Appeared in: *Computational Methods in Water Resources XIII, Vol. 2,* Bentley et al., eds. 907-912, Balkema, 2000.

A mass conservation analysis of the GWCE formulation

A. Aldama and A. Aguilar
Mexican Institute of Water Technology & National Autonomous University of Mexico

R. Kolar
University of Oklahoma

J. J. Westerink
University of Notre Dame

ABSTRACT: The Generalized Wave Continuity Equation (GWCE) formulation of the shallow water equations was developed in order to eliminate spurious node-to-node oscillations that were observable in finite element formulations of the primitive equations. In the GWCE formulation, the continuity equation is substituted by a linear combination of the mass conservation the momentum equations, which is cast in wave-like form. The resulting equation is coupled with either the conservative or the nonconservative version of the momentum equation, thus closing the system. A number of investigators have noticed that the GWCE formulation exhibits mass conservation errors that are exacerbated when fine grids are employed. A Fourier analysis is performed on the GWCE formulation. It is shown that the said formulation is not consistent with the mass conservation principle, a fact that explains the observational evidence. Numerical tests that confirm the theoretical findings are also included.

1 INTRODUCTION

Finite element discretizations of the shallow water equations in primitive form, produce solutions that exhibit spurious (spatial) node to node oscillations (Kinmark & Gray 1985). The continuity statement may be substituted by a linear combination of the continuity and the momentum equations, which is cast in wave-like form. The resulting equation combined with the momentum statement, gives rise to the generalized wave continuity equation (GWCE) formulation (Lynch & Gray 1979, Kinmark & Gray 1985).

The GWCE formulation is effective in suppressing the nodal oscillations, but several investigators have observed that it appears to exhibit some deficiencies in conserving mass (Kolar et al. 1994). This paper addresses the mass conservation issue by employing Fourier analyses of both the continuous and discrete versions of the GWCE formulation. In order to simplify the analyses and interpretation of the results, a one-dimensional version of the said formulation is used. Numerical examples that illustrate the theoretical findings are also included.

2 ONE-DIMENSIONAL CONTINUOUS GWCE FORMULATION

2.1 Equations of motion

As is well known, the Saint-Venant equations constitute the one-dimensional analogue of the shallow water equations. The former equations may be written as follows:

$$\mathcal{L}(A,Q) = \partial_t A + \partial_x Q = 0 \tag{1}$$

$$\mathcal{M}(A,Q) \equiv \partial_t Q + \partial_x \left(\frac{Q^2}{A} \right) + gA\left(\partial_x H + S_f\right) = 0 \tag{2}$$

where $\mathcal{L}(\cdot,\cdot)$ is the mass conservation or continuity operator; $\mathcal{M}(\cdot,\cdot)$, the momentum operator; A, the hydraulic area; Q, the discharge; H, the water surface elevation; g, the acceleration of gravity; S_f, the friction slope; x, the spatial coordinate; and t, time.

The GWCE is constructed as follows:

$$\mathcal{W}(A,Q) = \partial_t \mathcal{L}(A,Q) - \partial_x \mathcal{M}(A,Q) + G\mathcal{L}(A,Q) = 0 \tag{3}$$

where $\mathcal{W}(\cdot,\cdot)$ is the GWCE operator; and G, a constant parameter. As may be observed, as $G \to \infty$, $\mathcal{W} \to \mathcal{L}$, i.e., the GWCE (asymptotically) approaches the (primitive) mass conservation equation.

Substituting (1) y (2) en (3) results in:

$$\mathcal{W}(A,Q) = \partial_{tt}A - \partial_{xx}\left(\frac{Q^2}{A}\right) - g\partial_x\left[A\left(\partial_x H + S_f\right)\right] + G\left(\partial_t A + \partial_x Q\right) = 0 \quad (4)$$

Eq. (4), coupled with Eq.(2) constitutes the GWCE formulation for the one-dimensional free surface flow problem.

2.2 Perturbation analysis

Let the dependent variables A and Q in Eqs. (2) and (4) be decomposed as follows:

$$A = \overline{A} + a;\ \|a\| << \|\overline{A}\| \text{ and } Q = \overline{Q} + q;\ \|q\| << \|\overline{Q}\| \quad (5)$$

where $\overline{A}$ and $\overline{Q}$ are reference solutions of the system (2) and (4), and a and q are small perturbations to these solutions. It is further assumed that $\overline{A}$ and $\overline{Q}$ vary slowly in the scale of variation characteristic of a and q. Substituting (5) in (4) and (2) yields:

$$\partial_{tt}\left(\overline{A}+a\right) - \partial_{xx}\left[\frac{\left(\overline{Q}+q\right)^2}{\overline{A}+a}\right] - g\partial_x\left\{\left(\overline{A}+a\right)\times\left[\partial_x H\left(\overline{A}+a\right) + S_f\left(\overline{A}+a,\overline{Q}+q\right)\right]\right\} + G\left[\partial_t\left(\overline{A}+a\right) + \partial_x\left(\overline{Q}+q\right)\right] = 0 \quad (6)$$

$$\partial_t\left(\overline{Q}+q\right) + \partial_x\left[\frac{\left(\overline{Q}+q\right)^2}{\overline{A}+a}\right] + g\left(\overline{A}+a\right)\times\left[\partial_x H\left(\overline{A}+a\right) + S_f\left(\overline{A}+a,\overline{Q}+q\right)\right] = 0 \quad (7)$$

2.2 Localization

We will now employ the approach developed by Aldama and his collaborators (Aldama & Paniconi 1991, Aldama & Aguilar 1996, Aldama & Aparicio 1998). That approach consists of employing a Taylor-Fréchet expansion of Eqs. (6) and (7), using the fact that $\mathcal{W}(\overline{A},\overline{Q}) = 0$ and $\mathcal{M}(\overline{A},\overline{Q}) = 0$, recalling that $\overline{A}$ and $\overline{Q}$ vary slowly in the scale of variation characteristic of the perturbations, and neglecting terms of quadratic and higher orders. Thus the coefficients in the equations that govern the behavior of a and q may be localized at a reference point (x_0, t_0), yielding the following system of linear equations with constant coeficients:

$$\partial_{tt}a - 2U_o\partial_{xx}q + U_o\left(1 - F_{r0}^{-2}\right)\partial_{xx}a - F_1\partial_x q - \left(gS_{f_o} - F_2\right)\partial_x a + G\left[\partial_t a + \partial_x q\right] = 0 \quad (8)$$

$$\partial_t q + 2U_o\partial_x q - U_o\left(1 - F_{r0}^{-2}\right)\partial_x a + F_1 q + \left(gS_{f_o} - F_2\right)a = 0 \quad (9)$$

where the subindex "0" denotes the reference point (x_0, t_0), $U_0 = Q_0 / A_0$ is the flow velocity associated with such point; and $F_{r0} = U_0 / \sqrt{gD_0}$, is the Froude number, where D_0 represents the hydraulic depth. In addition, when Manning´s friction formula is employed:

$$F_1 = 2\alpha\left(\frac{k_s}{R_0}\right)^{1/3}\frac{Q_0}{R_0 A_0} \quad (10)$$

$$F_2 = \alpha\left(\frac{k_s}{R_0}\right)^{1/3}\frac{Q_0^2}{R_0 A_0}\left[\frac{1}{A_0} + \frac{4}{3R_0}\left(\frac{dR}{dA}\right)_0\right] \quad (11)$$

$$S_{f0} = \alpha\left(\frac{k_s}{R_0}\right)^{1/3}\frac{Q_0^2}{R_0 A_0^2} \quad (12)$$

where R_0 represents the hydraulic radius associated with the reference point; k_s, the Nikuradse equivalent roughness height; and $\alpha \cong 17/100$ (Aldama and Ocón 1998), and it is assumed that $Q_0 > 0$.

2.3 Fourier analysis

We will now study the propagation properties of the system (8)-(9) in Fourier space. Since the said system is linear and with constant coefficients, it is possible, without loss of generality, to assume that the perturbations are given by single Fourier modes of the following form:

$$a = \hat{a}(k)e^{i(kx-\omega t)} \quad (13)$$

$$q = \hat{q}(k)e^{i(kx-\omega t)} \quad (14)$$

where k represents wavenumber and ω, frequency. Substituting (13) and (14) in (8) and (9) yields:

$$\begin{bmatrix} \alpha_{1,1} & \alpha_{1,2} \\ \alpha_{2,1} & \alpha_{2,2} \end{bmatrix} \begin{bmatrix} \hat{a} \\ \hat{q} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad (15)$$

where

$$\alpha_{1,1} = \omega^2 + k^2 U_o^2 \left(1 - F_{r0}^{-2}\right) + ik\left(gS_{f_o} - F_2\right) + i\omega G \quad (16)$$

$$\alpha_{1,2} = -2k^2 U_o + ikF_1 - iGk \quad (17)$$

$$\alpha_{2,1} = kU_o^2 \left(1 - F_{r0}^{-2}\right) + i\left(gS_{f_o} - F_2\right) \quad (18)$$

$$\alpha_{2,2} = \omega - 2kU_o + iF_1 \quad (19)$$

Eq. (15) represents a linear and homogeneous system in $\hat{a}$ and $\hat{q}$ that possesses nontrivial solutions if and only if the determinant of its coefficient matrix vanishes. This consideration generates a cubic equation in ω, whose solutions represent the dispersion relation for the GWCE formulation. Those solutions are:

$$\omega = \begin{cases} -\frac{1}{2}\left[iF_1 - 2kU_o\right] \pm \left[\left(iF_1 - 2kU_o\right)^2 - \right. \\ \left. 4k\left(kU_o^2\left(1 - F_{r0}^{-2}\right) + i\left(gS_{f_o} - F_2\right)\right)\right]^{1/2} \\ \\ -iG \end{cases} \quad (20)$$

As may be observed in Eq. (20), ω is complex, of the form $\omega = \omega_r + i\,\omega_i$. Therefore, the solutions for a and q will be stable only when $\omega_i \leq 0$. For the first two values of ω, given by (20), this inequality results in the following well known condition for the stability of one dimensional free surface flow:

$$\left| V_e \right| \leq 1 \quad (21)$$

where $V_e = (2/3)(A_0 / R_0)(dR/dA)_0 F_{r0}$ is the Vedernikov number. For the case of the third value of ω, the solutions for the perturbation quantities are stable when $G \geq 0$.

2.4 Mass conservation

In order to evaluate whether the GWCE continuous formulation satisfies the mass conservation principle, let us substitute (5) in $\mathcal{L}(A,Q)$. Since $\mathcal{L}(\cdot,\cdot)$ is a linear operator, we get:

$$\mathcal{L}(A,Q) = \mathcal{L}(\overline{A},\overline{Q}) + \mathcal{L}(a,q) \quad (22)$$

Mass conservation requires that $\mathcal{L}(A,Q) = 0$. Since, according to (5), $a = o(\overline{A})$ and $q = o(\overline{Q})$, the last expression would require that both $\mathcal{L}(\overline{A},\overline{Q}) = 0$ and $\mathcal{L}(a,q) = 0$. Now, from (1), (13) and (14):

$$\mathcal{L}(a,q) = \partial_t a + \partial_x q = (-i\omega\hat{a} + ik\hat{q})e^{i(kx-\omega t)} \quad (23)$$

when the fist two values of ω, given by (20), are substituted in (23) and $\hat{q}$ is expressed in terms of â through the use of the second equation of the system (15), one gets the result $\mathcal{L}(a,q) = 0$, which means that the corresponding solutions for a and q conserve mass. Nevertheless, when the same procedure is employed, but now using the third value of ω, the following result is obtained:

$$\mathcal{L}(a,q) = \left\{ -G - ik\left[\frac{g\left(S_{f0} - F_2\right) + ik\left(gD_0 - U_0^2\right)}{-G + F_1 + 2iU_0 k}\right]\right\} \times \\ \times \hat{a}e^{-Gt}e^{ikx} \neq 0 \quad (24)$$

As is evident from the observation of (24), the solution for *a* and *q* corresponding to the third value of ω does not conserve mass, unless $G \to \infty$, which corresponds to the case when the GWCE formulation approaches the primitive formulation. Furthermore, Eq. (24) reveals that for finite values of *G*, when the wavenumber *k* increases, the mass conservation error in the GWCE formulation is expected to increase.

Therefore, since the general solution for a and q is expressed as a linear combination of the solutions corresponding to each of the three values of ω, the GWCE continuous formulation is not consistent with the mass conservation principle.

3 ONE-DIMENSIONAL DISCRETE GWCE FORMULATION

A time semidiscretization of the GWCE formulation (4) and (2) may be constructed by employing a weighted central difference approximation on (4) and a Crank-Nicolson approximation on (2) (Kinmark 1986; Luettich et al. 1992):

$$\frac{A^{n+1} - 2A^n + A^{n-1}}{\Delta t^2} + G\,\frac{A^{n+1} - A^{n-1}}{2\Delta t} -$$

$$\frac{1}{4}\left\{\partial_{xx}\left(\frac{Q^2}{A}\right)^{n+1}+g\partial_x\left[A^{n+1}(\partial_x H^{n+1}+S_f^{n+1})\right]\right\}-$$
$$\frac{1}{2}\left\{\partial_{xx}\left(\frac{Q^2}{A}\right)^{n}+g\partial_x\left[A^{n}(\partial_x H^{n}+S_f^{n})\right]\right\}- \qquad (25)$$
$$\frac{1}{4}\left\{\partial_{xx}\left(\frac{Q^2}{A}\right)^{n-1}+g\partial_x\left[A^{n-1}(\partial_x H^{n-1}+S_f^{n-1})\right]\right\}+$$
$$G\left(\frac{A^{n+1}-A^{n-1}}{\Delta t}+\frac{1}{4}\partial_x\left(Q^{n+1}+2Q^{n}+Q^{n-1}\right)\right)=0$$

$$2\frac{Q^{n+1}-Q^{n}}{\Delta t}+\partial_x\left(\frac{Q^2}{A}\right)^{n+1}+\partial_x\left(\frac{Q^2}{A}\right)^{n}+ \qquad (26)$$
$$gA^{n+1}\left(\partial_x H^{n+1}+S_f^{n+1}\right)+gA^{n}\left(\partial_x H^{n}+S_f^{n}\right)=0$$

where the superindices denote time level and Δt, a time interval.

Full discretization on the system (25)-(26) may be achieved by employing a linear finite element spatial approximation. When a constant element size equal to Δx is used, an approach similar to that explained for the continuous GWCE formulation may be employed. Thus, the solutions for area and discharge in the discrete system are decomposed in a reference solution and a perturbation. This decomposition is substituted in the discrete equations and a Taylor-Fréchet expansion of the result is performed. Higher order terms are neglected and the linearized equations are localized. Hence, a linear and constant coefficient system is obtained. A Fourier analysis may be performed on that system and a discrete dispersion relation may be derived. Once again, three values for the discrete frequency (or, equivalently, for the amplification factor) are obtained.

A mass conservation error may be derived in a manner similar to that described for the continuous case. The corresponding expression is rather lengthy and, for that reason, is not included here. Denoting the coefficient of $\hat{a}e^{-Gt}e^{ikx}$, cp. Eq. (24), in the expression for the mass conservation error corresponding to the discrete GWCE formulation with E_{mj}, where the subindex j denotes the first, second or third value of the discrete frequency (or amplification factor), a relation of the following form may be found:

$$\left\|E_{mj}\right\|_\infty = \max_{k\Delta x}\left|E_{mj}\right|;\quad j=1,2,3 \qquad (27)$$

where E_{mj} depends on the dimensionless parameters $k\Delta x$, $G\Delta t$, $U_o\Delta t/\Delta x$, F_{r0}, V_e and $gS_{f_0}\Delta t/U_0$. The values of $\left\|E_{m1}\right\|_\infty$ and $\left\|E_{m2}\right\|_\infty$ for a wide range of the parameters that these norms depend on are relatively small (of the order of 10^{-4}). Nevertheless, the values of $\left\|E_{m3}\right\|_\infty$ are significant. The behavior of this norm as a function of $G\Delta t$ is shown in Figure 1, for $U_0\Delta t/\Delta x = 0.2901$, $F_{r0}=0.4087$, $V_e=0.2609$ and $gS_{f0}\Delta t/U_0 = 0.0021$.

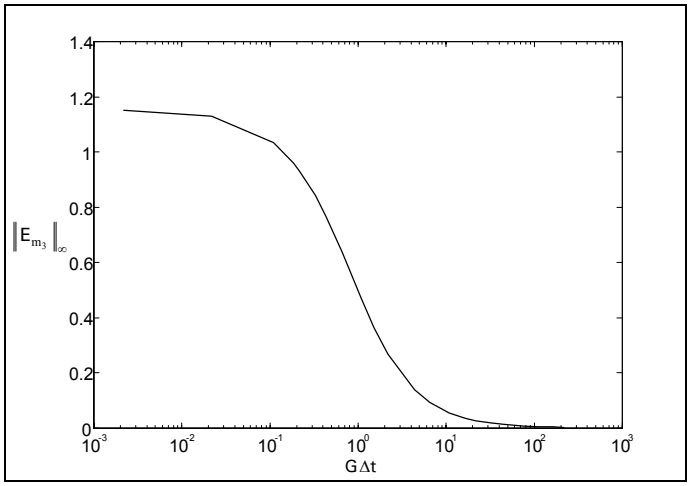

Figure 1. Infinity norm of the mass conservation error as a function of $G\Delta t$.

As expected, the mass conservation performance of the GWCE discrete formulation deteriorates with increasing values of *G*. The same type of behavior is observed for other values of the dimensionless parameters on which the mass conservation error depends. This confirms the findings presented earlier in relation to the lack of consistency of the GWCE formulation with the mass conservation principle.

4 VALIDATION

To validate the theoretical results presented in the previous sections, a series of simulations were conducted on four different one-dimensional domains: 1) a constant bathymetry channel that is 5 m deep and 50 km long; 2-3) a parabolic bathymetry with two different rates of rise, starting at 200 m in the deep ocean and rising to 3 m on the shelf; 4) east coast bathymetry that represents a reach of ocean from the east coast of the United States to the mid-Atlantic (minimum and maximum depths are 20 m and 5000 m). All simulations were conducted with a fully nonlinear finite element code based on the GWCE algorithm. Common conditions for the simulations were as follows: constant element size with a minimum wavelength to grid spacing ratio of

200 (for the M2 wave); land boundary on one end of the domain (zero normal flux); open ocean boundary on the other end of the domain with a 1 meter M2 forcing (principal semi-diurnal tidal constituent); no eddy viscosity; constant bottom friction parameter, τ, of 10^{-4}/s; constant time step of 10 seconds. For each, the model was spun up from a cold start over one M2 tidal cycle and the output was recorded over a second M2 cycle.

Mass conservation was evaluated using a direct integration of the primitive continuity equation (1) over the period of record. In particular, the two terms of equation (1) represent accumulation and net flux. Correspondingly, mass imbalance is recorded as the average difference of the absolute value between the net flux and accumulation terms over the period of record. This error measure was denoted as E_{avg}. For perfect mass balance, E_{avg} should equal zero. To facilitate comparison between domains (where conservation errors can vary by an order of magnitude), all errors were scaled so that their range of variation would be between 0 and 1, by using the following linear mapping:

$$Scaled\ Error = \frac{E_{avg} - E_{avg,min}}{E_{avg,max} - E_{avg,min}} \tag{28}$$

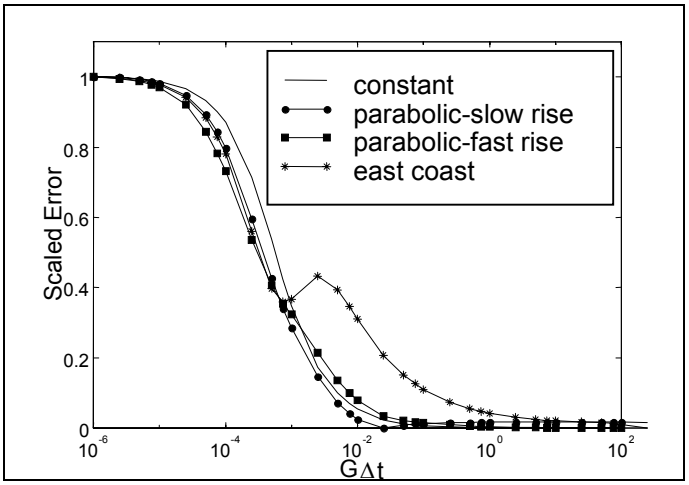

Figure 2 Scaled Error versus $G\Delta t$ for different domains.

Figure 2, which shows the Scaled Error versus $G\Delta t$ for each domain, illustrates the results of the validation. Of significance is the fact that the Scaled Error for all domains goes to zero as $G\Delta t$ gets large (we note that the absolute error, as measured by E_{avg}, also approaches zero). This is in accord with the theoretical results of equations 24 and 27 and the graph of Figure 1, which indicate that mass balance errors decrease as G increases. Also noteworthy is that the error plots in both Figures 1 and 2 follow the same pattern, i.e., little change in error for extremely low or high values of $G\Delta t$ with a smooth transition in between, that occurs over two to three log cycles (we are still investigating the cause of small "hiccup" in the east coast plot). One difference is that in Figure 2 the transition occurs between 10^{-4} to 10^{-2}, while in Figure 1, the transition occurs between 10^{-1} and 10^{1}. We attribute this to the difference in parameter values between the two graphs.

Finally, it is important to note that mass balance is only part of the picture. In effect, as G is increased, the GWCE equation becomes more primitive; if increased too much, the spurious oscillations that plague the primitive finite element formulation begin to appear. Thus, it should not be inferred that the "best" solution is obtained by choosing arbitrarily large $G\Delta t$ values. Rather, the overall quality of the solution, in terms of mass balance and numerical noise control, is very sensitive to the value of the parameter G, which should be chosen with discretion. Referring to the validation simulations of Figure 2, we note that most of the error is removed with a $G\Delta t$ value of 10^{-2}. With a Δt=10 seconds, this gives a G value of 10^{-3}/second, or a G/τ ratio of 10. Based on dispersion studies presented in Kolar et al. (1994), this ratio corresponds with the recommended upper bound on G in order to prevent spurious oscillations.

5 CONCLUSIONS

The GWCE formulation of the shallow water equations was developed in order to control spurious spatial oscillations that arise in numerical results generated by finite element discretizations of the primitive equations of motion. Even though success has been achieved in controlling the said oscillations via the GWCE formulation, it has also been observed that mass conservation errors arise in its application.

The present paper addresses the above described issue by employing a Taylor-Fréchet expansion of the operators corresponding to the GWCE and the momentum equation. The result is linearized and localized, thus making feasible the use of Fourier analysis. Hence a dispersion relation is obtained for both the continuous and the discrete versions of the GWCE formulation. It is shown that this formulation does not conserve mass and that the mass conservation error decreases as the G parameter increases. This result was expected as in the limit when $G \to \infty$, the GWCE formulation approaches the primitive formulation. Four numerical tests that confirm the theoretical findings are also included.

6 ACKNOWLEDGMENTS

Funding was provided in part by the National Science Foundation, Grant # ACI-9623592. We would also like to acknowledge the help of our students, Kendra Dresback and Casey Dietrich, in conducting the simulations. The assistance of Francisco Salinas in typesetting the manuscript is also acknowledged.

7 REFERENCES

Aldama, A. & C. Paniconi 1992. An Analysis of the Convergence of Picard Iterations for Implicit Approximations of Richard's Equation. *Proc.* IX Int. Conf. Comp. Meth. Water Res. 2: 521-528. Southampton: Comp. Mech. Publications

Aldama, A.A. &.A. Aguilar 1996. Stability Analysis of a General Preissmann Scheme. *Proc.* XI Int. Conf. Comp. Meth. Water Res. 2: 37-44. Southampton: Comp. Mech. Publications.

Aldama, A.A. & A. Ocón 1998. Algunas reflexiones sobre la resistencia al flujo en canales y sobre la fórmula de Manning. *Memorias del XV Congreso Nacional de Hidráulica*. Oaxaca, México.

Aldama, A. & J. Aparicio 1998. The effect of nonlinearities in the stability of numerical solutions of Richards equation. *Proc.* XII Int. Conf. Comp. Meth. Water Res. 1: 289-296. Southampton: Comp. Mech. Publications.

Kinnmark, P.E. & W. G. Gray 1985. The $2\Delta x$-test: A Tool for Analyzing Spurios Oscillations. *Adv. Water Resources*. (8):129-135.

Kinnmark, I 1986. *The Shallow Water Wave Equations: Formulation, Analysis and Application*. Berlin Heidelberg: Springer-Verlag, Lecture Notes in Engineering 15. 187 pp.

Kolar, R.L., J.J. Westerink, M.E. Cantekin & C.A. Blain 1994. Aspects of Nonlinear Simulations Using Shallow Water Models Based on the Wave Continuity Equation. *Computers and Fluids*. 23(3): 523-538.

Luettich, R. A., J.J. Westerink & N. W. Scheffner 1992. *ADCIRC: An Advanced Three-dimensional Circulation Model for Shelves, Coasts and Estuaries*, U.S. Whashington D.C.: Army Corps of Engineers, Technical report DRP-92-6. 137 pp.

Lynch D.R. & W.G. Gray 1979. A Wave Equation Model for Finite Element Tidal Computations, *Computers and Fluids*. (7): 207-228.