relative ease with which irregular boundaries and bathymetry can be handled.[1–4] For example, finite element schemes have been constructed for the primitive variable or mixed system as well as higher-order, wave equation formulations.[5] There have also been investigations of other approaches such as the harmonic-in-time method[6] and least squares mixed methods.[7]

In general the shallow water equations constitute a hyperbolic or incompletely parabolic system, solutions to which can exhibit discontinuities and steep layers. Petrov–Galerkin finite element schemes have been shown to be effective in handling the numerical difficulties associated with this class of problems, particularly if the system of conservation laws can be written in symmetric form. In the gas dynamics literature, for example, the streamline upwind Petrov–Galerkin (SUPG) method[8,9] and variants of this approach[10,11] have been developed and used to solve the symmetrized Euler and Navier–Stokes equations over a wide range of subcritical and supercritical flow conditions. An impediment to the application of this class of stabilized finite element methods for shallow water problems has been the unavailability of a symmetric form of the conservation equations.

CCC 0271–2091/96/010029–18

Received April 1995
Revised November 1995

We address this issue in the present work by presenting a symmetric, conservation form of the shallow water system. This is accomplished by the introduction of the total energy of the water column to construct a change of dependent variables. We then implement an SUPG finite element scheme. Next an analysis of the scheme including weak stability and the inclusion of horizontal viscosity are briefly considered. Numerical studies for three channel flows complete the investigation.

2. SYMMETRIC SHALLOW WATER SYSTEM

Consider a body of water with mean free surface level in the (x_1, x_2) plane. The bottom depth is given by the positive function $h(x_1, x_2)$ and the unknown surface elevation $\eta(x_1, x_2, t)$ is measured from the mean free surface. Thus the total height is $H = h + \eta$. A body force due to gravity is present and is assumed to act in the negative x_3-direction. The problem is to determine the free surface elevation $\eta(x_1, x_2, t)$ and depth-averaged velocity $\mathbf{u}(x_1, x_2, t) = (u_1, u_2)^{\mathrm{T}}$ for given bathymetry, boundary and initial conditions.

Under certain standard assumptions the incompressible Navier–Stokes equations may be depth averaged to obtain the shallow water equations (e.g. Reference 14). If the additional assumption of negligible Coriolis effects is made, these equations may be written in divergence form as

$$\mathbf{U}_{,t} + \nabla \cdot \mathbf{F}(\mathbf{U}) = \mathbf{S}(\mathbf{U}) \tag{1}$$

for the state vector $\mathbf{U} = (H, Hu_1, Hu_2)^{\mathrm{T}}$. In (1) the subscript comma denotes differentiation with respect to the indicated variable. The divergence term may be written as

$$\nabla \cdot \mathbf{F}(\mathbf{F}_1)_{,x_1} + (\mathbf{F}_2)_{,x_2}, \tag{2}$$

where

$$\mathbf{F}_1(\mathbf{U}) = (Hu_1, Hu_1^2 + gH^2/2, Hu_1u_2)^{\mathrm{T}}, \tag{3}$$

$$\mathbf{F}_2(\mathbf{U}) = (Hu_2, Hu_1u_2, Hu_2^2 + gH^2/2)^{\mathrm{T}} \tag{4}$$

and g is the acceleration due to gravity. Finally, the source term is

$$\mathbf{S}(\mathbf{U}) = (0, gHh_{,x_1} - b_1, gHh_{,x_2} - b_2)^{\mathrm{T}}, \tag{5}$$

where the bottom friction stresses are given by the multidimensional extension of the Manning–Chézy formula[15,16] as

$$b_i = \frac{gn^2u_iV}{\alpha H^{1/3}} \quad \text{for } i = 1, 2. \tag{6}$$

In (6), n is Manning's roughness coefficient, V is the depth-averaged speed, the conversion factor $\alpha = 1{\cdot}0$ for metric units and 2·21 for Imperial units and the assumption of a wide channel has been made so that the hydraulic radius is given by the depth.[15] We remark that Coriolis effects may be included by appropriately modifying $\mathbf{S}$.

Rather than the divergence form (1), the shallow water equations may also be written using the chain rule as the quasi-linear system

$$\mathbf{U}_{,t} + \mathbf{A}_1(\mathbf{U})\mathbf{U}_{,x_1} + \mathbf{A}_2(\mathbf{U})\mathbf{U}_{,x_2} = \mathbf{S}(\mathbf{U}), \tag{7}$$

where the flux Jacobian matrices $\mathbf{A}_1(\mathbf{U})$ and $\mathbf{A}_2(\mathbf{U})$ are given by

$$\mathbf{A}_1(\mathbf{U}) = \begin{pmatrix} 0 & 1 & 0 \\ gH - u_1^2 & 2u_1 & 0 \\ -u_1u_2 & u_2 & u_1 \end{pmatrix}, \qquad \mathbf{A}_2(\mathbf{U}) = \begin{pmatrix} 0 & 0 & 1 \\ -u_1u_2 & u_2 & u_1 \\ gH - u_2^2 & 0 & 2u_2 \end{pmatrix}. \tag{8}$$

It is well known that the system (7) (or equivalently (1)) is hyperbolic; i.e. that the Jacobians $\mathbf{A}_i$ have real eigenvalues and a complete set of eigenvectors (e.g. Reference 14). For most systems of this type that arise from convective transport models, the flux Jacobian matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ are not symmetric, nor can they usually be simultaneously diagonalized.[17] Indeed, this is the case for the shallow water system presently under consideration. Similar types of hyperbolic systems arise in gas dynamics and other applications. These systems may be symmetrized if an appropriate generalized entropy function exists as shown in References 18–20. Finite element methods based on these transformed systems written in terms of the so-called entropy variables have been developed for gas dynamics applications.[8,10,11] These methods are based on the analysis of Harten,[18] who investigated the symmetrization of conservation laws that have associated, generalized entropy functions. In the case of the compressible Navier–Stokes equations with heat conduction, Hughes *et al.*[19] have shown that symmetrization occurs only if the generalized entropy functin is at most trivially different from the physical entropy. With respect to the shallow water equations, Tadmor[20] used the total energy as a generalized entropy function and derived a skew-self-adjoint form of the shallow water equations in terms of the resulting variables. A theorem for hyperbolic conservation systems is presented in the same paper, which establishes an equivalence among the various properties of symmetrizability, having an entropy function and having a skew-self-adjoint form.

In Reference 21 we reviewed the derivation of the variables presented in Reference 20, derived the associated symmetric form of the system (7) and presented numerical results using a one-dimensional SUPG method. In the present study we give the symmetric formulation, generalize the SUPG method to two dimensions and compute steady state solutions for 2D channel flows. Finally, we show in the Appendix that the incompletely parabolic system of shallow water equations that results when viscous stresses are included in the formulation is also symmetrized when written in terms of these new variables.

To proceed, let us define a change of dependent variables, $\mathbf{U} = \mathbf{U}(\mathbf{V})$, under which the new flux Jacobians of the resulting system are symmetric. If the chain rule is applied to (7), the result may be written as

$$\mathbf{A}_0(\mathbf{V})\mathbf{V}_{,t} + \tilde{\mathbf{A}}_1(\mathbf{V})\mathbf{V}_{,x_1} + \tilde{\mathbf{A}}_2(\mathbf{V})\mathbf{V}_{,x_2} = \tilde{\mathbf{S}}(\mathbf{V}), \tag{9}$$

where

$$\mathbf{A}_0 = \frac{\partial \mathbf{U}}{\partial \mathbf{V}} \tag{10}$$

and $\tilde{\mathbf{A}}_i = \mathbf{A}_i\mathbf{A}_0$ for $i = 1, 2$. The change of variables is to be chosen so that $\mathbf{A}_0$ is symmetric and positive definite and the new flux Jacobians $\tilde{\mathbf{A}}_1$ and $\tilde{\mathbf{A}}_2$ are symmetric. Under these conditions, (9) is by definition a symmetric hyperbolic system. The symmetric form of the transport equations is interesting because weighted residual formulations based upon the symmetric form automatically possess certain stability properties.[19,20] We address this issue further in the Appendix.

Let us assume that an appropriate (convex), generalized entropy function $\mathscr{F} = \mathscr{F}(\mathbf{U})$ can be identified. Then the change of variables may be obtained by setting $\mathbf{V}^{\mathrm{T}} = \partial\mathscr{F}/\partial\mathbf{U}$. Convexity of $\mathscr{F}$ is necessary but not sufficient to guarantee symmetrization.[19] There must also exist scalar-valued functions σ_i associated with $\mathscr{F}$ such that

$$\frac{\partial\sigma_i}{\partial\mathbf{U}} = \mathbf{V}^{\mathrm{T}}\mathbf{A}_i. \tag{11}$$

These functions are called entropy fluxes. If $\mathscr{F}$ is convex and corresponding entropy fluxes exist, then the system can be symmetrized.

It is convenient to non-dimensionalize the governing equations before deriving the symmetric form for the shallow water system. This normalization is accomplished by introducing a length scale H_0 and a velocity scale $u_0 = \sqrt{(gH_0)}$. The time is non-dimensionalized by the ratio H_0/u_0. The resulting non-dimensional system is identical with (1), with the definition of $\mathbf{U}$ unchanged. Hence, for notational brevity, we do not introduce new symbols for the scaled system. The non-dimensional flux and source vectors are given by

$$\mathbf{F}_1(\mathbf{U}) = (Hu_1, Hu_1^2 + H^2/2, Hu_1u_2)^{\mathrm{T}}, \tag{12}$$

$$\mathbf{F}_2(\mathbf{U}) = (Hu_2, Hu_1u_2, Hu_2^2 + H^2/2)^{\mathrm{T}} \tag{13}$$

and

$$\mathbf{S}(\mathbf{U}) = (0, Hh_{,x_1} - b_1, Hh_{,x_2} - b_2)^{\mathrm{T}} \tag{14}$$

respectively, with

$$b_i = \frac{n_*^2 u_i \sqrt{(u_1^2 + u_2^2)}}{\alpha H^{1/3}} \quad \text{for } i = 1, 2 \tag{15}$$

and the quantity $n_*^2 = n^2 g/H_0^{1/3}$. The non-dimensional flux Jacobians are given by (8) with g replaced by unity. It should be understood that all subsequent equations are to be regarded as non-dimensional.

The depth-averaged sum of the potential and kinetic energies of the water column may be written as

$$\mathscr{F} = \frac{H^2 + HV^2}{2}. \tag{16}$$

Since this quantity must always decrease across a bore, we choose it to be the non-dimensional, generalized entropy function. This choice leads to the change of variables

$$\mathbf{V}^{\mathrm{T}} = \frac{\partial\mathscr{F}}{\partial\mathbf{U}} = (H - V^2/2, u_1, u_2). \tag{17}$$

It may be noted that the new variables are similar to the primitive variable form of the equations that is usually considered (e.g. Reference 16), in the sense that V_2 and V_3 are simply the Cartesian velocity components. However, the proposed variables differ in that instead of the surface elevation, V_1 may be interpreted as the difference in the potential and kinetic energies.

Differentiation of (17) with respect to $\mathbf{U}$ results in the symmetric matrix

$$\mathbf{A}_0^{-1} = \frac{\partial\mathbf{V}}{\partial\mathbf{U}} = \frac{\partial^2\mathscr{F}}{\partial\mathbf{U}^2} = \begin{pmatrix} 1 + V^2/H & -u_1/H & -u_2/H \\ & 1/H & 0 \\ \text{symm.} & & 1/H \end{pmatrix}. \tag{18}$$

It may be shown that the determinant of $\mathbf{A}_0^{-1}$ is $1/H^2$. Consequently, $\mathbf{A}_0^{-1}$ never becomes singular for physical values of $\mathbf{V}$. The symmetrizer may be found by inverting (18) to obtain

$$\mathbf{A}_0 = \begin{pmatrix} 1 & u_1 & u_2 \\ & H + u_1^2 & u_1 u_2 \\ \text{symm.} & & H + u_2^2 \end{pmatrix}. \tag{19}$$

The corresponding entropy fluxes follow on integrating (11) and may be written as

$$\sigma_i = u_i \left(H^2 + \frac{HV^2}{2} \right) \quad \text{for } i = 1, 2. \tag{20}$$

Finally, the symmetric, non-dimensional flux Jacobian matrices may be obtained after postmultiplying the flux Jacobian matrices $\mathbf{A}_i$, $i = 1, 2$, by $\mathbf{A}_0$. This multiplication leads to the matrices

$$\tilde{\mathbf{A}}_1 = \begin{pmatrix} u_1 & p_1 & u_1 u_2 \\ & u_1(3H + u_1^2) & u_2 p_1 \\ \text{symm.} & & u_1 p_2 \end{pmatrix}, \qquad \tilde{\mathbf{A}}_2 = \begin{pmatrix} u_2 & u_1 u_2 & p_2 \\ & u_2 p_1 & u_1 p_2 \\ \text{symm.} & & u_2(3H + u_2^2) \end{pmatrix}, \tag{21}$$

with $p_i = H + u_i^2$, $i = 1, 2$.

We remark that the flux vectors are not homogeneous functions of $\mathbf{V}$. If this were the case, then by Euler's theorem on homogeneous functions they would satisfy $\gamma \tilde{\mathbf{F}}_i(\mathbf{V}) = \tilde{\mathbf{A}}_i(\mathbf{V})\mathbf{V}$, where γ is the degree of homogeneity. For example, the Euler equations of gas dynamics written in the form (7) have homogeneous flux functions of degree one: they satisy $\mathbf{F}_i(\mathbf{U}) = \mathbf{A}_i(\mathbf{U})\mathbf{U}$. This property may be useful for performing stability analyses or implementing flux split algorithms. The lack of homogeneous flux functions is not a major disadvantage for the proposed symmetric formulation of the shallow water equations since the equations do not have this property even when written in the more familiar conservation form (7).

This completes the transformation to the symmetric form of the shallow water equations (9), with the definitions (17) and (19)–(21). In the next section we develop a streamline upwind Petrov–Galerkin finite element formulation based on this symmetric system.

3. PETROV–GALERKIN FORMULATION

The finite element method used in the present study is based on the approach of Hughes and Mallet,[8,10] who originally applied it to the Euler equations of gas dynamics. The variational formulation is obtained by taking a duality pairing of the transport equations with test functions and integrating over the domain. Thus (9) becomes

$$\int_\Omega \hat{\mathbf{W}}^{\mathrm{T}} (\mathbf{A}_0 \mathbf{V}_{,t} + \tilde{\mathbf{A}}^{\mathrm{T}} \nabla \mathbf{V} - \tilde{\mathbf{S}}) \, d\Omega = 0, \tag{22}$$

where $\tilde{\mathbf{A}}^{\mathrm{T}} = (\tilde{\mathbf{A}}_1, \tilde{\mathbf{A}}_2)$ and

$$\nabla \mathbf{V} = \begin{pmatrix} \mathbf{V}_{,x_1} \\ \mathbf{V}_{,x_2} \end{pmatrix}.$$

For the Petrov–Galerkin formulation the test functions are defined as the standard Galerkin test functions plus a bias term.[9] (For the class of methods considered here, this bias term may be regarded as a directional derivative.) More specifically, we set

$$\hat{\mathbf{W}} = \mathbf{W} + \tilde{\boldsymbol{\tau}}\tilde{\mathbf{A}}^{\mathrm{T}}\nabla\mathbf{W} + \mathscr{D}^{\mathrm{T}}\nabla\mathbf{W}, \tag{23}$$

where $\tilde{\boldsymbol{\tau}}$ denotes a symmetric, positive semidefinite matrix of intrinsic time scales which we discuss in detail later in this section. Briefly, it acts to normalize the directional derivative $\tilde{\mathbf{A}}^{\mathrm{T}}\nabla$. A discontinuity-capturing operator $\mathscr{D}^{\mathrm{T}} = (\mathscr{D}_1^{\mathrm{T}}, \mathscr{D}_2^{\mathrm{T}})$ is useful for eliminating spurious oscillations in the vicinity of local, steep gradients. Note that if $\tilde{\boldsymbol{\tau}} = \mathscr{D} = \mathbf{0}$, the Galerkin method is obtained.

If (23) is substituted into (22) and the Gauss divergence theorem is applied, the result may be written as

$$\int_\Omega \hat{\mathbf{W}}^{\mathrm{T}}\mathbf{A}_0\mathbf{V}_{,t}\,\mathrm{d}\Omega = -\int_{\partial\Omega}\mathbf{W}^{\mathrm{T}}\tilde{\mathbf{F}}_n\,\mathrm{d}\Gamma + \int_\Omega \nabla\mathbf{W}^{\mathrm{T}}\tilde{\mathbf{F}}\,\mathrm{d}\Omega - \int_\Omega \nabla\mathbf{W}^{\mathrm{T}}(\tilde{\mathbf{A}}\tilde{\boldsymbol{\tau}} + \mathscr{D})\tilde{\mathbf{A}}^{\mathrm{T}}\nabla\mathbf{V}\,\mathrm{d}\Omega + \int_\Omega \hat{\mathbf{W}}^{\mathrm{T}}\tilde{\mathbf{S}}\,\mathrm{d}\Omega, \tag{24}$$

where $\tilde{\mathbf{F}}_n = \tilde{\mathbf{F}}_1 n_1 + \tilde{\mathbf{F}}_2 n_2$ and n_1 and n_2 are the components of the local, outward, unit, normal vector. The term on the left side of (24) leads to a mass matrix. The first two terms on the right arise from the application of the divergence theorems to the convective flux vector. The third term on the right is due to the upwinding on the flux term, while the final term accounts for the effects of the source function.

The discretization proceeds by introducing the usual semidiscrete finite element expansion

$$\mathbf{V}_{\mathrm{h}} = \sum_{j=1}^{N} \mathbf{V}_j(t)\psi_j(x_1, x_2), \tag{25}$$

where N is the number of nodes in the finite element mesh and $\psi_j(x_1, x_2)$ are the basis functions. (Linear Lagrantge basis functions are used exclusively in the present work.) Substituting $\mathbf{V}_{\mathrm{h}}$ for $\mathbf{V}$ in (24) and setting the components of W_{h} successively to ψ_i, the semidiscrete system of ODEs has the form

$$\sum_{j=1}^{N} \mathbf{N}_{ij}\mathbf{V}_j'(t) = \mathbf{f}_i(\mathbf{V}(t)), \quad i = 1, \ldots, N, \tag{26}$$

where $\mathbf{V}_j'(t)$ indicates the time rate of change of the entropy variables associated with node j and $\mathbf{N}$ is the non-linear mass matrix whose 3×3 block associated with nodes i and j is given by

$$\mathbf{N}_{ij} = \int_\Omega \{[\mathbf{I}\psi_i + \psi_{i,x_1}(\tilde{\mathbf{A}}_1\tilde{\boldsymbol{\tau}} + \mathscr{D}_1) + \psi_{i,x_2}(\tilde{\mathbf{A}}_2\tilde{\boldsymbol{\tau}} + \mathscr{D}_2)]\mathbf{A}_0\}\psi_j\,\mathrm{d}\Omega. \tag{27}$$

Similarly, the block of the forcing function associated with node i is given by

$$\begin{aligned}\mathbf{f}_i = &-\int_{\partial\Omega}\psi_i\tilde{\mathbf{F}}_n\,\mathrm{d}\Gamma + \int_\Omega(\psi_{i,x_1}\tilde{\mathbf{F}}_1 + \psi_{i,x_2}\tilde{\mathbf{F}}_2)\,\mathrm{d}\Omega \\ &-\int_\Omega[\psi_{i,x_1}(\tilde{\mathbf{A}}_1\tilde{\boldsymbol{\tau}} + \mathscr{D}_1) + \psi_{i,x_2}(\tilde{\mathbf{A}}_2\tilde{\boldsymbol{\tau}} + \mathscr{D}_2)]\tilde{\mathbf{A}}^{\mathrm{T}}\nabla\mathbf{V}\,\mathrm{d}\Omega \\ &+\int_\Omega[\mathbf{I}\psi_i + \psi_{i,x_1}(\tilde{\mathbf{A}}_1\tilde{\boldsymbol{\tau}} + \mathscr{D}_1) + \psi_{i,x_2}(\tilde{\mathbf{A}}_2\tilde{\boldsymbol{\tau}} + \mathscr{D}_2)]\tilde{\mathbf{S}}\,\mathrm{d}\Omega.\end{aligned} \tag{28}$$

Except for the specification of the operators $\tilde{\boldsymbol{\tau}}$ and $\mathscr{D}$, the spatial discretization is complete. The selection of the matrix of intrinsic time scales, $\tilde{\boldsymbol{\tau}}$, is an open problem: linear error estimates, convergence proofs and dimensional analysis provide design conditions to be satisfied, but are insufficient to provide a unique definition.[11] Moreover, the choice of $\tilde{\boldsymbol{\tau}}$ is somewhat dependent on the problem being solved. Hughes and Mallet[8] have presented a formula that works well in practice for the compressible Navier–

Stokes equations. Shakib *et al.*[11] have proposed a more general form. In general, $\tilde{\boldsymbol{\tau}}$ is a symmetric, positive semidefinite matrix that, loosely speaking, acts to normalize the magnitude of the test function bias. For example, consider the one-dimensional, scalar convection equation. In this case $\tilde{\boldsymbol{\tau}}$ reduces to a scalar quantity and it may be shown that an optimal choice is the ratio of the local element size to the magnitude of the convective velocity.[8] For systems of equations the situation is complicated by the presence of multiple wave modes and in general an eigenproblem must be solved to obtain a formula for $\tilde{\boldsymbol{\tau}}$. This situation is discussed in Reference 11, from which we obtain the non-dimensional expression

$$\tilde{\boldsymbol{\tau}} = \frac{l}{2}\mathbf{A}_0^{-1}(\mathbf{A}_1^2 + \mathbf{A}_2^2)^{-1/2}, \tag{29}$$

where $l = \sqrt{(2A_e)}$ is the local estimate of the non-dimensional element length scale (A_e is the area of element e) and the $\mathbf{A}_i$ are the flux Jacobians (8). Note that the inverse square root is taken on a 3×3 matrix. Furthermore, $\tilde{\boldsymbol{\tau}}$ may be interpreted as an inverse norm of the rectangular, convective operator $\mathbf{A}$ with respect to $\mathbf{A}_0$. To compute $\tilde{\boldsymbol{\tau}}$, we write (29) as

$$\tilde{\boldsymbol{\tau}} = \mathbf{A}_0^{-1}\mathscr{B}^{-1/2}, \tag{30}$$

where

$$\mathscr{B} = \frac{4}{l^2}(\mathbf{A}_1^2 + \mathbf{A}_2^2). \tag{31}$$

Then we solve the associated eigenproblem to factor $\mathscr{B}$ using the similarity transformation

$$\mathscr{B} = \mathbf{M}\mathrm{diag}(\beta_k)\mathbf{M}^{-1}, \tag{32}$$

where $\mathbf{M}$ is a modal matrix, the columns of which are the eigenvectors $\mathbf{e}_k$, $k = 1, 2, 3$, and the β_k are the corresponding eigenvalues. It is important to note that the $\mathbf{e}_k$ are scaled so that $\mathbf{M}\mathbf{M}^{\mathrm{T}} = \mathbf{A}_0$.[8] In practice, since $\tilde{\boldsymbol{\tau}}$ is symmetric, it may be computed from the expansion

$$\tilde{\boldsymbol{\tau}} = \sum_{k=1}^{3} \tau_k \boldsymbol{\phi}_k \boldsymbol{\phi}_k^{\mathrm{T}}, \tag{33}$$

where $\tau_k = 1/\sqrt{\beta_k}$ and each eigenvector $\boldsymbol{\phi}_k = \mathbf{A}_0^{-1}\mathbf{e}_k$.

The eigenvalues τ_i, $i = 1, 2, 3$, are given by

$$\tau_{1,2} = \frac{\sqrt{2}l}{\sqrt{(3H + 2V^2 \mp \beta)}}, \qquad \tau_3 = \frac{l}{\sqrt{(V^2 + H)}}, \tag{34}$$

where $\beta = \sqrt{(H^2 + 16HV^2)}$. The eigenvectors are given by

$$\boldsymbol{\phi}_1 = \varphi_1\left(\frac{-4V^2 + H - \beta}{H - \beta}, \frac{4u_1}{H - \beta}, \frac{-u_2(H + \beta)}{4HV^2}\right)^{\mathrm{T}}, \tag{35}$$

$$\boldsymbol{\phi}_2 = \varphi_2\left(\frac{-4V^2 + H + \beta}{H + \beta}, \frac{4u_1}{H + \beta}, \frac{-u_2(H - \beta)}{4HV^2}\right)^{\mathrm{T}}, \tag{36}$$

$$\boldsymbol{\phi}_3 = \varphi_3(0, u_2, = -u_1), \tag{37}$$

with scale factors

$$\varphi_{1,2} = \sqrt{\left(\frac{\beta \mp H}{2\beta}\right)}, \qquad \varphi_3 = \frac{1}{V\sqrt{H}}. \tag{38}$$

The motivation and derivation of the discontinuity-capturing operator (DCO) used in the present study is given in Reference 10 for the Euler equations of gas dynamics. For completeness, we present below the final formulae. The DCO is computed as the product

$$\mathscr{D} = \mathbf{A}_{\|}\tilde{\boldsymbol{\tau}}_2. \tag{39}$$

The matrix $\mathbf{A}_{\|}$ may be interpreted as the projection of $\tilde{\mathbf{A}}$ onto the direction $\nabla\mathbf{V}$ and is defined by

$$\mathbf{A}_{\|} = \frac{\mathrm{diag}_2(\mathbf{A}_0)\nabla\mathbf{V}(\nabla\mathbf{V})^{\mathrm{T}}\tilde{\mathbf{A}}}{|\nabla\mathbf{V}|^2_{\mathbf{A}_0}} \tag{40}$$

where $\mathrm{diag}_2(\mathbf{A}_0)$ denotes a 6×6 operator with two copies of $\mathbf{A}_0$ on the diagonal and zeros elsewhere. In (40) we have also introduced the norm

$$|\nabla\mathbf{V}|_{\mathbf{A}_0} = \sqrt{[(\nabla\mathbf{V})^{\mathrm{T}}\mathrm{diag}_2(\mathbf{A}_0)\nabla\mathbf{V}]}. \tag{41}$$

Note that $\mathbf{A}_{\|}$ is not symmetric. The operator $\tilde{\boldsymbol{\tau}}_2$ in (39) is a symmetric, positive semidefinite, rank-one matrix that contains the time scales associated with the gradient information. It is computed according to

$$\tilde{\boldsymbol{\tau}}_2 = \max(0, \tau_{\|} - \tau)\boldsymbol{\phi}_{\|}\boldsymbol{\phi}_{\|}^{\mathrm{T}}, \tag{42}$$

with the eigenvector

$$\boldsymbol{\phi}_{\|} = \frac{\mathbf{A}_0^{-1}\tilde{\mathbf{A}}^{\mathrm{T}}\nabla\mathbf{V}}{|\tilde{\mathbf{A}}^{\mathrm{T}}\nabla\mathbf{V}|_{\mathbf{A}_0^{-1}}}. \tag{43}$$

Computing the eigenvalue for (42) first requires the evaluation of the scalar

$$\tau_{\|} = \frac{|\nabla\mathbf{V}|^2_{\mathbf{A}_0}}{|\tilde{\mathbf{A}}^{\mathrm{T}}\nabla\mathbf{V}|_{\mathbf{A}_0^{-1}}|\mathbf{D}\mathbf{V}|_{\mathbf{A}_0}}, \tag{44}$$

where the metric derivative

$$\mathbf{D}\mathbf{V} = \begin{pmatrix} \nabla\mathbf{V}^{\mathrm{T}}\nabla\xi_1 \\ \nabla\mathbf{V}^{\mathrm{T}}\nabla\xi_2 \end{pmatrix} \tag{45}$$

has been introduced, with ξ_1 and ξ_2 the computational co-ordinates. Then the definition of the eigenvalue in (42) is completed by constructing

$$\tau = |\mathbf{A}_0\boldsymbol{\phi}_{\|}|^2_{\tilde{\boldsymbol{\tau}}}. \tag{46}$$

This completes the specification of the contribution to the semidiscrete ODE system (26). Now (26) can be integrated time-accurately using standard ODE system integrators to compute $\mathbf{V}(t)$ and the elevation can be obtained by postprocessing $\mathbf{V}$. If a steady state solution exists, then from (26) it satisfies $\bar{\mathbf{f}}(\bar{\mathbf{V}}) = \mathbf{0}$. Here $\bar{\mathbf{f}}$ is a vector of length $3N$ whose block associated with node i is given by (28) and $\bar{\mathbf{V}}$ similarly denotes the assembled global vector of entropy variables. This non-linear system can be scaled or preconditioned in a variety of ways. For example we can premultiply by a preconditioning matrix derived from the mass matrix on the left side of (26); such strategies have been previously applied with some success. In practice the preconditioner $\bar{\mathbf{Q}}^{-1}$ should be simple and easily constructed; hence a diagonal matrix is frequently chosen. Assuming a suitable $\bar{\mathbf{Q}}^{-1}$ can be found, the stationary problem then involves solving $\bar{\mathbf{Q}}^{-1}\bar{\mathbf{f}} = \mathbf{0}$. This solution can be carried out using an appropriate iterative method such as Newton or Picard iteration.

An alternative approach that is gaining popularity is to develop a time-iterative recursion based on the form of (26) in which an iterate is 'time stepped' to the steady state solution. Since this may involve both modification of the matrix on the left side of (26) as well as large time steps, this approach may not be time-accurate. Instead it should be interpreted as a convenient choice of point- or block-iterative recursion with an associated preconditioner.[22] This idea is presented in more detail in the next section and is applied in the numerical studies presented later.

4. TIME-ITERATIVE SOLUTION

Since we are really interested in obtaining an effective iterative scheme for the preconditioned stationary problem, let us first scale (26) to obtain the global system.

$$\bar{\mathbf{Q}}^{-1}\bar{\mathbf{N}}\bar{\mathbf{V}}'(t) = \bar{\mathbf{Q}}^{-1}\bar{\mathbf{f}}. \tag{47}$$

Here $\bar{\mathbf{Q}}^{-1}$ (as well as its inverse $\bar{\mathbf{Q}}$) is a suitable global matrix of size $3N \times 3N$ with 3×3 blocks $\mathbf{Q}_i$ on the diagonal and $\bar{\mathbf{N}}$ is the $3N \times 3N$ matrix whose block elements N_{ij} are given by (27). Now the global mass matrix assembled from the blocks in (27) can be written as

$$\bar{\mathbf{N}} = \bar{\mathbf{N}}_{\mathrm{G}} + \bar{\mathbf{N}}_{\mathrm{U}}, \tag{48}$$

where $\bar{\mathbf{N}}_{\mathrm{G}}$ represents the Galerkin mass matrix for the transformed system whose blocks are given by

$$\mathbf{N}_{\mathrm{G}ij} = \int_{\Omega} \psi_i \psi_j \mathbf{A}_0 \, \mathrm{d}\Omega \tag{49}$$

and $\bar{\mathbf{N}}_{\mathrm{U}}$ denotes the contributions arising from the upwinding. In the present work we approximate $\bar{\mathbf{N}}_{\mathrm{G}}$ by $\bar{\mathbf{N}}_{\mathrm{L}}$ using underintegration via the appropriate Newton–Cotes rule (mass or capacitance lumping) and set $\bar{\mathbf{Q}} = \bar{\mathbf{N}}_{\mathrm{L}}$. Then the diagonal blocks of the preconditioner are given by $\mathbf{Q}_i = L_i \mathbf{A}_{0i}$, where $\mathbf{A}_{0i}$ represents (19) evaluated at mesh point i and L_i is the standard Galerkin lumped mass matrix term for node i. (For linear triangles, L_i is equal to one-third the sum of the area of the triangles whose support includes node i). Then

$$\mathbf{Q}_i^{-1} = \frac{1}{L_i}\mathbf{A}_{0i}^{-1} \tag{50}$$

is the block inverse for constructing the global preconditioner.

Hence (47) may be approximated as

$$(\mathbf{I} + \bar{\mathbf{Q}}^{-1}\bar{\mathbf{N}}_{\mathrm{U}})\bar{\mathbf{V}}'(t) = \bar{\mathbf{Q}}^{-1}\bar{\mathbf{f}}, \tag{51}$$

which can be integrated to obtain a time-accurate or steady state solution. We remark that several algorithms can be derived from (51). For example, the product $\bar{\mathbf{Q}}^{-1}\bar{\mathbf{N}}_{\mathrm{U}}\bar{\mathbf{V}}'(t)$ can be transposed to the right side and 'lagged' to obtain an approximate scheme that can be used for explicit, time-accurate solutions. Similarly, the diagonal blocks associated with these terms can be retained on the left and the off-diagonal blocks transposed to the right to develop a related approach. Of course, fully implicit, transient computations for the system (47) can also be implemented. Finally, for steady-state calculations, these contributions can simply be neglected to yield a block-iterative recursion. In the present study we neglect these contributions and use forward Euler time integration. This generates the following recursion for the block of nodal unknowns at grid point i: for iterate n, compute

$$\Delta \mathbf{V}_i = \frac{\Delta t}{L_i}(\mathbf{A}_{0i}^{-1}\mathbf{f}_i)^{(n)} \tag{52}$$

explicitly at each node of the grid, where $\Delta \mathbf{V}_i = \mathbf{V}_i^{(n+1)} - \mathbf{V}_i^{(n)}$. Since the scheme is not time-accurate, (52) should be interpreted as a basic iterative method with relaxation factor Δt. This suggests that convergence to steady state may be accelerated by varying Δt both between iterations (a varying global time step) and spatially with the grid points (a varying local time step Δt_i). Then (52) becomes

$$\Delta \mathbf{V}_i = \left(\frac{\Delta t_i}{L_i} \mathbf{A}_{0i}^{-1} \mathbf{f}_i \right)^{(n)}. \tag{53}$$

In the computations presented later, $\Delta t_i^{(n)}$ is computed elsewhere as follows. The time step is first initialized to a large value. Then for each element Ω_e we compute

$$(\Delta t)_i^{(n)} = \min_{\forall \Omega_e} \left[(\Delta t)_i, \left(\frac{2l\mathscr{C}}{4(V_n + \sqrt{H}) + cl} \right)^{(n)} \right], \tag{54}$$

where V_n is the magnitude of the velocity component normal to the edge opposite node i, $\mathscr{C}$ is the Courant number and $c = -2n_*^2 V/\alpha H^{4/3}$. (The Courant number is specified as input data for each problem and in practice the largest stable value should be used.)

5. BOUNDARY CONDITIONS

We consider two basic types of boundary conditions: the first type satisfies an inflow/outflow condition in which the normal component of the velocity is non-zero; the second type is a zero-mass-flux boundary which implies that the normal velocity component vanishes and the flow is tangential to the boundary. In each case we evaluate the normal flux $\tilde{\mathbf{F}}_n$ in the boundary integral that appears in (28) using the curretn iterate for V. Then we use the method of characteristic projections[23–25] to determine the number of boundary conditions to apply: for a supercritical inflow, Dirichlet data are applied on all equations at the associated node; for a supercritical outflow, no boundary conditions are applied; for a subcritical inflow, two boundary conditions are required; finally, only one boundary condition is necessary at a subcritical outflow. At a zero-mass-flux boundary there also is only one incoming mode and it can be treated as a special case of subcritical outflow.

The classification of the boundary type in two dimensions depends on the velocity component normal to the boundary, u_n. Accordingly, the projections are defined by the eigenvalues and left eigenvectors of the non-dimensional flux Jacobian

$$\mathbf{A}_n = \mathbf{A}_1 n_1 + \mathbf{A}_2 n_2 = \begin{pmatrix} 0 & n_1 & n_2 \\ Hn_1 - u_1 u_n & u_1 n_1 + u_n & u_1 n_2 \\ Hn_2 - u_2 u_n & u_2 n_1 & u_2 n_2 + u_n \end{pmatrix}. \tag{55}$$

The matrix (55) has eigenvalues

$$\lambda_{1,2} = u_n \mp \sqrt{H}, \qquad \lambda_3 = u_n, \tag{56}$$

with eigenvectors given by the columns of

$$\mathbf{T} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 & 0 \\ u_1 - n_1\sqrt{H} & u_1 + n_1\sqrt{H} & \sqrt{(2H)}n_2 \\ u_2 - n_2\sqrt{H} & u_2 + n_2\sqrt{H} & -\sqrt{(2H)}n_1 \end{pmatrix}. \tag{57}$$

We remark that the eigenvectors in (57) have been scaled so that $\mathbf{T}\mathbf{T}^{\mathrm{T}} = \mathbf{A}_0$. Let the change in the characteristic variables be denoted $\Delta\hat{\mathbf{U}}$. Then this projection may be written as

$$\Delta\hat{\mathbf{U}} = \mathbf{T}^{-1}\Delta\mathbf{U} = \mathbf{T}^{-1}\mathbf{A}_0\Delta\mathbf{V} \tag{58}$$

or

$$\Delta\hat{\mathbf{U}} = \mathbf{T}^{\mathrm{T}}\Delta\mathbf{V}. \tag{59}$$

Equation (59) can be written more explicitly in terms of changes in the depth and velocity components as

$$\Delta\hat{\mathbf{U}} = \begin{pmatrix} (\Delta H - \sqrt{H}\Delta u_n)/\sqrt{2} \\ (\Delta H + \sqrt{H}\Delta u_n)/\sqrt{2} \\ -\sqrt{H}\Delta u_s \end{pmatrix}, \tag{60}$$

where the tangential velocity component $u_s = -u_1 n_2 + u_2 n_1$.

Zero-mass-flux boundaries correspond to boundaries along which the flow is locally tangent. In this case only the first mode is incoming, so only one boundary condition need be applied. That condition is of course $u_n = 0$ equivalently $\Delta u_n = 0$. This condition can be enforced weakly by specifying $u_n = 0$ in the boundary integral that appears in (28). Weak implementations, while convenient, satisfy the boundary constraints only in an average sense along the boundary. In our experience, substantial non-zero normal velocity components can accumulate at these boundary nodes even though they vanish in an average sense. This situation can lead to instabilities, particularly on coarse meshes. For this reason we project the velocities at each iteration so that the zero-mass-flux condition is satisfied strongly at the boundary nodes. We do this by projecting the second and third components of $\Delta\mathbf{V}$ and leaving the first component unconstrained.

6. NUMERICAL RESULTS

In order to demonstrate the proposed method, we simulate the flow through three rectangular channels. For each of the following test cases the initial data consists of a Froude number Fr, a depth profile and a flow angle θ. Then the starting iterate at each node in the mesh is computed according to

$$\mathbf{V} = \begin{pmatrix} H(1 - Fr^2/2) \\ Fr\sqrt{H}\cos\theta \\ Fr\sqrt{H}\sin\theta \end{pmatrix}, \tag{61}$$

except for nodes on the zero-mass-flux boundaries, in which case the tangential component of (61) is taken. For each of the following examples, $\mathscr{C} \approx 0{\cdot}2$.

6.1. *Step Transition*

The first example is that of supercritical flow in a channel whose width is suddenly reduced from 9 to 7 m. The contraction occurs 9·5 m downstream of the inlet boundary. The inlet Froude number (based on the incoming unit depth) is 2·5, the channel is 28·5 m long, has zero bed slope and no bottom friction. This example is obviously not a practical design, but has features that make it an interesting test case. The mesh of linear triangular elements used for this calculation has 1120 nodes and 2073 triangles and is shown in Figure 1. The computed depth of the water column ranges from 0·8694 m to 3·441 m and contours are presented in Figure 2. It may be noted that the hydraulic jump is captured within a band of two elements in the streamwise direction; the irregularity in the contours in this region results from

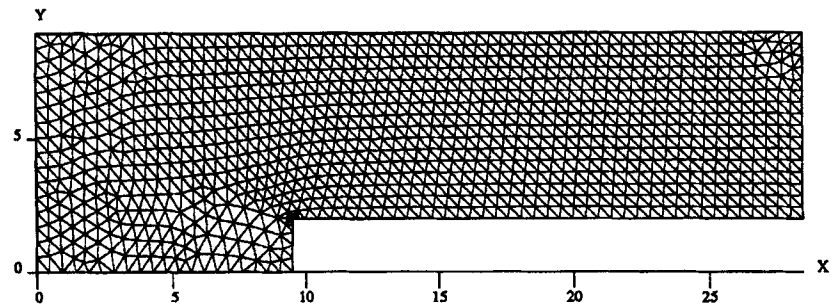

Figure 1. Mesh for step transition: 1120 nodes and 2073 triangles. All units are metres

the shape of this band. Immediately behind the jump, at the point $(x, y) = (1{\cdot}744, 5{\cdot}000)$ m, the depth is 3·099 m. This value compares well with the theoretical value of 3·071 m that should be obtained for a flow of unit depth and initial Froude number of 2·5.[26] The computed Froude numbers are between $2{\cdot}405 \times 10^{-3}$ and 2·500 and are plotted in Figure 3. After passing through the hydraulic jump, the flow becomes subcritical, then becomes supercritical again as it rounds the corner. Near $y = 0$ at the face of the step the flow is nearly stagnant, then accelerates rapidly around the top of the step. Finally, the flow becomes nearly uniform again about 10 m downstream of the step.

6.2. Curved Wall Transition

This problem corresponds to a supercritical transition from a rectangular flume model of width ranging from 2 to 1 ft. The reduction in area begins at $x = 20$ ft and is accomplished using two circular arcs of radius 75 in and a transition length of 41·375 in. The flume is smooth and has a constant bed slope so that a uniform flow is achieved upstream of the transition. We follow Reference 27 and use a bed slope of 0·0125. The Froude number is 4·0 based on the incoming depth of 0·1 ft. Consequently, for a uniform flow to exist upstream of the transition, we must have Manning's friction factor $n = 0{\cdot}005$ in (6).

The mesh used in the present study has 4585 nodes and 8628 linear triangles, a detail of which is shown in Figure 4. The computed Froude numbers (not shown) vary from 4·020 to 1·691. Computed depth contours are shown in Figure 5. Waves of the same family intersect and merge into an oblique hydraulic jump. This occurs at both walls so that the two jumps intersect and reflect from the opposite wall. This interaction continues downstream, becoming progressively weaker. In the numerical solution the depth ranges from 0·3606 to 0·09908 ft, with the peak depth just downstream of the first intersection of the two oblique jumps. These extrema compare well with the experimentally observed values of 0·40

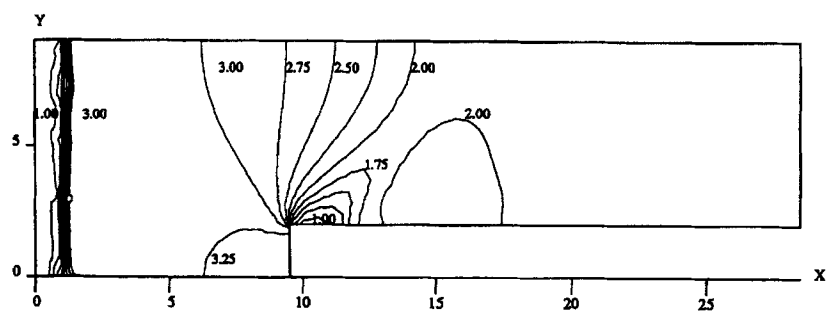

Figure 2. Depth contours for step transition problem. All units are metres

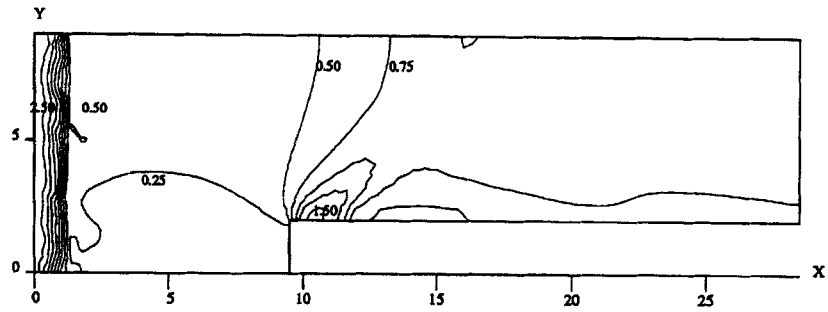

Figure 3. Contours of Froude number of step transition problem. All units are metres

and 0·10 ft respectively. Figure 6 presents these experimental observations along with the numerical results of Berger,[27] who used a related SUPG method to discretize the conservation form of the equations using bilinear quadrilateral elements. These results compare well with those obtained via the proposed method.

Finally, note that there is some mesh-related asymmetry in the contours shown in Figure 5. This may be explained by the fact that although the distribution of nodes is symmetric about the centreline of the flume, the orientation of triangles is not. This local grid orientation effect explains why the waves emanating from the lower wall are resolved more sharply relative to those which emanate from the upper wall. The waves generated by the lower boundary are roughly parallel to the triangle edges throughout most of the domain. In contrast, the waves generated by the upper boundary are roughly orthogonal to the triangle edges. Apparently, hydraulic jumps may be more highly resolved if the triangle edges are aligned with the front. This effect could be ameliorated through the use of an adaptive refinement algorithm which locally reduces the scale of the triangles (and therefore the numerical dissipation of the scheme) in the vicinity of the jumps. With respect to Figure 6, the experimental observations are asymmetric because of the difficulties associated with establishing a completely uniform flow in the laboratory; the numerical results are symmetric because of the use of the bilinear quadrilateral elements.

6.3. Straight Wall Transition

It is well known that the use of curved wall transitions (as in the above example) is generally a poor design practice for supercritical flows. From the standpoint of maximum wave height, a straight wall contraction is always a better choice for a given transition length.[28] This example illustrates this principle. The transition in the above flume model is altered to a straight wall of length 4·758 ft and a

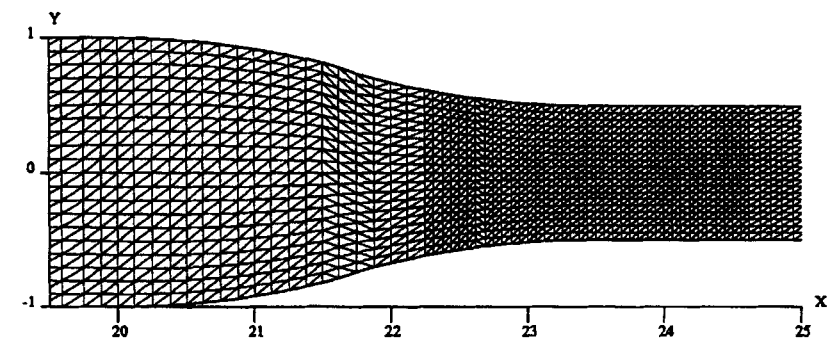

Figure 4. Mesh detail for curved wall transition: total of 4585 nodes and 8628 triangles. All units are feet

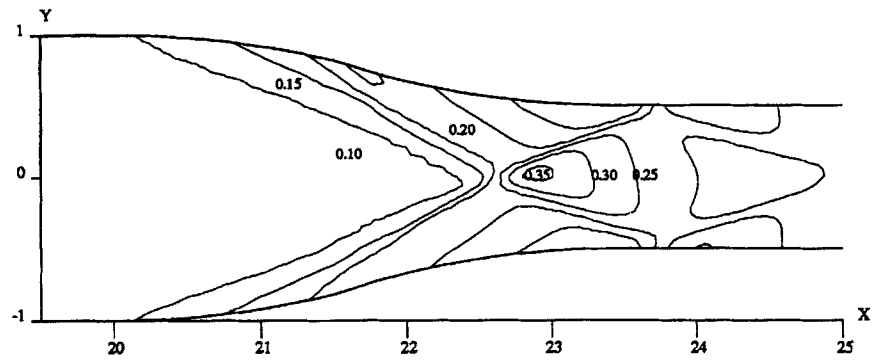

Figure 5. Detail of depth contours for curved wall transition problem. All units are feet

turning angle of 6°. The transition length is increased with respect to that described for the problem in Section 6.2 so that the numerical results of this subsection may be compared with the experimental observations of Ippen and Dawson.[28] The flume width, incoming Froude number and initial depth are unchanged from those given in Section 6.2.

Since a symmetric solution is expected, we consider only one half of the flume. The mesh used has 2100 nodes and 3828 linear triangles. A detail of this mesh is shown in Figure 7. Computed depth contours are shown in Figure 8. Comparing Figure 8 with Figure 5, we again see that an oblique jump forms at the leading edge of the transition and intersects its counterpart that is generated by the lower half of the flume. In the present case, however, negative disturbances[28] are generated at the trailing edge of the transition that interact with the jump downstream and weaken their effect. In the numerical solution the depth ranges from 0·09876 to 0·2394 ft. These extrema can be compared with the experimentally observed values of 0·10 and 0·25 ft respectively.[28] The Froude numbers (not shown) range from 2·390 to 4·023. For a turning angle of 6° and initial Froude number of 4·0 the theoretical

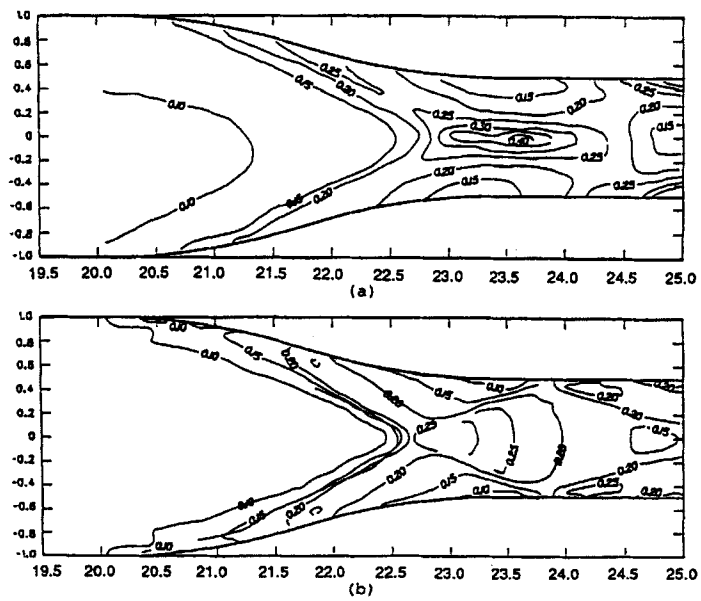

Figure 6. Depth contours for curved wall transition problem: (a) experimental observations of Ippen and Dawson;[28] (b) computational results of Berger.[27] All units are feet. After Reference 27

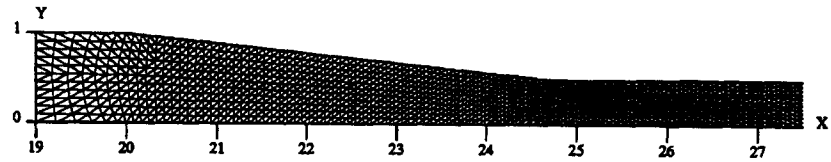

Figure 7. Mesh details for straight wall transition: total of 2397 nodes and 4480 triangles. All units are feet

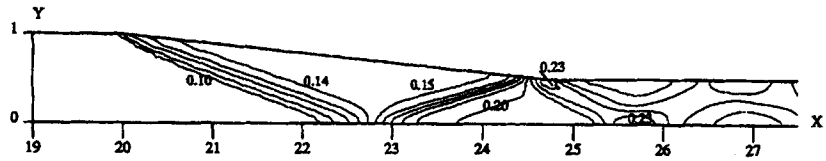

Figure 8. Detail of depth contours for straight wall transition problem. All units are feet

values of the leading wave angle and depth ratio are 20° and 1·498 respectively.[26] These values compare well with Figure 8. The maximum depth for the curved wall transition of Section 6.2 is much greater than that of the straight wall, because in the former case there are no negative disturbances to weaken the jumps.

7. CONCLUDING REMARKS

A symmetric form of the shallow water equations has been developed using variables derived from the total energy of the water column. In the Appendix we show that this choice of variables symmetrizes the system even in the presence of horizontal viscosity. The hyperbolic symmetric equations are used as a starting point for an SUPG finite element method. Forward Euler time integration with local time step adaptation was used to solve the resulting semidiscrete system of ODEs to steady state. Test cases in two dimensions were considered and the proposed algorithms were evaluated for representative channels with supercritical transitions. The results agree well with those published by other authors and demonstrate that the new method can accurately solve the shallow water equations.

The performance of the method for steady state problems is good, at least for the problems considered, and in practice is very stable. Each of the computations presented in Section 6 was performed in about 10–20 min of CPU time on a DEC 3000 Alpha workstation. Forward Euler time stepping, while easily implemented, is relatively expensive and our software could benefit from a more sophisticated time-stepping strategy. For stationary problems, non-standard Runge–Kutta strategies that offer increased stability in exchange for relaxed time accuracy could provide a more efficient steady state solution algorithm (e.g. Reference 22). We plan to explore this issue further in future studies.

The discontinuity-capturing operator considerably increases the robustness of the methodology. When all other factors were unchanged, the computations in Section 6 diverged if the DCO was not included. Solutions could still be obtained, but only for certain combinations of mesh size, starting iterate, time step, etc.

ACKNOWLEDGEMENTS

We are grateful to Dr. R. C. Berger at the U.S. Army Corps of Engineers' Waterway Experimental Station for helpful discussions and for providing the nodal co-ordinate set for the curved wall transition mesh. This work was supported in part by the National Science Foundation.

APPENDIX: INCOMPLETELY PARABOLIC SYSTEM

Symmetric Form

We now address the issue of symmetrizing the incompletely parabolic system

$$\mathbf{U}_{,t} + \mathbf{A}_1(\mathbf{U})\mathbf{U}_{,x_1} + \mathbf{A}_2(\mathbf{U})\mathbf{U}_{,x2} = \mathbf{S}(\mathbf{U}) + \nabla \cdot [\mathbf{K}(\mathbf{U}(\nabla \mathbf{U}], \tag{62}$$

where we have introduced the diffusivity tensor $\mathbf{K}(\mathbf{U})$. For the shallow water system under consideration, $\mathbf{K}(\mathbf{U})$ results from a combination of the molecular and turbulent Reynolds stresses. Typically, these effects are modelled empirically (see e.g. Reference 16) so that $\mathbf{K}(\mathbf{U})$ may be written as

$$\mathbf{K}(\mathbf{U}) = \begin{pmatrix} \mathbf{K}_{11}(\mathbf{U}) & \mathbf{K}_{12}(\mathbf{U}) \\ \mathbf{K}_{21}(\mathbf{U}) & \mathbf{K}_{22}(\mathbf{U}) \end{pmatrix}, \tag{63}$$

where

$$\mathbf{K}_{11}(\mathbf{U}) = \begin{pmatrix} 0 & 0 & 0 \\ -u_1\varepsilon_{11} & \varepsilon_{11} & 0 \\ -u_2\varepsilon_{21} & 0 & \varepsilon_{21} \end{pmatrix}, \qquad \mathbf{K}_{22}(\mathbf{U}) = \begin{pmatrix} 0 & 0 & 0 \\ -u_1\varepsilon_{12} & \varepsilon_{12} & 0 \\ -u_2\varepsilon_{22} & 0 & \varepsilon_{22} \end{pmatrix}, \tag{64}$$

$\mathbf{K}_{12}(\mathbf{U}) = \mathbf{K}_{21}(\mathbf{U}) = \mathbf{0}$ and the dispersion coefficients are represented by ε_{ij}, $i = 1, 2, j = 1, 2$. Note that the matrices in (64) are not symmetric. We have already shown in Section 2 that the generalized entropy function (16) symmetrizes the flux Jacobians $\mathbf{A}_i$, $i = 1, 2$. We show below that it also simultaneously symmetrizes $\mathbf{K}_{11}$ and $\mathbf{K}_{22}$.

The viscous term may be written as

$$\nabla \cdot (\mathbf{K}\nabla \mathbf{U}) = (\mathbf{K}_{11}\mathbf{U}_{,x_1})_{,x_1} + (\mathbf{K}_{22}\mathbf{U}_{,x_2})_{,x_2}. \tag{65}$$

Application of the chain rule yields

$$\nabla \cdot (\mathbf{K}\nabla \mathbf{U}) = (\mathbf{K}_{11}\mathbf{A}_0\mathbf{V}_{,x_1})_{,x_1} + (\mathbf{K}_{22}\mathbf{A}_0\mathbf{V}_{,x_2})_{,x_2}. \tag{66}$$

Now let

$$\tilde{\mathbf{K}}_{ij} = \mathbf{K}_{ij}\mathbf{A}_0. \tag{67}$$

Direct matrix multiplication reveals that

$$\tilde{\mathbf{K}}_{11} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \varepsilon_{11}H & 0 \\ 0 & 0 & \varepsilon_{21}H \end{pmatrix}, \qquad \tilde{\mathbf{K}}_{22} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \varepsilon_{12}H & 0 \\ 0 & 0 & \varepsilon_{22}H \end{pmatrix}. \tag{68}$$

Observe that the matrices in (68) are diagonal as well as symmetric and positive semidefinite. Hence the system (62) may be written as the symmetric, incompletely parabolic system

$$\mathbf{A}_0\mathbf{V}_{,t} + \tilde{\mathbf{A}}^{\mathrm{T}}\nabla \mathbf{V} = \tilde{\mathbf{S}} + \nabla \cdot (\tilde{\mathbf{K}}\nabla \mathbf{V}), \tag{69}$$

where

$$\tilde{\mathbf{K}} = \begin{pmatrix} \tilde{\mathbf{K}}_{11} & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{K}}_{22} \end{pmatrix}. \tag{70}$$

Stability Bound

Finally, we show that a weak stability result analogous to that presented for the compressible Navier–Stokes equations in Reference 19 can be demonstrated for the system (69). This derivation is more conveniently performed if indicial notation is first introduced. For example, let $\mathbf{V}_{,i} = \mathbf{V}_{,x_i}$ and also let a repeated subscript indicate summation. Then (69) may be written as

$$\mathbf{A}_0\mathbf{V}_{,t} + \tilde{\mathbf{A}}_i\mathbf{V}_{,i} = \tilde{\mathbf{S}} + (\tilde{\mathbf{K}}_{ij}\mathbf{V}_{,j})_{,i}.$$

The derivation proceeds from the weak statement

$$\int_\Omega \mathbf{W}^{\mathrm{T}}[\mathbf{A}_0\mathbf{V}_{,t} + \tilde{\mathbf{A}}_i\mathbf{V}_{,i} - \tilde{\mathbf{S}} - (\tilde{\mathbf{K}}_{ij}\mathbf{V}_{,j})_{,i}]\,\mathrm{d}\Omega = 0 \tag{71}$$

for all admissible test functions $\mathbf{W}$. In particular, setting $\mathbf{W} = \mathbf{V}$ in (71) and integrating the second-order term by parts, we obtain

$$\int_\Omega [\mathbf{V}^{\mathrm{T}}\mathbf{A}_0\mathbf{V}_{,t} + \mathbf{V}^{\mathrm{T}}\tilde{\mathbf{A}}_i\mathbf{V}_{,i} - \mathbf{V}^{\mathrm{T}}\tilde{\mathbf{S}} - (\mathbf{V}^{\mathrm{T}}\tilde{\mathbf{K}}_{ij}\mathbf{V}_{,j})_{,i}]\,\mathrm{d}\Omega = -\int_\Omega \mathbf{V}_{,i}^{\mathrm{T}}\tilde{\mathbf{K}}_{ij}\mathbf{V}_{,j}\,\mathrm{d}\Omega \leqslant 0, \tag{72}$$

since the $\tilde{\mathbf{K}}_{ij}$ are symmetric and positive semidefinite. Now, by construction,

$$\mathbf{V}^{\mathrm{T}}\mathbf{A}_0\mathbf{V}_{,t} = \left(\frac{\partial \mathscr{F}}{\partial \mathbf{U}}\right)^{\mathrm{T}}\frac{\partial \mathbf{U}}{\partial \mathbf{V}}\mathbf{V}_{,t} = \mathscr{F}_{,t} \tag{73}$$

and, using (11),

$$\mathbf{V}^{\mathrm{T}}\tilde{\mathbf{A}}_i\mathbf{V}_{,i} = \frac{\partial \sigma_i}{\partial \mathbf{U}}\frac{\partial \mathbf{U}}{\partial \mathbf{V}}\mathbf{V}_{,i} = \sigma_{i,i}. \tag{74}$$

By substitution of (14) and (17) the product $\mathbf{V}^T\tilde{\mathbf{S}}$ may be written as

$$\mathbf{V}^{\mathrm{T}}\tilde{\mathbf{S}} = Hu_ih_{,i} - u_ib_i. \tag{75}$$

Combining (73)–(75) in (72) we get

$$\int_\Omega [\mathscr{F}_{,t} + \sigma_{i,i} - Hu_ih_{,i} - u_ib_i - (\mathbf{V}^{\mathrm{T}}\tilde{\mathbf{K}}_{ij}\mathbf{V}_{,j})_{,i}]\,\mathrm{d}\Omega = -\int_\Omega \mathbf{V}_{,i}^{\mathrm{T}}\tilde{\mathbf{K}}_{ij}\mathbf{V}_{,j}\,\mathrm{d}\Omega \leqslant 0. \tag{76}$$

Hence the weak solution $\mathbf{V}$ to the incompletely parabolic problem satisfies this growth inequality.

REFERENCES

1. R. F. Dressler, 'New nonlinear shallow flow equations with curvature', *J. Hydraul. Res.*, **16**, 205–272 (1978).
2. W. G. Gray and D. R. Lynch, 'Time-stepping schemes for finite element tidal model computations', *Adv. Water Resources*, **2**, 83–95 (1977).
3. M. Kawahara, H. Hirano, K. Tsubota and K. Inagaki, 'Selective lumping finite element method for shallow water flow', *Int. j. numer. methods fluids*. **2**, 89–112 (1982).
4. R. Walters, 'A model for tides and currents in the English Channel and Southern North Sea', *Adv. Water Resources*, **10**, 138–148 (1987).
5. R. L. Kolar, J. J. Westerink, M. E. Catekin and C. A. Blain, 'Aspects of nonlinear simulations using shallow water models based on the wave continuity equation', *Comput. Fluids*, **23**, 523–538 (1994).
6. R. Walters, 'Numerically induced oscillations in finite element approximations to the shallow water equations', *Int. j. numer. methods fluids*, **3**, 591–604 (1983).
7. 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).
8. T. J. R. Hughes and M. Mallet, 'A new finite element formulation for computational fluid dynamics: III. The generalized streamline operator for multidimensional advective–diffusive systems', *Comput. Methods Appl. Mech. Eng.*, **58**, 305–328 (1986).

9. C. Johnson, 'Streamline diffusion methods for problems in fluid mechanics', in *Finite Elements in Fluids*, Wiley, Chichester, Vol. 6, 1985, pp. 251–261.
10. T. J. R. Hughes and M. Mallet, 'A new finite element formulation for computational fluid dynamics: IV. A discontinuity-capturing operator for multidimensional advective–diffusive systems', *Comput. Methods Appl. Mech. Eng.*, **58**, 329–336 (1986).
11. F. Shakib, T. J. R. Hughes and Z. Johan, 'A new finite element formulation for computational fluid dynamics: X. The compressible Euler and Navier–Stokes equations', *Comput. Methods Appl. Mech. Eng.*, **89**, 141–219 (1991).
12. R. C. Berger and G. F. Carey, 'A perturbation analysis and finite element approximate model for free surface flow over curved beds', *Int. j. numer. methods eng.*, **31**, 493–507 (1991).
13. R. Walters and G. F. Carey, 'Analysis of spurious oscillation modes for the shallow water and Navier–Stokes equations', *Comput. Fluids*, **11**, 51–68 (1983).
14. J. J. Stoker, *Water Waves*, Interscience, New York, 1957.
15. V. T. Chow, *Open Channel Hydraulics*, McGraw-Hill, New York, 1959.
16. R. Walters and R. Cheng, 'A two-dimensional hydrodynamic model of a tidal estuary', *Adv. Water Resources*, **2**, 177–184 (1979).
17. R. F. Warming, R. M. Beam and B. J. Hyett, 'Diagonalization and simultaneous symmetrization of the gas-dynamic matrices', *Math. Comput.*, **29**, 1037–1045 (1975).
18. A. Harten, 'On the symmetric form of systems of conservation laws with entropy', *J. Comput. Phys.*, **49**, 151–164 (1983).
19. T. J. R. Hughes, L. P. Franca and M. Mallet, 'A new finite element formulation for computational fluid dynamics: I. Symmetric forms of the compressible Euler and Navier–Stokes equations and the second law of thermodynamics', *Comput. Methods Appl. Mech. Eng.*, **54**, 223–234 (1986).
20. E. Tadmore, 'Skew-selfadjoint form for systems of conservation laws', *J. Math. Anal. Appl.*, **103**, 428–442 (1984).
21. S. W. Bova and G. F. Carey, 'A new symmetrized formulation and SUPG scheme for the shallow water equations', *Adv. Water Resources*, (in press).
22. A. A. Lorber, G. F. Carey and W. D. Joubert, 'ODE recursions and iterative solvers for linear equations', *SIAM J. Sci. Comput.*, **17**, 1 (1996).
23. B. Engquist and A. Majda, 'Absorbing boundary conditions for the numerical simulation of waves', *Math. Comput.*, **31**, 629–651 (1977).
24. R. L. Higdon, 'Initial boundary value problems for linear hyperbolic systems', *SIAM Rev.*, **28**, 177–217 (1986).
25. H. O. Kreiss, 'Initial boundary value problems for hyperbolic systems', *Commun. Pure Appl. Math.*, **23**, 277–298 (1970).
26. A. T. Ippen, 'Mechanics of supercritical flow', *Trans. ASCE*, **116**, 268–295 (1951).
27. R. C. Berger, 'Free-surface flow over curved surfaces', *Ph.D. Thesis*, University of Texas at Austin, 1992.
28. A. T. Ippen and J. H. Dawson, 'Design of channel contractions', *Trans. ASCE*, **116**, 326–346 (1951).