**Hervouet, Jean-Michel; Pavan, Sara; Ata, Riadh**

# Distributive advection schemes and dry zones, new solutions

# Distributive advection schemes and dry zones, new solutions

Jean-Michel Hervouet [(1)], Sara Pavan [(1,2)], Riadh Ata [(1,2)]
[(1)] Laboratoire National d'Hydraulique et Environnement, Electricité de France
[(2)] Laboratoire d'Hydraulique Saint-Venant
6 Quai Watier, 78400 Chatou, France
Email of corresponding author: j-m.hervouet@edf.fr

***Abstract*—This paper is a continuation of "Ongoing research on advection schemes", published in 2014 in this series of proceedings. It is restricted to distributive schemes and comes after the description of the new predictor-corrector introduced in the previous paper. The developments and tests were done with Telemac-2D but can be easily applied also to 3D. First a second order in time version of this predictor-corrector is developed. Then a new criterion for proving monotonicity is coined, which allows to perform as many correction steps as we want, with an arbitrary predictor which is just maintained within a given range and is not even subjected to mass conservation. With 4 extra correction steps the rotating cone grows from 0.5331 to 0.75. At this level the problem of dry zones still remains. To solve it, it is first shown that a fully implicit distributive scheme is unconditionally stable, even on dry zones. However the numerical diffusion is largely increased, losing all the benefits previously gained. Then a locally implicit predictor-corrector scheme is designed, with full implicitation only in the dry zones. An unexpected consequence of this new scheme is that we can choose an arbitrary time-step, and this allows to use the distributive schemes in conditions where they perform better, e.g. the rotating cone height after one rotation is now 0.79 in the latest tests. This is much larger than the 0.39 of the NERD scheme which was before the only distributive scheme working with tidal flats. A new test case with bridge piers and an island treated as a dry zone is presented. Monotonicity is well preserved and mass conservation is obtained at machine accuracy.**

## I. Introduction

Mass conservation, monotonicity and dry zones are now fairly well handled in the Telemac system, so that the numerical diffusion of advection schemes becomes the new frontier where progress is necessary to improve the quality of studies. For example the study of pollutants in rivers, the stability of stratifications, and the numerical simulation of non linear waves are highly dependent on the quality of advection schemes, and on the space and time orders. Improving on this topic is not an easy task, since on one hand a couple of theorems show that simple linear schemes cannot do the job, and on the other hand this subject has been already heavily investigated by many teams. Moreover we face additional problems due to the free-surface flows, like the depth-averaged or moving grid context, and still the treatment of tidal flats, that at first sight precludes most existing solutions, since divisions by the depth appear in many solution procedures.

In the 2014 Telemac User Club we presented several improvements. In finite volumes an approximate Riemann solver, the Harten-Lax-van Leer-Contact scheme (HLLC, see [12]) with $1^{st}$ and $2^{nd}$ order was presented. In finite elements, the classical N and PSI distributive schemes could be improved by adding the derivative in time in the upwinding process. It was done in a predictor-corrector procedure, after the recent publication by Mario Ricchiuto [11]. The predictor gives an approximation of the derivative in time of the tracer, which is then used in the corrector step. Three test cases were presented: a pollutant plume in a steady state river, the transport of a stain, and the rotating cone. The height of the cone after one rotation, which should theoretically be 1, was 0.2136 for the classical PSI scheme, 0.4710 for the HLLC second order scheme, and 0.5331 for the new predictor-corrector PSI scheme. The conclusion of this first paper announced: "We now work on tidal flats, which could be dealt with by an implicit predictor-corrector distributive scheme, as shown by preliminary tests not treated here. Another promising issue is the possibility of iterating the corrector step, which would give even less numerical diffusion, which is also shown by preliminary tests". The present paper will now detail in a sequence the three main improvements obtained since the first paper: a second order in time predictor-corrector scheme, then the possibility of iterating the corrections, and in the end a new approach, a locally implicit predictor-corrector distributive scheme. The rotating cone test and a new test case with bridge piers and an island will show the new features. All the developments and tests are done with Telemac-2D but the theory applies also to 3D, as the varying volumes around points in 3D play the same mathematical role as the varying depth in 2D.

## II. A second order in time predictor-corrector distributive scheme

In the previous paper we reported theoretical mass conservation problems to get a second order in time predictor-corrector scheme in the depth-averaged context, as was done in a simpler context by Ricchiuto in his original paper. We now have found a correct derivation, with boundary and source terms now always taken into account in all the steps. We start from the same predictor step, which is the classical PSI scheme:

$$\frac{S_i h_i^{n+1} C_i^* - S_i h_i^{n+1} C_i^n}{\Delta t} =$$

$$-\sum_j \min(\Phi_{ij}^{psi}(C^n), 0)\left(C_j^n - C_i^n\right) \tag{1}$$

$$-\min(b_i, 0)\left(C_i^{boundary} - C_i^n\right)$$

$$+\max(Sce_i, 0)\left(C_i^{sce} - C_i^n\right)$$

We recall that $h_i^n$ and $h_i^{n+1}$ are respectively the depths at point $i$ at the beginning and at the end of the time step, $S_i$ is the integral of the test function, $\Phi_{ij}^N$ and $\Phi_{ij}^{psi}$ are the fluxues between points given by respectively the N and PSI scheme. $C_i^n$ is the initial value of the tracer at point $i$, $C_i^{n+1}$ the final value, and $C_i^*$ the value at the predictor step. $\Delta t$ is the time step, $b_i$ is the boundary flux if $i$ is on a boundary and $Sce_i$ a possible source term inside the domain, while $C_i^{boundary}$ is the prescribed value of $C$ at the boundary, and $C_i^{sce}$ the value of the tracer at a source.

The rather long derivation of the corrector step will not be given here, it is obtained with the construction of a fully implicit and a fully explicit scheme, and then by blending them with the implicitation coefficient $\theta$. When $C_i^{n+1}$ is involved in the fluxes, it is replaced by $C_i^*$, which does not spoil the mass conservation if this is correctly done at the level of the conservative form. We eventually find the following equation, which is by construction mass conservative:

$$S_i h_i^{n+1}\left(C_i^{n+1} - C_i^*\right) =$$

$$-\theta\overleftarrow{S_i h_i^n\left(C_i^* - C_i^n\right)} - (1-\theta)\,\overleftarrow{S_i h_i^{n+1}\left(C_i^* - C_i^n\right)}$$

$$-\theta\Delta t\overleftarrow{\sum_j \min(\Phi_{ij}^{psi}(C^*), 0)\left(C_j^* - C_i^*\right)} \tag{2}$$

$$-(1-\theta)\,\Delta t\overleftarrow{\sum_j \min(\Phi_{ij}^{psi}(C^n), 0)\left(C_j^n - C_i^n\right)}$$

$$+Sce_i\ \Delta t\left(C_i^{sce} - (1-\theta)\,C_i^n - \theta C_i^*\right)$$

$$-b_i\ \Delta t\left(C_i^{boundary} - (1-\theta)\,C_i^n - \theta C_i^*\right)$$

Backward arrows are put on terms which are treated altogether with upwinding, at element level, in the same way that leads from N to PSI scheme. At element level derivatives in time are first equally shared between the 3 points of the triangle, this is considered to be the equivalent of a N scheme, then the PSI limitation is applied to the whole contribution that includes the fluxes. Mass conservation is rather easy to prove, with the help of the discretised continuity equation, but a proof of monotonicity was impossible to find, unless some restrictions are applied to $C^*$, namely that $C^*$ is not too far from $C^n$, and this idea will be also used for iterating the corrector. A very important point is that the mass conservation is ensured whatever the mass of $C^*$, because it is both in the left- and righ-hand side and can be cancelled, except in fluxes that do not contribute to a change of mass. The monotonicity proof can thus be done with an arbitrary $C^*$. We write the corrector in the following way, as already done in the previous paper:

$$S_i h_i^{n+1} C_i^{n+1} = S_i h_i^{n+1} C_i^* - f_i S_i h_i^{n+1-\theta}\left(C_i^* - C_i^n\right)$$

$$-(1-\theta)\Delta t\sum_j \mu_j\left(C_j^n - C_i^n\right)\min(\Phi_{ij}^{psi}(C^n), 0)$$

$$-\theta\Delta t\sum_j \mu_j\left(C_j^* - C_i^*\right)\min(\Phi_{ij}^{psi}(C^*), 0) \tag{3}$$

$$+\Delta t\max(Sce_i, 0)\left(C_i^{sce} - (1-\theta)\,C_i^n - \theta C_i^*\right)$$

$$-\Delta t\min(b_i, 0)\left(C_i^{boundary} - (1-\theta)\,C_i^n - \theta C_i^*\right)$$

All $f_i$ and $\mu_j$ are in the range [0,1] to account for the upwinding limitation. $h_i^{n+1-\theta}$ is a notation for $(1-\theta)\,h_i^{n+1} + \theta h_i^n$. Note that if $C^* = C^n$ we fall back to the classical N or PSI scheme, which is stable, so we can expect to keep this stability if $C_i^*$ is chosen not too far from $C_i^n$. We now want to have positive coefficients for all values of $C$ in the right-hand side. Only the coefficients of $C_i^*$ and $C_i^n$ are questionable. They are:

Coefficient of $C_i^*$:

$$a^* = S_i h_i^{n+1} - f_i S_i h_i^{n+1-\theta}$$

$$+\theta\Delta t\sum_j \mu_j\min(\Phi_{ij}^{psi}(C^*), 0) \tag{4}$$

$$-\theta\Delta t\left(\max(Sce_i, 0)\ - \min(b_i, 0)\right)$$

Coefficient of $C_i^n$:

$$a^n = f_i S_i h_i^{n+1-\theta}$$

$$+(1-\theta)\Delta t\sum_j \mu_j\min(\Phi_{ij}^{psi}(C^n), 0) \tag{5}$$

$$-(1-\theta)\,\Delta t\left(\max(Sce_i, 0)\ - \min(b_i, 0)\right)$$

$a^*$ or $a^n$ may be negative but the positivity of $a^* + a^n$ is largely ensured by the stability condition of the predictor, as we have:

$$a^* + a^n = S_i h_i^{n+1} + \theta\Delta t\sum_j \mu_j\min(\Phi_{ij}^{psi}(C^*), 0)$$

$$+ (1-\theta)\,\Delta t \sum_j \mu_j \min(\Phi_{ij}^{psi}(C^n), 0) \tag{6}$$

$$+\Delta t \left[ -\max(Sce_i, 0)\ + \min(b_i, 0) \right]$$

As a matter of fact, we can take $\mu_j = 1$ (worst case scenario), and replace the $\Phi_{ij}^{psi}(C^*)$ and $\Phi_{ij}^{psi}(C^n)$ by $\Phi_{ij}^N$, and we fall back to the classical stability condition.

We now write:

$$C_i^* = C^{\min} + \alpha \left( C^{\max} - C^{\min} \right) \tag{7}$$

$$C_i^n = C^{\min} + \beta \left( C^{\max} - C^{\min} \right) \tag{8}$$

with $\alpha$ and $\beta$ in the range [0,1]. $C^{\min}$ and $C^{\max}$ are the local extrema that should not be trespassed, computed with the neighbouring values of $C^n$ and $C^*$. We want to find the solutions under which:

$$a^* C_i^* + a^n C_i^n = (a^* + a^n) C_i^{average} \tag{9}$$

with: $C_i^{average} = C^{\min} + \gamma \left( C^{\max} - C^{\min} \right)$, and $\gamma$ in the range [0,1]. In fact there is not always a solution, even with very small time steps, and we had to change the strategy. Choosing $\theta = \frac{1}{2}$ and under the stability condition of the first order in time of the predictor-corrector, we looked for a condition on $\alpha$ as a function of $\beta$, and it gave:

$$2\beta - 1 \le \alpha \le 2\beta \tag{10}$$

$$\frac{\beta}{3} \le \alpha \le \frac{2}{3} + \frac{\beta}{3} \tag{11}$$

which is equivalent to:

$$2C_i^n - C^{\max} \le C_i^* \le 2C_i^n - C^{\min} \tag{12}$$

$$\frac{2C^{\min}}{3} + \frac{C_i^n}{3} \le C_i^* \le \frac{2C^{\max}}{3} + \frac{C_i^n}{3} \tag{13}$$

Our solution resorts to imposing these conditions to every $C_i^*$, which, as we have said, does not spoil the mass conservation even if we change the mass of $C^*$. In some severe conditions, when the restrictions apply, the second order will simply not be reached.

## III. Iterating the corrections

We have shown in the previous section that any predictor value can be used in the corrector step, provided that it remains within a certain distance from the initial value $C^n$. The corrector can thus be applied as many times as we want, taking every time as new predictor the value of the last iteration. The same principle can be applied also to the first order in time predictor-corrector scheme, but the condition appears to be different:

$$C_i^n + \frac{C^{\min} - C_i^n}{2} \le C_i^* \le C_i^n + \frac{C^{\max} - C_i^n}{2} \tag{14}$$

It can also be shown that this condition is naturally ensured by the PSI scheme which is our predictor, so the limitation does not need to be applied at the first iteration. Iterating the corrector proves to be very efficient, as shown by the rotating cone test. We recall that in this case the mesh is a 20.1 m x 20.1 m square composed of 4489 squares of side 0.3 m, each one split into two triangles. With the first order scheme we find after one rotation:

| number of corrections | cone height after one rotation |
|---|---|
| 0 | 0.21 (PSI scheme) |
| 1 | 0.53 |
| 2 | 0.69 |
| 3 | 0.74 |
| 4 | 0.75 |
| 21 | 0.75 |

It seems that we have rapidly a dramatic improvement, after very few iterations of the corrector. The state-of-the-art obtained last year, 0.53, is boosted to 0.75. Comparing order 1 and order 2 of the N predictor-corrector with corrections scheme yields:

| corrections | cone height, order 1 | cone height, order 2 |
|---|---|---|
| 0 | 0.18 (N scheme) | 0.18 (N scheme) |
| 1 | 0.50 | 0.48 |
| 2 | 0.68 | 0.60 |
| 3 | 0.74 | 0.63 |
| 4 | 0.75 | 0.64 |
| 5 | 0.76 | 0.64 |
| 6 | 0.77 | 0.65 |

Figure 1, for order 1 and Figure 2 for order 2 show the cone after one rotation of the N predictor-corrector with six corrections. The shape is different but there is no clear advantage of order 2 in this case. However the convergence tests, not shown in this paper, show the gain in order, though order 2 is not exactly achieved, as was already found with unstructured meshes.

## IV. Dry zones: a locally implicit predictor-corrector scheme

It can be shown that when the tracer is semi-implicited in the fluxes with a coefficient $\theta$, the stability criterion on the time is divided by $1 - \theta$ and becomes:

$$\Delta t < \frac{1}{1-\theta} \frac{S_i h_i^n}{\left( \sum_j \max\left(\Phi_{ij}, 0\right) + \max\left(b_i, 0\right) - \min\left(Sce_i, 0\right) \right)} \tag{15}$$

A fully implicit distributive scheme becomes unconditionnally stable, even on dry zones. However tests show that such a scheme is far too diffusive. This is why we looked for a

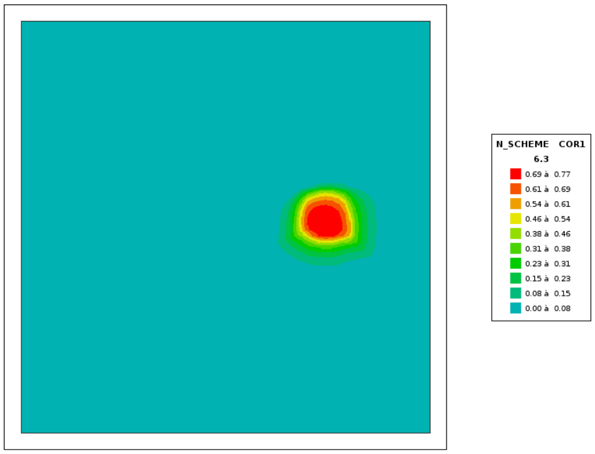

Fig. 1. Rotating cone test, cone after one rotation. N predictor-corrector with 6 corrections, order 1.

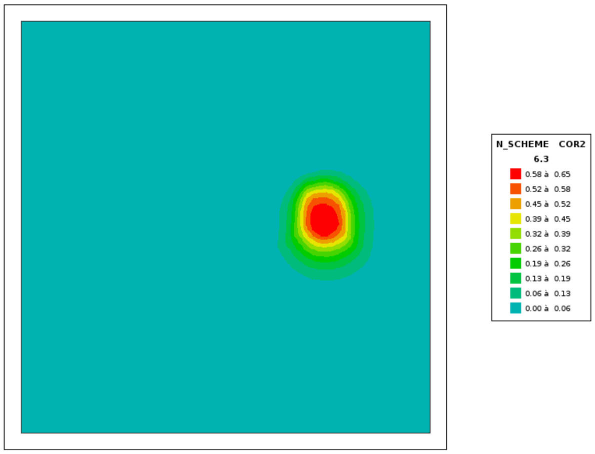

Fig. 2. Rotating cone test, cone after one rotation. N predictor-corrector with 6 corrections, order 2.

scheme that would be locally implicit, with full implicitation only on dry zones.

### A. Semi-implicit predictor

We choose to solve in the predictor step the following equation:

$$
\begin{aligned}
& S_i h_i^{n+1-\theta_i} C_i^* - S_i h_i^{n+1-\theta_i} C_i^n = \\
& -\Delta t \sum_j \left(\theta_j C_j^* + (1-\theta_j)\, C_j^n\right) \min\left(\Phi_{ij}^N, 0\right) \\
& +\Delta t \sum_j \left(\theta_i C_i^* + (1-\theta_i)\, C_i^n\right) \min\left(\Phi_{ij}^N, 0\right) \qquad (16) \\
& +\Delta t \max\left(Sce_i, 0\right)\left(C_i^{sce} - (\theta_i C_i^* + (1-\theta_i) C_i^n)\right) \\
& -\Delta t \min\left(b_i, 0\right)\left(C_i^{boundary} - (\theta_i C_i^* + (1-\theta_i) C_i^n)\right)
\end{aligned}
$$

### B. Corrector

Now that we have an approximation $C_i^*$ of the final concentration, we can write the original derivative in time in the form:

$$
S_i h_i^{n+1-\theta_i}\left(C_i^{n+1} - C_i^* + C_i^* - C_i^n\right) \qquad (17)
$$

where the term $S_i h_i^{n+1-\theta_i}\left(C_i^* - C_i^n\right)$ can be transfered in the right-hand side. Separating the contribution of fluxes between explicit and implicit terms, we get:

$$
\begin{aligned}
& S_i h_i^{n+1-\theta_i}\left(C_i^{n+1} - C_i^*\right) = -S_i h_i^{n+1-\theta_i}\left(C_i^* - C_i^n\right) \\
& -\Delta t \sum_j \left(\theta_j C_j^{n+1} - \theta_i C_i^{n+1}\right) \min\left(\Phi_{ij}^N, 0\right) \\
& -\Delta t \sum_j \left((1-\theta_j)\, C_j^n - (1-\theta_i)\, C_i^n\right) \min\left(\Phi_{ij}^N, 0\right) \qquad (18) \\
& +\Delta t \max\left(Sce_i, 0\right)\left(C_i^{sce} - \left(\theta_i C_i^{n+1} + (1-\theta_i) C_i^n\right)\right) \\
& -\Delta t \min\left(b_i, 0\right)\left(C_i^{boundary} - \left(\theta_i C_i^{n+1} + (1-\theta_i) C_i^n\right)\right)
\end{aligned}
$$

We now want to add upwinding to the derivative in time, and we also include in the upwinding the explicit part of the flux contributions. It gives, still using our backward arrays notation:

$$
\begin{aligned}
& S_i h_i^{n+1-\theta_i}\left(C_i^{n+1} - C_i^*\right) = \overleftarrow{-S_i h_i^{n+1-\theta_i}\left(C_i^* - C_i^n\right)} \\
& -\Delta t \sum_j \left(\theta_j C_j^{n+1} - \theta_i C_i^{n+1}\right) \min\left(\Phi_{ij}^N, 0\right) \\
& \overleftarrow{-\Delta t \sum_j \left((1-\theta_j)\, C_j^n - (1-\theta_i)\, C_i^n\right) \min\left(\Phi_{ij}^N, 0\right)} \qquad (19) \\
& +\Delta t \left(\max\left(Sce_i, 0\right)\left(C_i^{sce} - \left(\theta_i C_i^{n+1} + (1-\theta_i) C_i^n\right)\right)\right) \\
& -\Delta t \min\left(b_i, 0\right)\left(C_i^{boundary} - \left(\theta_i C_i^{n+1} + (1-\theta_i) C_i^n\right)\right)
\end{aligned}
$$

Note that a tentatively second order upwinded contribution should be:

$$
\begin{aligned}
& \overleftarrow{-S_i h_i^{n+1-\theta_i}\left(C_i^* - C_i^n\right)} \\
& \overleftarrow{-\Delta t \sum_j \left(\theta_j C_j^* + (1-\theta_j) C_j^n\right) \min\left(\Phi_{ij}^{psi}, 0\right)} \qquad (20) \\
& \overleftarrow{+\Delta t \sum_j \left(\theta_i C_i^* + (1-\theta_i) C_i^n\right) \min\left(\Phi_{ij}^{psi}, 0\right)}
\end{aligned}
$$

but it is not what is naturally given by the derivation, the reason being that this would lead to mass errors, because $\Phi_{ij}^{psi}$ is built with $C^n$ and can replace $\Phi_{ij}^N$ safely only when used with $C^n$, not with $C^*$.

### C. Monotonicity

As the mass is correct by construction, the only remaining question is the monotonicity. We now rewrite our corrector step so that only positive coefficients of values of $C$ appear. We also introduce coefficient $f_i$ and $\mu_{ij}$ as before to account for the PSI reduction of the upwinded terms, it yields:

$$\left(S_i h_i^{n+1-\theta_i} - \theta_i \Delta t \sum_j \min\left(\Phi_{ij}^N, 0\right)\right) C_i^{n+1}$$
$$+\theta_i \Delta t \left(\max\left(Sce_i, 0\right) - \min\left(b_i, 0\right)\right) C_i^{n+1} =$$
$$\Delta t \left(\max\left(Sce_i, 0\right) C_i^{sce} - \min\left(b_i, 0\right) C_i^{boundary}\right)$$
$$-\Delta t \sum_j \theta_j C_j^{n+1} \min\left(\Phi_{ij}^N, 0\right) \tag{21}$$
$$-\mu_{ij} \Delta t \sum_j \left(1-\theta_j\right) C_j^n \min\left(\Phi_{ij}^N, 0\right)$$
$$+C_i^* \left(1-f_i\right) S_i h_i^{n+1-\theta_i} + C_i^n f_i S_i h_i^{n+1-\theta_i}$$
$$+(1-\theta_i) C_i^n \Delta t \sum_j \mu_{ij} \min\left(\Phi_{ij}^N, 0\right)$$
$$-(1-\theta_i) C_i^n \Delta t \left[\max\left(Sce_i, 0\right) - \min\left(b_i, 0\right)\right]$$

With this form we see that the only risk of negative coefficients happens with $C_i^n$. The coefficient of $C_i^{n+1}$ is positive thanks to the stability condition that has been previously chosen. Without the extra derivative in time, we would have to ensure the positivity of:

$$B_{ii} = S_i h_i^{n+1-\theta_i} - \Delta t (1-\theta_i)\ flux(i)$$

Denoting:

$$flux(i) = \max\left(Sce_i, 0\right) - \sum_j \min\left(\Phi_{ij}^N, 0\right) - \min\left(b_i, 0\right) \tag{22}$$

which leads to the condition:

$$\Delta t < \frac{1}{1-\theta_i} \frac{S_i h_i^n}{\left(flux(i) + \sum_j \Phi_{ij} + b_i - Sce_i\right)} \tag{23}$$

Now we see that there is a risk of negative coefficient of $C_i^n$, unless we consider a limitation of $C_i^*$. As the terms depending on $\mu_{ij}$ are negative in the coefficient of $C_i^n$ we remain on the safe side by choosing $\mu_{ij} = 1$. As before, we now introduce:

$$C_i^* = C^{\min} + \alpha \left(C^{\max} - C^{\min}\right) \tag{24}$$

$$C_i^n = C^{\min} + \beta \left(C^{\max} - C^{\min}\right) \tag{25}$$

We are left with proving that:

$$C_i^* \left(1-f_i\right) S_i h_i^{n+1-\theta_i} + C_i^n f_i S_i h_i^{n+1-\theta_i}$$
$$+(1-\theta_i) C_i^n \Delta t \sum_j \min\left(\Phi_{ij}^N, 0\right)$$
$$-(1-\theta_i) C_i^n \Delta t \left(\max\left(Sce_i, 0\right) - \min\left(b_i, 0\right)\right) = \tag{26}$$
$$S_i h_i^{n+1-\theta_i} C_i^{average}$$
$$+(1-\theta_i) \Delta t \sum_j \min\left(\Phi_{ij}^N, 0\right) C_i^{average}$$
$$-(1-\theta_i) \Delta t \left(\max\left(Sce_i, 0\right) - \min\left(b_i, 0\right)\right) C_i^{average}$$

we denote:

$$\gamma = S_i h_i^{n+1-\theta_i}$$
$$+(1-\theta_i) \Delta t \sum_j \min\left(\Phi_{ij}^N, 0\right) \tag{27}$$
$$-(1-\theta_i) \Delta t \left(\max\left(Sce_i, 0\right) - \min\left(b_i, 0\right)\right)$$

It eventually yields:

$$\gamma C_i^{average} = C_i^* \left(1-f_i\right) S_i h_i^{n+1-\theta_i}$$
$$+C_i^n \left(f_i S_i h_i^{n+1-\theta_i} + \gamma - S_i h_i^{n+1-\theta_i}\right) \tag{28}$$

or:

$$\gamma C_i^{average} =$$
$$\left(\gamma - S_i h_i^{n+1-\theta_i}\right) \left(C^{\min} + \beta \left(C^{\max} - C^{\min}\right)\right)$$
$$+S_i h_i^{n+1-\theta} \left(C^{\min} + \alpha \left(C^{\max} - C^{\min}\right)\right)$$
$$-\left(f_i S_i h_i^{n+1-\theta_i} \left(C^{\min} + \alpha \left(C^{\max} - C^{\min}\right)\right)\right) \tag{29}$$
$$+\left(f_i S_i h_i^{n+1-\theta_i} \left(C^{\min} + \beta \left(C^{\max} - C^{\min}\right)\right)\right)$$

which is:

$$C_i^{average} = C^{\min}$$

$$+\frac{\beta\left(\gamma - S_i h_i^{n+1-\theta_i}\right)}{\gamma}\left(C^{\max} - C^{\min}\right) \tag{30}$$

$$+\frac{\alpha\left(1-f_i\right)S_i h_i^{n+1-\theta_i} + \beta f_i S_i h_i^{n+1-\theta_i}}{\gamma}\left(C^{\max} - C^{\min}\right)$$

We see that need to have:

$$0 < \beta\gamma + \left(\alpha - \beta\right)\left(1 - f_i\right)S_i h_i^{n+1-\theta_i} < \gamma \tag{31}$$

If $\alpha > \beta$: positivity is ensured and then the worst situation happens when $f_i = 0$, in which case we get the condition:

$$\beta\gamma + \left(\alpha - \beta\right)S_i h_i^{n+1-\theta_i} < \gamma \tag{32}$$

which also reads:

$$\alpha S_i h_i^{n+1-\theta_i} < \gamma\left(1 - \beta\right) + \beta S_i h_i^{n+1-\theta_i} \tag{33}$$

We now assume that the time step was chosen so that:

$$\Delta t < \frac{1}{2\left(1 - \theta_i\right)}\frac{S_i h_i^n}{\left(flux(i) + \sum_j \Phi_{ij} + b_i - Sce_i\right)} \tag{34}$$

which gives the property:

$$\gamma > \frac{S_i h_i^{n+1-\theta_i}}{2} \tag{35}$$

Our most demanding condition for $\alpha$ is then (the smallest $\gamma$ is to be considered):

$$\alpha < \frac{1}{2} + \frac{\beta}{k} \tag{36}$$

If $\alpha < \beta$: only the positivity gives a condition and again the worst condition is $f_i = 0$ and we get the condition:

$$0 < \beta\gamma + \left(\alpha - \beta\right)S_i h_i^{n+1-\theta_i}$$

where the stronger condition, again obtained with the minimum $\gamma$, is:

$$\frac{\beta}{2} < \alpha \tag{37}$$

We end up with the general condition:

$$\frac{\beta}{2} < \alpha < \frac{1}{2} + \frac{\beta}{k} \tag{38}$$

which is also:

$$C_i^n + \frac{1}{2}\left(C^{\min} - C_i^n\right) < C_i^* < C_i^n + \frac{1}{2}\left(C^{\max} - C_i^n\right) \tag{39}$$

Now the next question is: is this property ensured by $C_i^*$ when we use a semi-implicit predictor? We have:

$$S_i h_i^{n+1-\theta_i} C_i^* - S_i h_i^{n+1-\theta_i} C_i^n =$$

$$\Delta t\left(\max\left(Sce_i, 0\right)\left(C_i^{sce} - \left(\theta_i C_i^* + (1-\theta_i)C_i^n\right)\right)\right)$$

$$-\Delta t\sum_j\left(\theta_j C_j^* + (1-\theta_j)C_j^n\right)\min\left(\Phi_{ij}, 0\right) \tag{40}$$

$$+\Delta t\sum_j\left(\theta_i C_i^* + (1-\theta_i)C_i^n\right)\min\left(\Phi_{ij}, 0\right)$$

$$-\Delta t\min\left(b_i, 0\right)\left(C_i^{boundary} - \left(\theta_i C_i^* + (1-\theta_i)C_i^n\right)\right)$$

which is equivalent to:

$$\left[S_i h_i^{n+1-\theta_i} + \theta_i\Delta t\left(-\min\left(\Phi_{ij}, 0\right)\right)\right]C_i^*$$

$$+\theta_i\Delta t\left(\max\left(Sce_i, 0\right) - \min\left(b_i, 0\right)\right)C_i^*$$

$$-\left[S_i h_i^{n+1-\theta_i} + \theta_i\Delta t\left(-\min\left(\Phi_{ij}, 0\right)\right)\right]C_i^n$$

$$-\theta_i\Delta t\left(\max\left(Sce_i, 0\right) - \min\left(b_i, 0\right)\right)C_i^n = \tag{41}$$

$$\Delta t\left(\max\left(Sce_i, 0\right)\left(C_i^{sce} - C_i^n\right)\right)$$

$$-\Delta t\sum_j\left(\theta_j C_j^* + (1-\theta_j)C_j^n - C_i^n\right)\min\left(\Phi_{ij}, 0\right)$$

$$-\Delta t\min\left(b_i, 0\right)\left(C_i^{boundary} - C_i^n\right)$$

Denoting:

$$\lambda = \Delta t\left(\max\left(Sce_i, 0\right) - \min\left(\Phi_{ij}, 0\right) - \min\left(b_i, 0\right)\right)$$

and remarking that in the right-hand side all terms $C_i^n$ are balanced by a $-C$ of some sort, we can write:

$$C_i^n + \frac{\lambda}{\left(S_i h_i^{n+1-\theta_i} + \theta_i\lambda\right)}\left(C_i^{\min} - C_i^n\right) < C_i^* \tag{42}$$

and:

$$C_i^* < C_i^n + \frac{\lambda}{\left(S_i h_i^{n+1-\theta_i} + \theta_i\lambda\right)}\left(C_i^{\max} - C_i^n\right) \tag{43}$$

The maximum of $\frac{\lambda}{\left(S_i h_i^{n+1-\theta_i} + \theta_i\lambda\right)}$ is obtained with the maximum of $\lambda$. Under the condition 34 this maximum is $\frac{1}{2+\theta_i}$ which is less than $\frac{1}{2}$. So we get indeed the property:

$$C_i^n + \frac{1}{k}\left(C_i^{\min} - C_i^n\right) < C_i^* < C_i^n + \frac{1}{k}\left(C_i^{\max} - C_i^n\right) \tag{44}$$

which is the condition found for the explicit predictor, and which could be even stricter if we impose a non zero minimum of $\theta_i$.

With $k = 2$ we arrive at:

$$C_i^n + \frac{1}{2}\left(C_i^{\min} - C_i^n\right) < C_i^* < C_i^n + \frac{1}{2}\left(C_i^{\max} - C_i^n\right) \quad (45)$$

which is identical to the property found for the explicit predictor. This long derivation shows that the locally implicit scheme basically behaves like the explicit option. However, we have so far only half of the monotonicity proof, because a new and unexpected problem occurs: the sum of the coefficients of values of $C$ is no longer correct after PSI reduction. This problem is addressed in the next paragraph.

### D. A correct sum of coefficients

It is easy to see that our final linear system is in the form $S_i h_i^{n+1-\theta_i} C_i^{n+1} = S_i h_i^{n+1-\theta_i} C_i^* +$ other terms which all contain well balanced differences of values of $C$, for example $\Delta t \left(\max\left(Sce_i, 0\right)\left(C_i^{sce} - C_i^n\right)\right)$. It can be deduced by this that we have in the end $C_i^{n+1}$ = a correct interpolation of values of $C$, with the sum of coefficients equal to 1. This is however not the case if such balanced terms are reduced by a PSI limitation in an unbalanced way. In what precedes it is the case with the term:

$$\overleftarrow{-\Delta t \sum_j \left(\left(1-\theta_j\right) C_j^n - \left(1-\theta_i\right) C_i^n\right) \min\left(\Phi_{ij}^N, 0\right)}$$

The balance of $\left(1-\theta_j\right) C_j^n - \left(1-\theta_i\right) C_i^n$ is ensured by terms $-\ \theta_j C_j^{n+1} - \theta_i C_i^{n+1}$ and this is no longer the case after PSI reduction of only the explicit part. We are thus doomed to reduce only true differences of $C$ values. In the case of term:

$$\overleftarrow{-\Delta t \sum_j \left(\left(1-\theta_j\right) C_j^n - \left(1-\theta_i\right) C_i^n\right) \min\left(\Phi_{ij}^N, 0\right)} \quad (46)$$

a solution consists in not upwinding all the terms, but only those that can be balanced in the PSI reduction, denoting:

$$\min\theta(i,j) = \min(1-\theta_j, 1-\theta_i) \quad (47)$$

we replace our term by:

$$-\Delta t \sum_j \left(\left(1-\theta_j - \min\theta(i,j)\right) C_j^n\right) \min\left(\Phi_{ij}^N, 0\right)$$

$$+\Delta t \sum_j \left(\left(1-\theta_i - \min\theta(i,j)\right) C_i^n\right) \min\left(\Phi_{ij}^N, 0\right) \quad (48)$$

$$\overleftarrow{-\Delta t \sum_j \min\theta(i,j)\left(C_j^n - C_i^n\right) \min\left(\Phi_{ij}^N, 0\right)}$$

This can be done at the element level when doing the PSI reduction.

### E. Choosing the local semi-implicitation

Assuming that the classical condition of the explicit N scheme gives the limitation:

$$\Delta t_{stab}(i) < \frac{S_i h_i^n}{\left(flux(i) + \sum_j \Phi_{ij} + b_i - Sce_i\right)} \quad (49)$$

which is the condition 23 with $\theta = 0$, and prescribing a number of $n$ steps into a time step $\Delta t$ we now want for the implicit predictor-corrector:

$$\frac{1}{1-\theta_i}\frac{\Delta t_{stab}(i)}{2} = \frac{\Delta t}{n} \quad (50)$$

which yields:

$$\theta_i = \max(0, 1 - \frac{n \Delta t_{stab}(i)}{2\Delta t}) \quad (51)$$

To get the same implicitation as the one step semi-implicit N we thus just need to multiply the number of time steps by 2.

Choosing the N scheme, a number of corrections of 5, the height of the rotating cone after 1 rotation, depending on the number of substeps $n$, gives:

| $n$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| height | 0.09 | 0.12 | 0.14 | 0.16 | 0.18 | 0.20 | 0.24 | 0.28 | 0.33 | 0.41 | 0.46 | 0.53 | 0.59 | 0.64 | 0.69 | 0.72 | 0.75 | 0.77 | 0.77 | 0.78 | 0.78 |

After $n = 20$ it gradually decreases, so 20 is an optimum. With $n = 20$, if we now vary the number of corrections we get:

| corrections | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cone height | 0.54 | 0.71 | 0.76 | 0.77 | 0.78 | 0.79 | 0.79 | 0.79 | 0.79 | 0.79 | 0.79 | 0.79 |

Six iterations here already give an optimum result. It is noteable that we get a slightly better result than the previous predictor-corrector approach. It is due to the fact that we can now look for the better time stepping, independently of any stability condition.

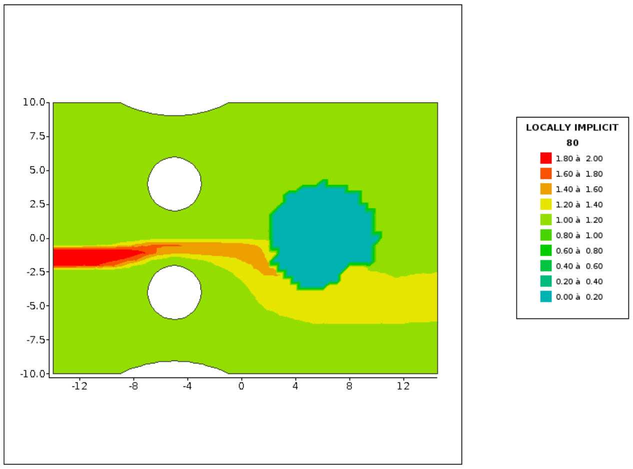

Fig. 3. The bridge pier test case with a tracer and an island.

## V. A test case with dry zones

The test case called "pildepon", a flow around bridge piers, in the portfolio of examples has been chosen, but the bottom has been modified so that a part of the domain is dry, thus forming an island. To achieve this a disc of radius 4 m has been carved out around the point of coordinates (6,0), by setting the bottom elevation at 5 m instead of 0. In Figure 3 the tracer on the island has been artificially set to 0 after the computation, to visualise the island. Otherwise the values are between 1 and 2, according to the initial and boundary conditions. The island contour is uneven due to the mesh roughness. Being a steady state, this case is not really meant for the predictor-corrector approach since the derivative in time is 0, but we show the ability of the locally implicit scheme to cope with dry areas. For this case the number of corrections is 0 and there is no sub-stepping.

## VI. Conclusion

Thanks to a local semi-implicitation depending on the local stability condition we could eventually build a distributive advection solver with a number of interesting properties:

- Mass conservation
- Monotonicity
- Low numerical diffusion
- Ability to cope with dry zones
- Unconditional stability

The height of the cone after one rotation is now more than 3 times higher than what we get with the original PSI scheme, also higher than the method of characteristics. There is no extra problem with domain decomposition parallelism. The only drawback so far is the fact that there are linear systems to solve. Given the fact that the algorithm is potentially unconditionnally stable, the number of sub-steps, which was originally given by the stability analysis, is now a tuning parameter yielding more or less numerical diffusion. The number of corrections after the predictor step is also a parameter, but it seems that no more than 5 to 6 iterations is enough to get optimum results. A problem remains: the locally implicit scheme is only a first order scheme, because so far we could not get $2^{nd}$ order without getting non linear terms in the final system.

We shall now try to apply these ideas to 3D. It should not be too difficult, as we already know that the varying depth is replaced in 3D by the varying volumes around points, so that all our theory is readily applicable.

A potential improvement would be to avoid solving too many linear systems. In the corrector steps, taking advantage of the fact that a good predictor mass is not a problem, except for the last correction, it could be possible to downgrade the accuracy, or every correction could be considered as an iteration in a Newton-Raphson process, this is left for further researches.

## References

[1] HERVOUET J.-M., PHAM C.-T.: Telemac version 5.7, release notes. Telemac-2D and Telemac-3D. 2007.

[2] HERVOUET J.-M., RAZAFINDRAKOTO E., VILLARET C.: Telemac version 5.8, release notes. Telemac-2D, Telemac-3D and Sisyphe. 2008.

[3] HERVOUET J.-M.: Telemac version 5.9, release notes. Bief, Telemac-2D, Telemac-3D and Sisyphe. 2009.

[4] HERVOUET J.-M.: Telemac version 6.0, release notes. Telemac-2D and Telemac-3D. 2010.

[5] HERVOUET J.-M., RAZAFINDRAKOTO E., VILLARET C.: Telemac version 6.1, release notes. Telemac-2D, Telemac-3D and Sisyphe. 2011.

[6] ATA R., HERVOUET J.-M.: Telemac version 6.2, release notes. Telemac-2D, Telemac-3D. 2012.

[7] HERVOUET J.-M., PAVAN S.: Telemac version 6.3, release notes. Telemac-2D, Telemac-3D. 2013.

[8] http://www.opentelemac.org/

[9] HERVOUET J.-M.: Hydrodynamics of free surface flows, modelling with the finite element method. Wiley & sons. 2007.

[10] ABGRALL R., MEZINE M.: Construction of second order accurate monotone and stable residual distribution schemes for unsteady flow problems. Journal of Computational Physics. 188:16-55. 2003.

[11] RICCHIUTO M.: An explicit residual based approach for shallow water flows. Inria Research Report n°8350, Project-Team Bacchus, September 2013.

[12] TORO E.F.: Riemann Solvers and Numerical Methods for Fluid Dynamics. Springer, 2009.