Residual distribution advection schemes in Telemac

Jean-Michel Hervouet, Sara Pavan, Mario Ricchiuto

Residual distribution advection schemes in Telemac

J.-M. Hervouet, S. Pavan, M. Ricchiuto

RESEARCH REPORT
N° 9087
August 2017
Project-Team CARDAMOM

ISSN 0249-6399   ISRN INRIA/RR--9087--FR+ENG

Residual distribution advection schemes in Telemac

J.-M. Hervouet[*], S. Pavan[*], M. Ricchiuto[†]

Project-Team CARDAMOM

Research Report n° 9087 — August 2017 — 70 pages

Abstract: This report gives an overview of the different implementations of residual distribution schemes for the advection equation in Telemac (www.opentelemac.org). The formulations considered are obtained starting from the predictor-corrector method initially proposed in (Ricchiuto et Abgrall, JCP 2010). Several iteration techniques (NERD, LIPS and ERIA) are proposed and tested in terms of accuracy and efficiency. The basic idea of NERD is a transfer of fluxes done segment by segment, surprisingly this results in an unconditional stability. LIPS is based upon a local implicitation coefficient, ERIA inspires from NERD and treats the fluxes triangle by triangle. The main advances are the low numerical diffusion coupled with an unconditional stability that allows to deal with shallow or even dry zones in a computational domain.

Key-words: advection, residual distribution schemes, Telemac

[*] LHSV and EDF

[†] Team CARDAMOM, Inria BSO

RESEARCH CENTRE
BORDEAUX – SUD-OUEST

351, Cours de la Libération
Bâtiment A 29
33405 Talence Cedex

Residual distribution advection schemes in Telemac

Résumé : Ce rapport présente toutes les variantes des schémas aux résidus distribués utilisés pour les termes de convection dans le système hydroinformatique Telemac (www.opentelemac.org). Les différentes formulations considérées se basent sur une reécriture de la méthode de prédiction-correction (Ricchiuto et Abgrall, JCP 2010). Plusieurs techniques itératives (nommées NERD, LIPS et ERIA) sont proposées et étudiées en termes de précision et efficacité. NERD exploite l'idée d'un passage de flux segment par segment et obtient ainsi une stabilité inconditionnelle, LIPS met en oeuvre un coefficient d'implicitation local, ERIA reprend l'idée de NERD mais en l'appliquant à un traitement des flux triangle par triangle. Les acquis importants sont la faible diffusion numérique et la capacité de fonctionner sur des zones à hauteur d'eau faible ou nulle.

Mots-clés : convection, schémas aux résidus distribués, Telemac

Contents

1 Introduction

Since the publication of Reference [6] in 2007, considerable improvements have been brougth to the distributive advection schemes in Telemac. In this now ten years old book the depth averaged context in 2D was not considered, leading to errors of mass-conservation. Only the N and PSI schemes were known at that time, with rather high numerical diffusion and stability criteria excluding tidal flats, a deadly drawback. The fact that adaptation to depth-averaged equations and to moving grids in 3D was the same mathematical problem was not seen at that time. Moreover, in the while, new techniques have emerged, like the idea of adding the derivative in time to the PSI limitation process, which is developed in Reference [5] and its simple explicit implementation with a predictor-corrector approach described in [16]. Concurrently, the idea of the NERD scheme was proposed in 2011 (Reference [13]), opening the way to tidal flats and dry zones. More recently, during Sara Pavan's Ph.D. at LNHE, in the years 2014-2016 (Reference [20]), several decisive progresses were made, noticeably improving the predictor-corrector approach. Coupling the idea of the NERD scheme and the predictor-corrector technique eventually lead to the ERIA scheme, which took advantage of all the recent advances, leading to what is our best solution so far, with all the numerical properties that we can dream of for our free surface applications:

- Mas conservation
- Monotonicity
- Suitable for depth averaged context or moving grids in 3D
- With very low numerical diffusion
- Compatible with massive parallelim
- Without solving linear systems

It was thus necessary to provide a new presentation of these advection schemes and to introduce the latest improvements. It is done here in a comprehensive way, starting from the basic conservative tracer equation, with extensive derivations that will allow the reader to have at hand in a single document all what is necessary for a full understanding of these new and promising schemes, including in the end hints for further improvements.

During most of the document we shall restrict ourselves to 2D domains and shallow water equations. It will be shown in the end that everything can be easily extended to free surface 3D flows and Navier-Stokes equations. In Reference [6] the presentation of distributive schemes was rather classical, with a geometrical approach of upwinding. It will be done here in a very different way, emphasizing on the fluxes between points and showing that distributive schemes are an outsider approach in between finite elements and finite volumes, which brings valuable solutions for unstructured grids.

2 Starting from finite elements

In all what follows we deal with linear functions h (depth), C (tracer), $\overrightarrow{u}$ (velocity field), that can be derived once, hence belonging to the space called H_1. We start from the tracer equation in depth-averaged context, without diffusion (treated in another fractional step) and including source terms:

$$\frac{\partial(hC)}{\partial t} + \text{div}(hC\overrightarrow{u}) = Sce\ C^{sce} \tag{1}$$

where Sce is given in m/s. It represents punctual sources of water, rain, etc. C^{sce} is the value of the tracer at the source when the source corresponds to water entering into the domain. If the source is a sink, the value of C^{sce} will be discarded, the tracer exiting the domain keeping the local value at the position of the exit. We recall that the Saint-Venant continuity equation reads:

$$\frac{\partial h}{\partial t} + \text{div}(h\overrightarrow{u}) = Sce \tag{2}$$

It is obtained with the only assumption of the impermeability of the bottom and the free surface. We start now with finite elements and do the variational formulation of the continuity equation:

$$\int_\Omega \Psi_i \frac{\partial h}{\partial t}\ \ d\Omega = -\int_\Omega \Psi_i\ \ \text{div}(h\overrightarrow{u})\ d\Omega + \int_\Omega \Psi_i Sce\ \ d\Omega \tag{3}$$

where Ω is the 2-dimensional domain and Ψ_i is the test function of point i. Unlike classical presentations of distributive schemes and in view of proving exactly the mass conservation, we do an integration by parts of the term containing $\text{div}(h\overrightarrow{u})$:

$$\int_\Omega \Psi_i \frac{\partial h}{\partial t}\ \ d\Omega = -\int_\Gamma \Psi_i\ h\ \overrightarrow{u}.\overrightarrow{n}\ d\Gamma + \int_\Omega h\overrightarrow{u}.\overrightarrow{\text{grad}}(\Psi_i)\ d\Omega + \int_\Omega \Psi_i Sce\ \ d\Omega \tag{4}$$

where Γ is the domain boundary and $\overrightarrow{n}$ the outward normal to the boundary. In a similar way we do the variational formulation of the tracer equation:

$$\int_\Omega \Psi_i \frac{\partial(hC)}{\partial t}\ \ d\Omega + \int_\Omega \Psi_i\,\text{div}(hC\overrightarrow{u})\ d\Omega = \int_\Omega \Psi_i Sce\ C^{sce}\ d\Omega \tag{5}$$

which gives after integration by parts of the divergence term:

$$\int_\Omega \Psi_i \frac{\partial(hC)}{\partial t}\ \ d\Omega = -\int_\Gamma \Psi_i\ hC\ \overrightarrow{u}.\overrightarrow{n}\ d\Gamma + \int_\Omega hC\overrightarrow{u}.\overrightarrow{\text{grad}}(\Psi_i)\ d\Omega + \int_\Omega \Psi_i Sce\ C^{sce}\ d\Omega \tag{6}$$

Leaving now pure finite elements, $\frac{\partial h}{\partial t}$ in the variational formulation is discretised in the form:

$$\frac{\partial h}{\partial t} \simeq \frac{\left(h_i^{n+1} - h_i^n\right)}{\Delta t} \tag{7}$$

and the continuity equation is written:

$$\frac{S_i\left(h_i^{n+1}-h_i^n\right)}{\Delta t} = Sce_i + \int_\Omega h\overrightarrow{u}.\overrightarrow{\text{grad}}(\Psi_i)\ d\Omega - b_i \tag{8}$$

where S_i is $\displaystyle\int_\Omega \Psi_i\ d\Omega$, the integral of test functions and $b_i = \displaystyle\int_\Gamma \Psi_i\ h\ \overrightarrow{u}.\overrightarrow{n}\ d\Gamma$, are the fluxes at the open boundaries, counted negatively if the water enters the domain. Sce_i are the fluxes at sources, counted positively if the water enters the domain. This equation can be interpreted as a water balance: $S_i h_i^n$ is the water carried by point i at the beginning of the time step, $S_i h_i^{n+1}$ is the water of point i at the end of the time step. We see in the right-hand side the fluxes arriving to point i, through sources or through the boundary. The term $-\displaystyle\int_\Omega h\overrightarrow{u}.\overrightarrow{\text{grad}}(\Psi_i)\ d\Omega$ can be interpreted as the total flux which leaves from point i to go to other neighbouring points, it is thus the opposite of the sum of fluxes arriving from all these points. If we define the fluxes between points i and j as Φ_{ij} ($\Phi_{ij} > 0$ if the flux is from i to j) we have:

$$\frac{S_i\left(h_i^{n+1}-h_i^n\right)}{\Delta t} = Sce_i - \sum_j \Phi_{ij} - b_i \tag{9}$$

A heavy use of these fluxes between points will be done later for the derivation of distributive schemes, but a problem arises: the finite element theory can easily compute $-\displaystyle\int_\Omega h\overrightarrow{u}.\overrightarrow{\text{grad}}(\Psi_i)\ d\Omega$, but how the fluxes between points can be deduced?

3 Fluxes from points and fluxes between points

On the finite element side, the terms $-\displaystyle\int_\Omega h\overrightarrow{u}.\overrightarrow{\text{grad}}(\Psi_i)\ d\Omega$ are computed at element level for the 3 points of a triangle, giving 3 contributions Φ_1, Φ_2 and Φ_3 which are the fluxes leaving points 1, 2 and 3 of the triangle. These fluxes are partial as they are restricted to the triangle and would have to be assembled after with other triangles, if we were to compute the total fluxes leaving points. We can notice that:

$$\Phi_1 + \Phi_2 + \Phi_3 = 0 \tag{10}$$

because on a triangle the sum of the test functions:

$$\Psi_1 + \Psi_2 + \Psi_3 = 1 \tag{11}$$

hence the sum of their gradients is 0. The fluxes between points are defined on Figure 1. There are simple relations between fluxes leaving points and fluxes between points, namely:

$$\Phi_1 = \Phi_{12} - \Phi_{31} \tag{12}$$

$$\Phi_2 = \Phi_{23} - \Phi_{12} \tag{13}$$

$$\Phi_3 = \Phi_{31} - \Phi_{23} \tag{14}$$

However the fluxes between points cannot be readily deduced, there is a degree of freedom. Any constant added to the fluxes between points will not change the fluxes leaving points. In other words adding a circulation of water within a triangle will not change the mass balance. However it will change a lot the mixing of tracers! We are

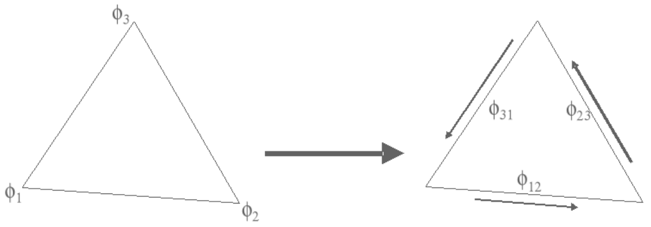

Figure 1: Fluxes leaving points (left) and fluxes between points (right)

thus facing an infinite number of solutions and decide to pick up the less diffusive. The less diffusive solution will be the one with the smallest fluxes between points. The 3 less diffusive solutions are those which cancel one flux. That is:

Solution 1:

$$\Phi_{12} = -\Phi_2 \quad \Phi_{23} = 0 \quad \Phi_{31} = \Phi_3 \tag{15}$$

Solution 2:

$$\Phi_{12} = \Phi_1 \quad \Phi_{23} = -\Phi_3 \quad \Phi_{31} = 0 \tag{16}$$

Solution 3:

$$\Phi_{12} = 0 \quad \Phi_{23} = \Phi_2 \quad \Phi_{31} = -\Phi_1 \tag{17}$$

It can be checked easily that these solutions are compatible with Equations 12 to 14, using Equation 10. Now which solution among these 3 is the less diffusive? In every we find in the fluxes between points two of the original fluxes leaving points, one with sign changed, so every time one of the initial fluxes leaving points is cancelled. The best solution, with respect to numerical diffusion, will be the one that cancels the largest flux. We thus come to the following algorithm given in Fortran style:

```
IF(ABS($\Phi_1$).GE.ABS($\Phi_2$).AND.ABS($\Phi_1$).GE.ABS($\Phi_3$)) THEN
    $\Phi_{12} = -\ \Phi_2$
    $\Phi_{23} = 0$
    $\Phi_{31} = \Phi_3$
ELSEIF(ABS($\Phi_2$).GE.ABS(f1).AND.ABS($\Phi_2$).GE.ABS($\Phi_3$)) THEN
    $\Phi_{12} =\Phi_1$
    $\Phi_{23}=$ - $\Phi_3$
    $\Phi_{31} = 0$
ELSEIF(ABS($\Phi_3$).GE.ABS($\Phi_1$).AND.ABS($\Phi_3$).GE.ABS($\Phi_2$)) THEN
    $\Phi_{12} =0$
    $\Phi_{23}= \Phi_2$
    $\Phi_{31} = -\ \Phi_1$
```

ENDIF

This algorithm has first been introduced by Leo Postma and was briefly described in a geometrical way as the "nearest projection method" in the Reference [4], though it was not given *in extenso*. These fluxes are the fluxes of the N advection scheme, that will be presented later. A more general and equivalent form is generally given in literature, and will be valid for other elements:

```
N= MIN($\Phi_1$,0.D0)+ MIN($\Phi_2$,0.D0)+ MIN($\Phi_3$,0.D0)
$\Phi_{12}$=MAX($\Phi_1$,0.D0)*MIN($\Phi_2$,0.D0)/N
$\Phi_{23}$=MAX($\Phi_2$,0.D0)*MIN($\Phi_3$,0.D0)/N
$\Phi_{31}$=MAX($\Phi_3$,0.D0)*MIN($\Phi_1$,0.D0)/N
```

Another form can be found in Telemac and is of unknown origin. It could be an unpublished discovery by Jean-Marc Janin at EDF:

```
$\Phi_{12}$=MAX(MIN($\Phi_1$,-$\Phi_2$),0.D0)- MAX(MIN($\Phi_2$,-$\Phi_1$),0.D0)
$\Phi_{23}$=MAX(MIN($\Phi_2$,-$\Phi_3$),0.D0)- MAX(MIN($\Phi_3$,-$\Phi_2$),0.D0)
$\Phi_{31}$=MAX(MIN($\Phi_3$,-$\Phi_1$),0.D0)- MAX(MIN($\Phi_1$,-$\Phi_3$),0.D0)
```

This latter form is valid for triangles only. We shall from now on refer to these fluxes as "N fluxes". For proving the equivalence of the three forms, we found no other way than testing all cases. It actually happens that there are only two cases, as explained in the next section.

When assembled the element fluxes Φ_{12}, Φ_{23} and Φ_{31} will give a set of fluxes given per segment, denoted Φ_{ij}, the assembled fluxes counted positively from i to j, which are defined for all points i and j belonging to a same segment, with the property:

$$\Phi_{ij} + \Phi_{ji} = 0 \tag{18}$$

4 One-target case and two-target case

The three possible solutions 15 to 17 are summed up in the top of Figure 2.

To avoid minus signs we can revert the arrows (bottom of same figure). It is then obvious that all the 3 solutions are alike: the two remaining fluxes are directed towards the same point. It is now very important to remember that the 3 original fluxes from points sum to 0, and that the one with largest absolute value has been cancelled. IT MEANS THAT THE TWO REMAINING HAVE THE SAME SIGN! Depending on this sign we have either the situation where two points send water to the third one (one-target case) or one point is sending water to the two others (two-target case). This fact will have large practical consequences in the derivation of distributive schemes, especially for the ERIA scheme. It is already obvious that within a triangle, upstream and downstream are clearly identified.

5 Discretising the tracer advection equation

We now take for granted that the fluxes between points are known. We choose to discretise the derivative in time of Equation 1 in the form (which implies mass-lumping):

$$\frac{S_i h_i^{n+1} C_i^{n+1} - S_i h_i^n C_i^n}{\Delta t} \tag{19}$$

The variational formulation of $Sce\ C^{sce}$ gives $\int_\Omega Sce\ C^{sce}\ \Psi_i\ d\Omega$, also simplified into $Sce_i C_i^{sce}$, if we write $Sce_i = \int_\Omega Sce\ \ \Psi_i\ d\Omega$, i.e. the discharge of the source in m$^3/s$. The

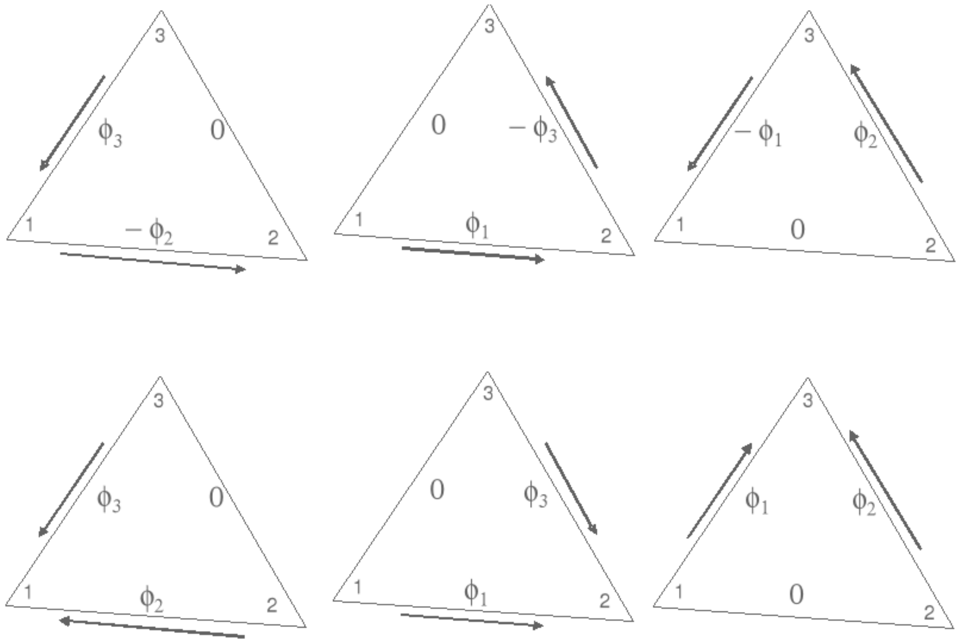

Figure 2: The 3 possible combinations of fluxes (top) rearranged to show that both remaining fluxes are directed to the same point or (if negative) leaving the same point (bottom)

variational formulation of $\mathrm{div}(hC\overrightarrow{u})$ gives after an integration by parts:

$$\int_\Omega \mathrm{div}(hC\overrightarrow{u})\ \Psi_i\ d\Omega = \int_\Gamma \Psi_i\ hC\ \overrightarrow{u}.\overrightarrow{n}\ d\Gamma - \int_\Omega hC\overrightarrow{u}.\overrightarrow{\mathrm{grad}}(\Psi_i)\ d\Omega \tag{20}$$

The first term on the right-hand side, which represents the fluxes at boundaries, is treated hereafter in the form $b_i C_i^{boundary}$, which implies also a mass-lumping, b_i being $\int_\Gamma \Psi_i\ h\ \overrightarrow{u}.\overrightarrow{n}\ d\Gamma$ the boundary flux at point i. $C_i^{boundary}$ itself will depend on the sign of b_i. The terms $-\int_\Omega hC\overrightarrow{u}.\overrightarrow{\mathrm{grad}}(\Psi_i)\ d\Omega$ are the internal fluxes of tracer, namely the fluxes that leave points when they are positive. Before assembling, the sum of these terms are 0 on every triangle (within a triangle and without source terms, mass is conserved). If we now use the N fluxes between points, namely Φ_{ij} the flux between point i and j, positive if it goes from i to j, we choose to write:

$$-\int_\Omega hC\overrightarrow{u}.\overrightarrow{\mathrm{grad}}(\Psi_i)\ d\Omega = \sum_j C_{ij}\Phi_{ij} \tag{21}$$

each flux Φ_{ij} carrying a tracer with value C_{ij}, to be defined (it will be done considering the flow direction). Because this tracer leaving point i will be received by point j as C_{ji}, we need to have:

$$C_{ji} = C_{ij} \tag{22}$$

i.e. a value linked to the segments, not to the points.

Our way of writing Equation 21 is based on the fact that our N-scheme fluxes Φ_{ij} have been designed to give:

$$-\int_\Omega h\overrightarrow{u}.\overrightarrow{\mathrm{grad}}(\Psi_i)\ d\Omega = \sum_j \Phi_{ij} \tag{23}$$

which are the fluxes of water leaving i. The tracer flux travelling from i to j is logically considered to be the water flux multiplied by C_{ij}.

Note 1: as we have $\Phi_{ij} = -\Phi_{ji}$ we still have the property:

$$-\sum_i \int_\Omega hC\overrightarrow{u}.\overrightarrow{\mathrm{grad}}(\Psi_i)\ d\Omega = \sum_i\sum_j C_{ij}\Phi_{ij} = \sum_i\sum_j \min\left(\Phi_{ij},0\right)\left(C_{ij}-C_{ji}\right) = 0 \tag{24}$$

We arrive then to the upwind finite volume advection scheme already presented in Reference [7], with a slightly re-arranged Equation 4.15:

$$S_i h_i^{n+1} C_i^{n+1} - S_i h_i^n C_i^n = \Delta t \left(Sce_i C_i^{sce} - \sum_j C_{ij}\Phi_{ij} - b_i C_i^{boundary} \right) \tag{25}$$

This discretised equation is mass-conservative, whatever the values of Φ_{ij}, provided that they obey Equation 18 because when we sum over i, we get:

$$\sum_i S_i h_i^{n+1} C_i^{n+1} - \sum_i S_i h_i^n C_i^n = \Delta t \left(\sum_i Sce_i C_i^{sce} - \sum_i b_i C_i^{boundary} \right) \tag{26}$$

which is a balance of mass taking into account the sources and the boundaries.

6 Deriving a locally semi-implicit upwind distributive scheme

We do here the basic derivation from which most of the schemes will be deduced. The general principle is that we start from the conservative equation 26 and we move to a non conservative form (that will be strictly equivalent, thus also mass-conservative). This derivation is close to what would be done in the continuum. Only after we shall decide a choice of C_{ij}. In the derivation we introduce a semi-implicit value of the tracer C at point i: $\theta_i C_i^{n+1} + (1-\theta_i)C_i^n$, where C_i^n is the initial value of C at point i and C_i^{n+1} the final value, which is yet unknown. θ_i is a local implicitation that will be chosen later. It was brought to our attention in February 2016 that it is also an idea developed by Paulien van Slingerland in her thesis in 2007 (Reference [8]). In her case it is however a coefficient θ_{ij} linked to the segments, probably due to finite volumes specific requirements. In our case upwinding is not hindered since the tracer that travels via a segment only depends on the upstream point, and the semi-implicitation does not depend on the conditions downstream.

We start from Equation 25 and add on both sides the following quantity:

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

$$-\Delta t\ Sce_i \left(\theta_i C_i^{n+1} + (1-\theta_i)C_i^n\right) + \Delta t\ b_i \left(\theta_i C_i^{n+1} + (1-\theta_i)C_i^n\right) \tag{27}$$

We get:

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

$$-S_i \left(h_i^{n} - (1-\theta_i) \frac{\Delta t}{S_i} \left(\sum_j \Phi_{ij} - \Delta t\ Sce_i + b_i \right)\right) C_i^{n} =$$

$$\Delta t \sum_j \left(\theta_i C_i^{n+1} + (1-\theta_i)C_i^n\right)\Phi_{ij} - \Delta t \sum_j C_{ij}\Phi_{ij} \tag{28}$$

$$+\Delta t\ Sce_i \left(C_i^{sce} - \left(\theta_i C_i^{n+1} + (1-\theta_i)C_i^n\right)\right) - \Delta t\ b_i \left(C_i^{boundary} - \left(\theta_i C_i^{n+1} + (1-\theta_i)C_i^n\right)\right)$$

From the discretised continuity Equation 9 we deduce that:

$$\frac{\Delta t}{S_i}\left(\sum_j \Phi_{ij} - Sce_i + b_i\right) = h_i^n - h_i^{n+1} \tag{29}$$

and defining $h_i^{n+\theta}$ as the depth at time $t^n + \theta\Delta t$:

$$h_i^{n+\theta} = (1-\theta)h_i^n + \theta h_i^{n+1} \tag{30}$$

our equation becomes :

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

$$\sum_j \left(\theta_i C_i^{n+1} + (1-\theta_i)C_i^n\right)\Phi_{ij} - \sum_j C_{ij}\Phi_{ij} \tag{31}$$

$$+\ Sce_i\left(C_i^{sce} - \left(\theta_i C_i^{n+1} + (1-\theta_i)C_i^n\right)\right) -\ b_i\left(C_i^{boundary} - \left(\theta_i C_i^{n+1} + (1-\theta_i)C_i^n\right)\right)$$

where the non conservative derivative in time appears in the left-hand side.

Now we want a semi-implicit upwind scheme: we consider that C_{ij} is equal to $\theta_i C_i^{n+1} + (1-\theta_i)C_i^n$ if Φ_{ij} is positive, i.e. from i to j, and C_{ij} is equal to $\theta_j C_j^{n+1} + (1-\theta_j)C_j^n$ if Φ_{ij} is negative.

We also consider that exiting sources ($Sce_i < 0$) or boundary fluxes ($b_i > 0$) will have a value of C_i^{sce} or $C_i^{boundary}$ equal to $\theta_i C_i^{n+1} + (1-\theta_i)C_i^n$. We get:

$$S_i h_i^{n+1-\theta_i}\frac{\left(C_i^{n+1} - C_i^n\right)}{\Delta t} =$$

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

$$-\sum_j \left(\theta_i C_j^{n+1} + (1-\theta_i)C_j^n - \theta_i C_i^{n+1} - (1-\theta_i)C_i^n\right)\min\left(\Phi_{ij}, 0\right) \tag{32}$$

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

That is, if we now put implicit terms in the left-hand side and explicit terms in the right-hand side:

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

$$+\sum_j \theta_j C_j^{n+1}\min\left(\Phi_{ij}, 0\right) =$$

$$\frac{S_i h_i^{n+1-\theta_i}}{\Delta t}C_i^n - \sum_j \left(\left(1-\theta_j\right)C_j^n - \left(1-\theta_i\right)C_i^n\right)\min\left(\Phi_{ij}, 0\right) \tag{33}$$

$$+\max\left(Sce_i, 0\right)\left(C_i^{sce} - (1-\theta_i)C_i^n\right) - \min\left(b_i, 0\right)\left(C_i^{boundary} - (1-\theta_i)C_i^n\right)$$

7 Properties of the locally semi-implicit upwind distributive scheme

Our numerical scheme is mass conservative by construction, we now need to see if it obeys the maximum principle.

Equation 33 can be put in the form of a linear system:

$$AC^{n+1} = BC^n + D \tag{34}$$

Where A and B are matrices and D is a diagonal. Namely:

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

$$A_{ij} = \theta_j \ \min(\Phi_{ij}, 0)$$

$$B_{ii} = \frac{S_i h_i^{n+1-\theta_i}}{\Delta t} - (1-\theta_i)\left(\max\left(Sce_i, 0\right) - \sum_j \min\left(\Phi_{ij}, 0\right) - \min\left(b_i, 0\right)\right) \tag{35}$$

$$B_{ij} = - \ (1-\theta_j)\min(\Phi_{ij}, 0)$$

$$D_i = - \ \min(b_i, 0) C_i^{boundary} + \ \max(Sce_i, 0) C_i^{sce}$$

It happens that matrix A is a M-matrix, which means that A^{-1} has only positive elements. This is due to the fact that all diagonal terms of A are positive and all its off-diagonal terms are negative (this is not the definition of a M-matrix but matrices like this are M-matrices). When solved, the system will give for C_i^{n+1} a combination of values of various C at time t^n and t^{n+1}, with a sum of coefficients equal to 1. This can become a proof of monotonicity if we can show that all the coefficients are positive. The fact that the combination involves values taken at t^{n+1} is not a hack in the proof and is covered by the positivity properties of M-matrices.

Actually only B_{ii} raises a problem, which leads to a Courant-Friedrich-Levy (CFL) condition. We can in fact write B_{ii} in the form:

$$B_{ii} = \frac{S_i h_i^n}{\Delta t} + (1-\theta_i)\left(\min\left(Sce_i, 0\right) - \sum_j \max\left(\Phi_{ij}, 0\right) - \max\left(b_i, 0\right)\right) \tag{36}$$

which is just using the fact that:

$$S_i h_i^n = S_i h_i^{n+1-\theta_i} - (1-\theta_i)\,\Delta t \left(Sce_i - \sum_j \Phi_{ij} - b_i\right) \tag{37}$$

and that (example for b_i, but same treatment for the other terms):

$$b_i = \min\left(b_i, 0\right) + \max\left(b_i, 0\right) \tag{38}$$

we immediately get the criterion:

$$\Delta t_{stab} < \frac{1}{1-\theta_i} \frac{S_i h_i^{start}}{\left(\sum_j \max\left(\Phi_{ij}, 0\right) + \max\left(b_i, 0\right) - \min\left(Sce_i, 0\right)\right)} \tag{39}$$

or alternatively the equivalent form:

$$\Delta t_{stab} < \frac{1}{1-\theta_i} \frac{S_i h_i^{end}}{\left(-\sum_j \min\left(\Phi_{ij}, 0\right) - \min\left(b_i, 0\right) + \max\left(Sce_i, 0\right)\right)} \tag{40}$$

Here we have replaced h_i^n by h_i^{start} and h_i^{n+1} by h_i^{end} because a stability condition may lead us to iterate within a time step and in this process the starting depth will be h_i^n only at the first iteration. Under this stability conditions all the coefficients of values of C are positive, and as their sum is 1 they are also all smaller than 1.

From Formula 39 we deduce two important facts:

- If we want a constant θ, only $\theta = 1$ will be able to give an unconditional stability.
- Explicit schemes will not work with dry zones.

Note: in distributive schemes publications the stability is ensured at element level, before assembling, which looks more restrictive. We work here on assembled fluxes.

8 Compatible fluxes for the locally implicit scheme

To find the fluxes to be taken into account for a verification of mass conservation we go back to Equation 25. When summed over all points the terms $\sum_j C_{ij}\Phi_{ij}$ cancel because we have:

$$\sum_i \sum_j C_{ij}\Phi_{ij} = \sum_i \sum_{j<i} C_{ij}\Phi_{ij} + C_{ji}\Phi_{ji} = \sum_i \sum_{j<i} C_{ij}\left(\Phi_{ij} + \Phi_{ji}\right) = 0 \tag{41}$$

With our decision on exiting sources and boundary terms we then write in fact:

$$b_i C_i^{boundary} = \min(b_i, 0) C_i^{boundary} + \max(b_i, 0)\left(\theta_i C_i^{n+1} + (1-\theta_i) C_i^n\right) \tag{42}$$

and:

$$Sce_i C_i^{sce} = \max(Sce_i, 0) C_i^{sce} + \min(Sce_i, 0)\left(\theta_i C_i^{n+1} + (1-\theta_i) C_i^n\right) \tag{43}$$

The tracer flux to be taken into account will then be:

$$\begin{aligned} &\textstyle\sum_i \left[\max(Sce_i, 0) C_i^{sce} + \min(Sce_i, 0)\left(\theta_i C_i^{n+1} + (1-\theta_i) C_i^n\right)\right] \\ &-\textstyle\sum_i \left[\min(b_i, 0) C_i^{boundary} + \max(b_i, 0)\left(\theta_i C_i^{n+1} + (1-\theta_i) C_i^n\right)\right] \end{aligned} \tag{44}$$

If $b_i < 0$ and the boundary point is not of Dirichlet type, we are in the case of an output with free velocity, then the value of $C_i^{boundary}$ might not be given by the user. In this case it will be assumed to be the previous known value, i.e. C_i^n.

We have now derived our basic semi-implicit upwind distributive scheme and the way to control mass conservation. It allows us to start presenting our series of schemes, but first we shall present test-cases that will allow us to evaluate on the spot the quality of every scheme regarding numerical diffusion, monotonicity and mass conservation.

9 Test cases

9.1 The rotating cone test-case

This first test case is done in the context of a divergence free rotating velocity field. It consists of the advection of a tracer in a solid rotation velocity field. Namely the computational domain is a square between abscissae 0 and 20.1 m and between ordinates 0 and 20.1 m. The mesh is composed of 4489 squares of side 0.3 m split into two triangles, which gives 8978 elements. The velocity field in m/s has the following two components u and v:

$$u(x, y) = 10.05 - y \tag{45}$$

$$v(x, y) = x - 10.05 \tag{46}$$

The initial tracer value is between 0 and 1, of the following Gaussian shape:

$$C^0(x, y) = e^{-\left[(x-15)^2 + (y-10.05)^2\right]//2} \tag{47}$$

The original maximum height of the cone is 1 (see Figure 3). The principle of the test is to simulate one rotation of the tracer around the center of the square. With an ideal solver, there should be no variation of the tracer after one rotation. We thus have a trivial solution which is anything but trivial for the numerical schemes. We do here one rotation in 32 iterations, with a time step of 0.196349541 s which is in fact $\pi/16$. The Courant number is about 7 if we consider sides of triangles as the mesh size. Due to their stability condition our distributive schemes will revert to sub-iterations within the requested time step. As we know the analytical solution, we can also compute the standard deviation of C^1, the result after one rotation, as:

$$Error = \sqrt{\frac{\displaystyle\sum_{i=1}^{npoin} S_i \left(C_i^1 - C_i^0\right)^2}{\displaystyle\sum_{i=1}^{npoin} S_i}} \tag{48}$$

where *npoin* is the number of points in the mesh. This parameter is however difficult to interpret, as it mixes two different errors, the amplitude error and the phase error. The results of the method of characteristcs, in strong form and in weak form (Reference [17]), are given in Figures 4 and 5. We recall here that none of these two forms is mass conservative, and that the weak form does not obey the maximum principle, they are thus discarded in studies with tracers. The cone heights after one rotation are respectively 0.6778 and 0.9936, and the standard deviations are 27.60 10^{-3} and 20.21 10^{-3}. Despite the incredible result of the weak form in terms of amplitude, its standard deviation reveals a phase error. As a matter of fact the cone is slightly shifted towards the centre of the square, due to the first-order of the Runge-Kutta method that computes the path-lines. In terms of amplitude the distributive schemes will do better than the strong form of characteristics, and in terms of phase error they will do better than the weak form.

A convergence study will be done with five levels. Level 0 is the original mesh of 4624 points and 8978 elements. At every new level of refinement the mesh size is divided by 2 and the time step by 2. All meshes are similar, only the mesh size changes. Of course we still do one rotation. We give in the table below the number of points and elements, and the time step chosen to get an unchanged CFL number. The number of points is:

$$npoin = \left(2^{level} 67 + 1\right)^2 \tag{49}$$

The number of elements, denoted *nelem*, is:

$$nelem = 2\ 67^2 4^{level} \tag{50}$$

level	number of points	number of elements	Δt
0	4624	8978	0.196349541 s
1	18225	35912	0.098174770 s
2	72361	143648	0.049087385 s
3	288369	574592	0.024543693 s
4	1151329	2298368	0.012271846 s

We give in the table below the results of strong and weak form of characteristics on our 5 levels:

level	cone height, strong	standard deviation, strong	cone height, weak	standard deviation, weak
0	0.6816	22.07 10^{-3}	0.995931	1.23 10^{-3}
1	0.8124	12.75 10^{-3}	0.996516	1.25 10^{-3}
2	0.8978	6.96 10^{-3}	0.999402	1.23 10^{-3}
3	0.9468	3.78 10^{-3}	0.999739	1.25 10^{-3}
4	0.9727	2.23 10^{-3}	0.999968	1.29 10^{-3}

In this series of 5 runs an attempt was done to keep the same accuracy in the computation of the trajectories. This accuracy depends on the average number of sub-steps

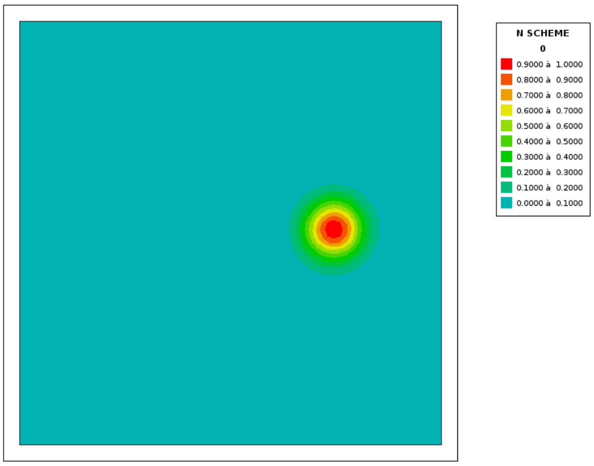

Figure 3: Initial shape of the rotating cone

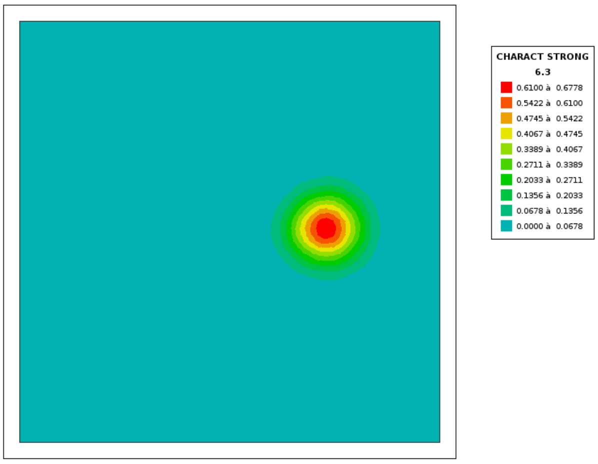

Figure 4: Characteristics in strong form. Cone after one rotation

per element of the first-order Runge-Kutta method. This number is a parameter of the method of characteristics. It is set respectively to 48, 24, 12, 6 and 3 for the levels 0 to 4. It gives interesting results. For the strong form, the standard deviation is consistant with a first order in space and time, the deviation is roughly divided by 2 when the mesh size is divided by 2. With the weak form the deviation is astonishingly constant. It is in fact the phase error. The amplitude error is so small that it does not change significantly the deviation.

9.2 Flow around bridge piers

Our second test case will be a flow around bridge piers, which is representative of a majority of quasi-steady flows in river applications. It is taken from the porfolio of examples provided in the Telemac package (see [23]), namely the test called "pildepon". The mesh was originally a regular curvilinear grid and every rectangle has been split into two triangles. The computational domain is in the range [-14,+14.5] horizontally and [-10,+10] vertically. There are 2280 points and 4304 elements. The bathymetry varies from -4 m to -1 m (Figure 6) and the depth from about 1 m to more than 4.25 m.

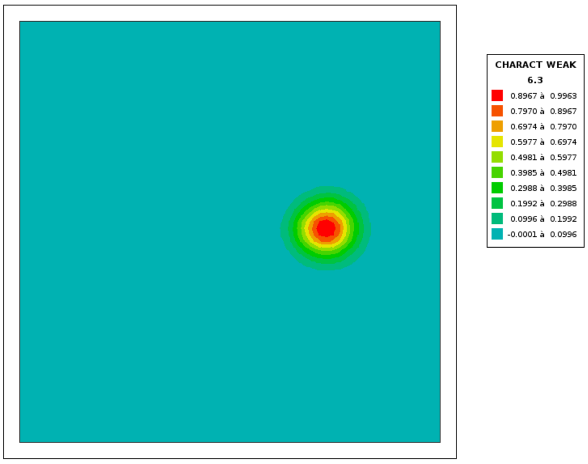

Figure 5: Characteristics in weak form. Cone after one rotation

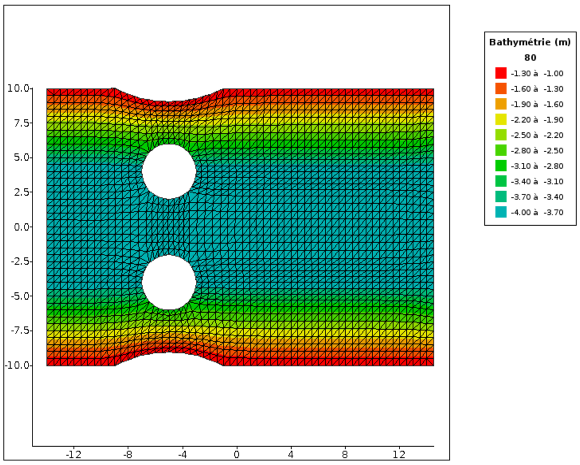

Figure 6: Mesh and bathymetry of the bridge piers test case.

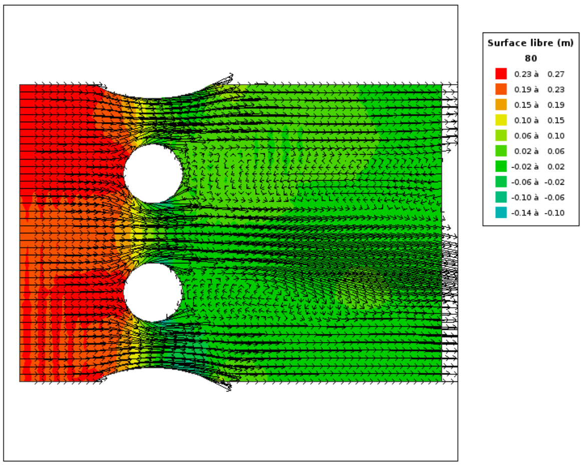

Figure 7: Velocity field and free surface after 80 s

100 time steps of 0.8 s are computed. The discharge on the left boundary is 0 at the beginning, then linearly raised to 62 m^3/s in 10 s and then left at this value during the remaining 70 s. The free surface elevation at the right boundary (exit) is 0. The flow is not steady since there are von Karman eddies behind the bridges, and sometimes (see e.g. Figure 7) the vortex shedding will trigger re-entering velocities at the exit. In this case the tracer boundary conditions are changed and it is considered that the value of the re-entering tracer is the last computed. To better track would-be errors of monotonicity, the tracer diffusion is set to 0.

With this test case, mass conservations (water and tracer) and monotonicity will be checked. A tracer with value 2 is entered upstream, whereas the initial value is 1. The advection scheme for velocities will be kept constant across all the tests, so that the velocity is left unchanged. The original test has no tidal flats nor dry zones, but when this is required the bottom will be modified so that a part of the domain is dry, thus forming an island. To achieve this a disc of radius 4 m will be carved out around the point of coordinates (6,0), by setting the bottom elevation at 5 m instead of 0.

For reference we show on Figure 8 the result given by the strong form of the method of characteristics. Monotonicity is preserved but not mass since the mass-balance reveals a relative error of 0.38 10^{-1}. The loss is in fact 67.01 (if our tracer value is considered without dimension, the unit would be m^3/s) and the total quantity of tracer at the end of the computation is 1786.812.

10 The N scheme

The N scheme is simply obtained from a fully explicit form of Equation 33:

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

$$-\sum_j\left(C_j^n-C_i^n\right)\min\left(\Phi_{ij},0\right)+\max\left(Sce_i,0\right)\left(C_i^{sce}-C_i^n\right)-\min\left(b_i,0\right)\left(C_i^{boundary}-C_i^n\right) \tag{51}$$

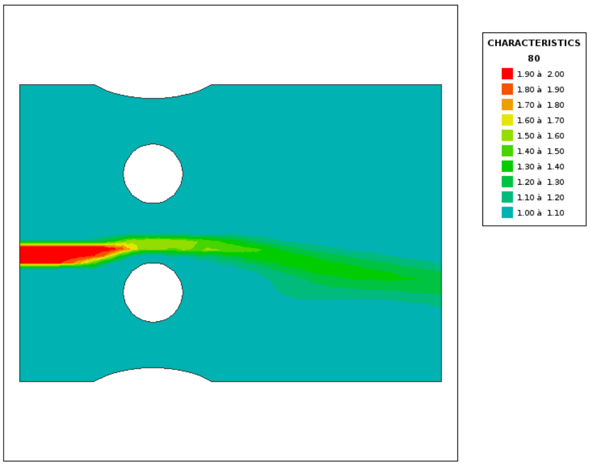

Figure 8: Bridge piers test case. Tracer advected with the method of characteristics

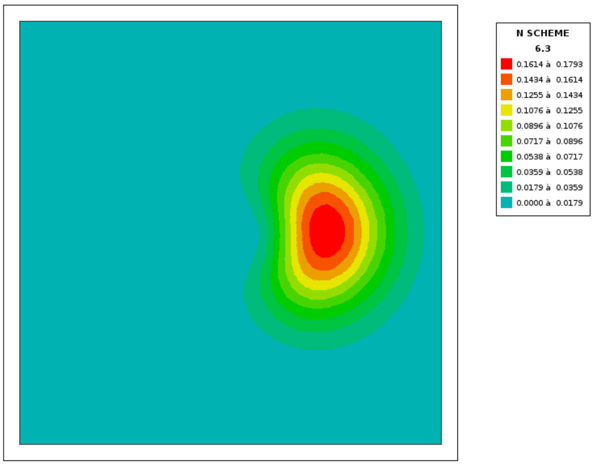

Figure 9: N scheme. The cone after one rotation.

with the stability criterion:

$$\Delta t_{stab} < \frac{S_i h_i^{start}}{\sum_j \max\left(\Phi_{ij}, 0\right) + \max\left(b_i, 0\right) - \min\left(Sce_i, 0\right)} \tag{52}$$

Figure 9 shows the cone after one rotation. The cone height is 0.1793 and the standard deviation 67.30 10^{-3}. Despite the fact that the N fluxes have been computed to minimise numerical diffusion, the N scheme is indeed very diffusive. In this case it is also due to the error in time. Note that the colour scale does not range from 0 to 1 but from 0 to 0.1793.

Figure 10 shows the tracer in the bridge piers test case. Monotonicity is obeyed and, unlike the method of characteristics, the relative error on the mass conservation of tracer is 0.19 10^{-14}, which can be considered to be the machine accuracy, allowing a few truncation errors. However there is no improvement on numerical diffusion, e.g. the yellow color of the range [1.60,1.70] does not go further downstream, whereas we would expect that, without diffusion, the value of 2 travels until the exit. Due to the stability condition, 20 sub-iterations are done at every time step, so that the real number of time steps is

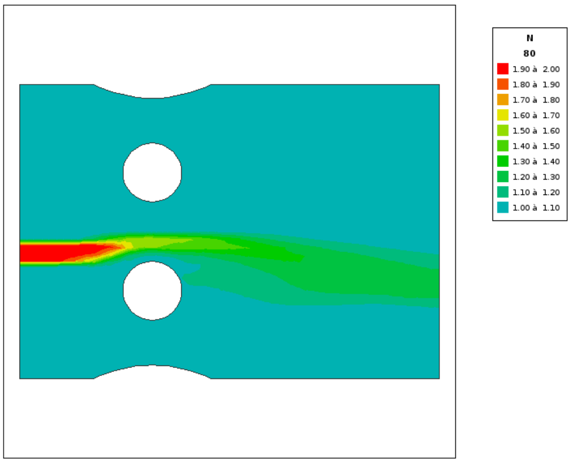

Figure 10: Bridge piers test case. Tracer advected with the N scheme.

actually 640.

For the different levels of refinement, the cone height after one rotation and the standard deviation are given in the following table:

level	N scheme, cone height	N scheme, standard deviation
0	0.1793	67.30 10^{-3}
1	0.2997	55.18 10^{-3}
2	0.4549	41.07 10^{-3}
3	0.6195	27.50 10^{-3}
4	0.7613	16.68 10^{-3}

These results show that the N scheme is hardly of order one in space. Actually we are plagued by the order in time that prevents us from finding the order in space.

11 The Positive Streamwise Invariant (PSI) scheme

Up to now we have worked on assembled fluxes Φ_{ij}. We shall now work at element level, considering fluxes limited to one element e, denoted Φ^e_{ij}. In a triangle, with the N scheme, the contribution of internal fluxes to the final right-hand side of Equation 51 is for a point i:

$$\Phi^e_i = -\sum_{j=1}^{3} \left(C^n_j - C^n_i\right) \min\left(\Phi^e_{ij}, 0\right) \tag{53}$$

We consider here the local numbering of points in the triangle, thus only numbers from 1 to 3. The total contribution of the triangle for internal fluxes will be:

$$\Phi^e = -\sum_{i=1}^{3} \sum_{j=1}^{3} \left(C^n_j - C^n_i\right) \min\left(\Phi^e_{ij}, 0\right) \tag{54}$$

With this definition we can consider that the N scheme is a distribution between 3 points of the total Φ^e with coefficients:

$$\beta_i^N = \frac{-\sum_{j=1}^{3}\left(C_j^n - C_i^n\right)\min\left(\Phi_{ij}^e, 0\right)}{\Phi^e} \tag{55}$$

The idea of the PSI scheme stems from the remark that these coefficients are not bounded and that they can be changed, provided that the sum remains 1 and that they are all positive. As a matter of fact on one hand only the total contribution Φ^e is used in a proof of mass conservation, and on the other hand the coefficient of C_i^n will not be threatened to become negative. This can be achieved by a "MinMod limiter", i.e. by choosing new coefficients:

$$\beta_i^{PSI} = \max(\min(\beta_i^N, 0), 0) \tag{56}$$

An equivalent form consists in considering reduced fluxes $\Phi_{ij}^{e\;psi}$such that:

$$\Phi_{ij}^{e\;psi} \text{ is replaced with } \beta_i^{PSI}\Phi_{ij}^e \tag{57}$$

After assembling it will give the reduced fluxes, denoted $\Phi_{ij}^{psi}\left(C^n\right)$ to clearly state that they depend on the advected function.

The MinMod limiter is valid only because we have N fluxes, in which case one of the 3 coefficients is zero. As a matter of fact only the points that receive water within the triangle have a non zero coefficient, and we have only a one-target and a 2-target case. If we have only 2 non zero coefficients, for example a set -1.1, 0 and 2.1, the MinMod limiter will give 0, 0 and 1, yielding positive coefficients with sum unchanged. When we deal with predictor-corrector schemes all the 3 local contributions may be non 0, so the MinMod limiter will not work anymore. We give here a more general algorithm that will work for any element and for any kind of contributions. For a triangle it consists in choosing:

If the total contribution Φ^e is positive:

$$\beta_i^{PSI} = \frac{\max(\Phi_i^e, 0)}{\max(\Phi_1^e, 0) + \max(\Phi_2^e, 0) + \max(\Phi_3^e, 0)} \tag{58}$$

If the total contribution Φ^e is negative:

$$\beta_i^{PSI} = \frac{\min(\Phi_i^e, 0)}{\min(\Phi_1^e, 0) + \min(\Phi_2^e, 0) + \min(\Phi_3^e, 0)} \tag{59}$$

Considering that the PSI reduction reduces the N fluxes, at least when they are taken at element level, and that the fluxes mix the tracers, it is a hint that the PSI scheme will have less numerical diffusion than the N scheme. The PSI reduction is non-linear, and it depends on the tracer C^n. It has been shown (Reference ???) that the PSI scheme is second-order in space. It is the non-linearity that allows to go beyond the limitations of the Godunov theorem (Reference [1]), stating that:

Linear numerical schemes for solving partial differential equations (PDE's), having the property of not generating new extrema (monotone scheme), can be at most first-order accurate.

From now on, the PSI reduction will be denoted with a backward arrow, and the reduced form of $-\sum_j\left(C_j^n - C_i^n\right)\min\left(\Phi_{ij}, 0\right)$ will be written $-\overleftarrow{\sum_j\left(C_j^n - C_i^n\right)\min\left(\Phi_{ij}, 0\right)}$. This notation will have to be handled with care, first because $\overleftarrow{a+b} \neq \overleftarrow{a} + \overleftarrow{b}$, and then because this reduction will not always be a reduction, as will show the stability analysis. The PSI scheme now reads:

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

$$-\overleftarrow{\sum_j \left(C_j^n - C_i^n\right)\min\left(\Phi_{ij}, 0\right)} \tag{60}$$

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

To study the stability we need to understand what is the effect of the PSI reduction. We will use the fact that the PSI reduction multiplies every contribution at element level by a coefficient in the range [0,1], but this must be looked at carefully. A flux Φ_{ij} may be the sum of 2 fluxes Φ_{ij}^{e1} and Φ_{ij}^{e2} on either side of a segment, and these two fluxes may be of different signs. In this case one is of the same sign than Φ_{ij} and of larger absolute value, and the other has the opposite sign. Let us suppose for example that in the equation $\Phi_{ij} = \Phi_{ij}^{e1} + \Phi_{ij}^{e2}$ we have $\Phi_{ij} < 0$, $\Phi_{ij}^{e1} < 0$, and $\Phi_{ij}^{e2} > 0$. The terms that will be reduced will be $\left(C_j^n - C_i^n\right)\min\left(\Phi_{ij}^{e1}, 0\right)$ on one side and 0 on the other side. The resulting sum may be larger than the original $\left(C_j^n - C_i^n\right)\min\left(\Phi_{ij}, 0\right)$. Thus the stability analysis done for the N scheme with assembled fluxes is no longer valid. To avoid this it is decided that we forbid such situations:

$$\text{Opposite fluxes, at element level, on either side of a segment, are forbidden!} \tag{61}$$

It is simple to handle such situations, at least for a mesh of triangles, as soon as the assembled value is known when doing the PSI reduction at element level. When a flux contributing to Φ_{ij} is of different sign, it is ignored. When a flux contributing to Φ_{ij} is of same sign but with larger absolute value, it is taken equal to Φ_{ij}. In this way the sum is unchanged and the local flux always has the right sign. Another possibility would consist in sharing the assembled fluxes between elements in a way that conserve the sign, for example proportionally to the triangle area. This would be however a slightly different numerical scheme. This condition has not been applied in Telemac for N and PSI schemes, yet no violation of monotonicity due to advection was ever reported. This is due to the fact that these schemes have a numerical diffusion that hides this problem. The cases with opposite fluxes are also very rare, e.g. it never happens in the rotating cone test. With less numerical diffusion, as the schemes that we shall describe now, counter-examples popped up and it was necessary to enforce Condition 61. We will still use the backward arrow notation, keeping in mind that the details of assembly must be looked at carefully.

Figure 11 shows the cone after one rotation. The cone height is 0.2137 and the standard deviation 63.30 10^{-3}. It is hardly better than the N scheme, which is disappointing. The reason is that this test case is an unsteady case, and the PSI scheme remains first-order in time. As we have kept the stability condition of the N scheme, there are also 20 sub-iterations at every time step.

Figure 12 shows the tracer in the bridge piers test case. Monotonicity is obeyed and the relative error on the mass of tracer is 0.18 10^{-14}, not significantly different from the N scheme. There is an improvement on numerical diffusion, the yellow color of the range [1.60,1.70] goes further downstream. We see here the effect of the second order in space. For the different levels of refinement the cone height after one rotation and the standard deviation are given in the following table:

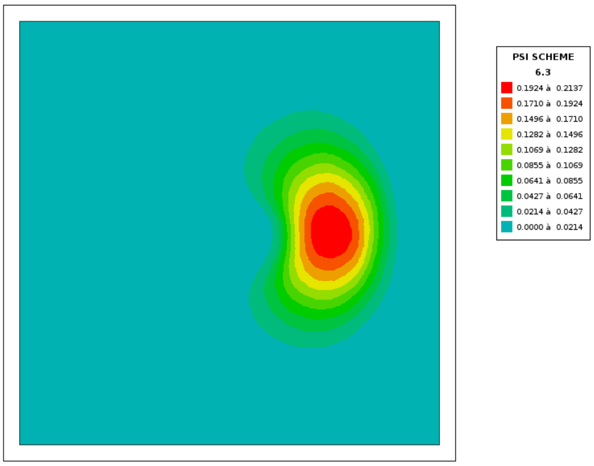

Figure 11: PSI scheme. Cone after one rotation

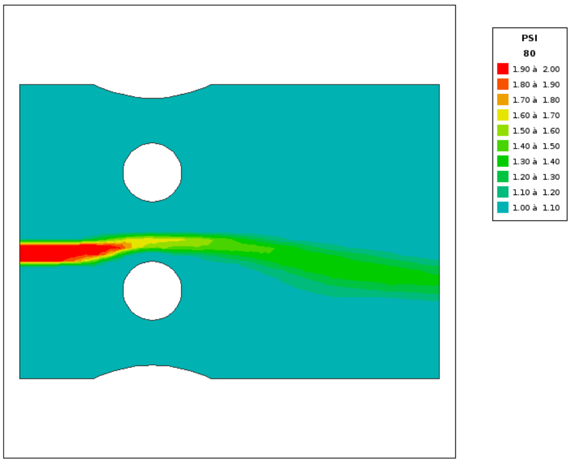

Figure 12: Bridge piers test case. Tracer advected with the PSI scheme

level	PSI scheme, cone height	PSI scheme, standard deviation
0	0.2137	$63.30\ 10^{-3}$
1	0.3357	$51.16\ 10^{-3}$
2	0.4859	$37.72\ 10^{-3}$
3	0.6417	$25.11\ 10^{-3}$
4	0.7751	$15.17\ 10^{-3}$

The improvement on the N sheme is anything but dramatic! Again it is due to the order in time.

12 Predictor-corrector distributive scheme

To avoid non-linear terms or even solving linear systems, we now follow the ideas issued by Mario Ricchiuto (Reference [16]), with a predictor-corrector scheme that approximates a semi-implicit scheme and moreover includes an important property: "upwinding" the derivative in time, where upwinding is a misleading term since it will consist only in including the derivative in time in the PSI reduction. The predictor step is, to start with, a classical explicit N scheme.

12.1 First order in time predictor-corrector scheme

The predictor step aims at finding an estimate of the final concentration C^{n+1}, which is denoted C^*. This step is just a classical explicit N scheme:

$$\begin{gathered}\frac{S_i h_i^{n+1} C_i^* - S_i h_i^{n+1} C_i^n}{\Delta t} = \\ -\sum_j \min(\Phi_{ij},0)\left(C_j^n - C_i^n\right) \qquad (62) \\ -\min(b_i,0)\left(C_i^{boundary} - C_i^n\right) + \max(Sce_i,0)\left(C_i^{sce} - C_i^n\right)\end{gathered}$$

The choice of Δt remains to be defined but the time step obeys at least Equation 52 to ensure monotonicity. The corrector step is first written, for the sake of explanation:

$$\begin{gathered}\frac{S_i h_i^{n+1} C_i^{n+1} - S_i h_i^{n+1} C_i^*}{\Delta t} = -\left(\frac{S_i h_i^{n+1} C_i^* - S_i h_i^{n+1} C_i^n}{\Delta t}\right) \\ -\sum_j \min(\Phi_{ij},0)\left(C_j^n - C_i^n\right) - \min(b_i,0)\left(C_i^{boundary} - C_i^n\right) + \max(Sce_i,0)\left(C_i^{sce} - C_i^n\right) \\ (63)\end{gathered}$$

On both sides the term $-S_i h_i^{n+1} C_i^* / \Delta t$ has been added, which is of no effect so far, but we see at the beginning of the right-hand side the opposite of the predictor left-hand side. The key idea is that this derivative in time will be added to the flux contribution and PSI-reduced together with it. Namely the corrector will be:

$$\begin{gathered}\frac{S_i h_i^{n+1} C_i^{n+1} - S_i h_i^{n+1} C_i^*}{\Delta t} = \\ -\left(\overleftarrow{\frac{S_i h_i^{n+1} C_i^* - S_i h_i^{n+1} C_i^n}{\Delta t} + \sum_j \min(\Phi_{ij},0)\left(C_j^n - C_i^n\right)}\right) \qquad (64) \\ -\min(b_i,0)\left(C_i^{boundary} - C_i^n\right) + \max(Sce_i,0)\left(C_i^{sce} - C_i^n\right)\end{gathered}$$

When summing this equation over all points i in the mesh, as the PSI reduction does not change the total contribution at element level, we can remove the backward arrow, and then remove the term $-S_i h_i^{n+1} C_i^* / \Delta t$ on both sides, and we get the same proof of mass conservation as the classical N scheme. We thus just need to examine the stability of our scheme.

12.2 Stability of the first-order predictor-corrector scheme

To study the stability we again need to understand precisely what is the effect of the PSI reduction. If we take the term $-S_i h_i^{n+1}\left(C_i^* - C_i^n\right)/\Delta t$ under the backward arrow, in triangles containing i it will appear as $-S_T h_i^{n+1}\left(C_i^* - C_i^n\right)/3\Delta t$ where S_T is the area of the triangle. As a matter of fact, $S_T/3$ is the integral at element level of the test function of point i. Then $S_T h_i^{n+1}\left(C_i^* - C_i^n\right)/3\Delta t$ will be multiplied by a coefficient in the range [0,1], due to the PSI reduction. When these local coefficients will be assembled, their sum will not be greater in absolute value than $S_i h_i^{n+1}\left(C_i^* - C_i^n\right)/\Delta t$, which would be their sum without reduction. The global effect of the PSI reduction on the term $(S_ih_i^{n+1}C_i^* - S_ih_i^{n+1}C_i^n)/\Delta t$ is thus to multiply it by a coefficient denoted f_i, in the range [0,1]. As we have already seen the situation is more complicated for the term $\sum_j \min(\Phi_{ij},0)\left(C_j^n - C_i^n\right)$. However, under the condition 61, Φ_{ij} is the sum of at most 2 fluxes Φ_{ij}^{e1} and Φ_{ij}^{e2} of same sign. If $\Phi_{ij} < 0$, only case that gives a contribution, assembling the locally reduced values will give a term:

$$a\min(\Phi_{ij}^{e1},0)\left(C_j^n - C_i^n\right) + b\min(\Phi_{ij}^{e2},0)\left(C_j^n - C_i^n\right)$$

where a is the reduction factor of point i in element $e1$ and b the reduction factor of point i in element $e2$. This term can be written $\mu_{ij}\min(\Phi_{ij},0)\left(C_j^n - C_i^n\right)$, with μ_{ij} in the range [0,1], and even between a and b, as it is:

$$\mu_{ij} = \frac{a\Phi_{ij}^{e1} + b\Phi_{ij}^{e2}}{\Phi_{ij}^{e1} + \Phi_{ij}^{e2}}$$

We thus have to prove the monotonicity of the following scheme:

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

$$-f_i(\frac{S_ih_i^{n+1}C_i^* - S_ih_i^{n+1}C_i^n}{\Delta t}) - \mu_{ij}\sum_j \min(\Phi_{ij},0)\left(C_j^n - C_i^n\right) \tag{65}$$

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

where f_i and μ_{ij} are random numbers in the range [0,1]. Actually these two numbers are not totally independent, e.g. one cannot be 0 if the other is 1, but this will not used in the proof. The worrying fact is that they are different because μ_{ij} stems from the reduction on 2 elements at most, while f_i stems from the reduction on all the elements containing point i. If we look at the coefficients of the different values of C that will give C_i^{n+1} we see that the sum is $S_ih_i^{n+1}/\Delta t$, i.e. the coefficient of C_i^{n+1}, so that the monotonicity will be proven if all the coefficients are positive. Actually there is only a risk with the coefficient of C_i^n. For example the coefficient of C_i^* is $S_ih_i^{n+1}(1-f_i)/\Delta t$, which is always positive or 0. The coefficient of C_i^n is:

$$f_i\frac{S_ih_i^{n+1}}{\Delta t} + \mu_{ij}\sum_j \min(\Phi_{ij},0) + \min(b_i,0) - \max(Sce_i,0)$$

There is actually no hope of finding a value of Δt that will give the positivity of this number, since f_i may be 0. Is it a dead end? No, the positivity of coefficients is sufficient but not necessary. If we come back to the maximum principle and if we have local extrema $C_i^{\min}$ and $C_i^{\max}$ that must not be trespassed on, we just need to have C_i^{n+1} in the range $[C_i^{\min}$,$C_i^{\max}]$. Given our formula these extrema will be:

$$C_i^{\min} = \min(C_i^*, C_i^n, \text{all } C_j^n, C_i^{boundary}, C_i^{sce}) \tag{66}$$

$$C_i^{\max} = \max(C_i^*, C_i^n, \text{all } C_j^n, C_i^{boundary}, C_i^{sce}) \tag{67}$$

j including all other points in elements containing i. Actually the values of C appearing in the min and max functions are taken in the right-handside of Formula 64. In this formula the C_j^n may be random numbers independent of C_i^n, but C_i^* stems from the predictor. If Δt tends to 0 it will tend to C_i^n, so with a stability condition that remains to be defined for the predictor, C_i^* will not be too far from C_i^n. This can contribute to the monotonicity. Namely if we can show that we have:

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

$$\left(\frac{S_i h_i^{n+1}}{\Delta t} + \sum_j \mu_{ij} \min(\Phi_{ij},0) + \min(b_i,0) - \max(Sce_i,0) \right) C_i^{average} \qquad (68)$$

and that we have the two conditions:

- $C_i^{average}$ is in the range $[C_i^{\min}$,$C_i^{\max}]$,
- $\frac{S_i h_i^{n+1}}{\Delta t} + \sum_j \mu_{ij} \min(\Phi_{ij},0) + \min(b_i,0) - \max(Sce_i,0)$ is positive.

we shall have a new proof of monotonicity. Because $\mu_{ij} \leq 1$ the second condition is ensured as soon as the time step has been chosen for the stability of the N or PSI scheme, which is the minimum required for the predictor. We thus come to two conditions on C_i^* :

$$\frac{(1-f_i)\, S_i h_i^{n+1}}{\Delta t} C_i^* + \left(\frac{f_i S_i h_i^{n+1}}{\Delta t} + \mu_{ij} \sum_j \min(\Phi_{ij},0) + \min(b_i,0) - \max(Sce_i,0) \right) C_i^n \leq$$

$$\left(\frac{S_i h_i^{n+1}}{\Delta t} + \sum_j \mu_{ij} \min(\Phi_{ij},0) + \min(b_i,0) - \max(Sce_i,0) \right) C_i^{\max} \qquad (69)$$

and:

$$\frac{(1-f_i)\, S_i h_i^{n+1}}{\Delta t} C_i^* + \left(\frac{f_i S_i h_i^{n+1}}{\Delta t} + \mu_{ij} \sum_j \min(\Phi_{ij},0) + \min(b_i,0) - \max(Sce_i,0) \right) C_i^n \geq$$

$$\left(\frac{S_i h_i^{n+1}}{\Delta t} + \sum_j \mu_{ij} \min(\Phi_{ij},0) + \min(b_i,0) - \max(Sce_i,0) \right) C_i^{\min} \qquad (70)$$

Let us look at the first condition. It is naturally true if $f_i = 1$. The risk of a result larger than $C_i^{\max}$ exists only if C_i^* is larger than C_i^n (otherwise the left-hand side decreases as soon as f_i decreases). The most risky situation happens with $f_i = 0$ and $\mu_{ij} = 1$. Our condition then becomes:

$$C_i^* \leq C_i^{\max} + \frac{\Delta t}{S_i h_i^{n+1}} \left(\sum_j \min(\Phi_{ij},0) + \min(b_i,0) - \max(Sce_i,0) \right) \left(C_i^{\max} - C_i^n\right) \quad (71)$$

and we find in the same way:

$$C_i^* \geq C_i^{\min} + \frac{\Delta t}{S_i h_i^{n+1}} \left(\sum_j \min(\Phi_{ij},0) + \min(b_i,0) - \max(Sce_i,0) \right) \left(C_i^{\min} - C_i^n\right) \quad (72)$$

Let us suppose now that the time step of the predictor is chosen in the form:

$$\Delta t_{stab} < \frac{1}{k} \frac{S_i h_i^{n+1}}{\left(-\sum_j \min\left(\Phi_{ij}, 0\right) - \min\left(b_i, 0\right) + \max\left(Sce_i, 0\right)\right)} \tag{73}$$

which is equivalent (see Equations 39 and 40) to:

$$\Delta t_{stab} < \frac{1}{k} \frac{S_i h_i^{n}}{\left(\sum_j \max\left(\Phi_{ij}, 0\right) + \max\left(b_i, 0\right) - \min\left(Sce_i, 0\right)\right)} \tag{74}$$

This leads us to the following stability condition of the corrector:

$$C_i^{\min} + \frac{1}{k}\left(C_i^n - C_i^{\min}\right) \leq C_i^* \leq C_i^{\max} + \frac{1}{k}\left(C_i^n - C_i^{\max}\right) \tag{75}$$

With $k = 1$ it is clear that the corrector would be equal to the predictor since it imposes $C_i^* = C_i^n$. We thus need a reduced time step in the predictor, compared to the N scheme. We shall now look for a value of k in the predictor stability condition that would imply also the stability of the corrector with Condition 75. Under Condition 73, we look at the predictor value C_i^* written in the form:

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

$$-\Delta t \sum_j \min(\Phi_{ij}(C^n), 0) C_j^n - \Delta t \min(b_i, 0) C_i^{boundary} + \Delta t \max(Sce_i, 0) C_i^{sce} \tag{76}$$

If we replace C_j^n, $C_i^{boundary}$ and C_i^{sce} with $C_i^{\max}$ in the right-hand side it will give a maximum value of $S_i h_i^{n+1} C_i^*$. Replacing them by $C_i^{\min}$ will give a minimum value. We have thus:

$$S_i h_i^{n+1} C_i^* \leq S_i h_i^{n+1} C_i^n$$

$$+ \left(-\Delta t \sum_j \min(\Phi_{ij}(C^n), 0) C_j^n - \Delta t \min(b_i, 0) C_i^{boundary} + \Delta t \max(Sce_i, 0) C_i^{sce}\right) \left(C_i^{\max} - C_i^n\right) \tag{77}$$

$$S_i h_i^{n+1} C_i^* \geq S_i h_i^{n+1} C_i^n$$

$$+ \left(-\Delta t \sum_j \min(\Phi_{ij}(C^n), 0) C_j^n - \Delta t \min(b_i, 0) C_i^{boundary} + \Delta t \max(Sce_i, 0) C_i^{sce}\right) \left(C_i^{\min} - C_i^n\right) \tag{78}$$

If these two inequalities are true, they will be also true with original N fluxes which are larger, and Condition 73 written differently states that:

$$-\Delta t \sum_j \min\left(\Phi_{ij}, 0\right) - \Delta t \min\left(b_i, 0\right) + \Delta t \max\left(Sce_i, 0\right) < \frac{1}{k} S_i h_i^{n+1} \tag{79}$$

We can deduce eventually that:

$$C_i^n + \frac{1}{k}\left(C_i^{\min} - C_i^n\right) \leq C_i^* \leq C_i^n + \frac{1}{k}\left(C_i^{\max} - C_i^n\right) \tag{80}$$

or:

$$\left(1 - \frac{1}{k}\right) C_i^n + \frac{1}{k} C_i^{\min} \leq C_i^* \leq \left(1 - \frac{1}{k}\right) C_i^n + \frac{1}{k} C_i^{\max} \tag{81}$$

This is the property obtained with the predictor, to be compared with Property 75 requested for the corrector, which can be written:

$$\left(1 - \frac{1}{k}\right) C_i^{\min} + \frac{1}{k} C_i^n \leq C_i^* \leq \left(1 - \frac{1}{k}\right) C_i^{\max} + \frac{1}{k} C_i^n \tag{82}$$

These two conditions coincide if $1 - \frac{1}{k} = \frac{1}{k}$, i.e. if $k = 2$. We have thus found that:

$$\text{The first-order explicit predictor-corrector is stable with half the time-step of N and PSI schemes} \tag{83}$$

Results: with the rotating cone test, we get a cone height of 0.47 after one rotation. This is a tremendous progress, the result of the PSI scheme is more than doubled.

Note: the stability condition is slightly different if we assume that C_i^* obeys the predictor equation, as a matter of fact the proof can then be done by adding the predictor and the corrector and it gives:

$$\Delta t_{stab} < \frac{S_i h_i^{n+1}}{\left(\sum_j \max\left(\Phi_{ij}, 0\right) - \sum_j \min\left(\Phi_{ij}, 0\right) - 2\min\left(b_i, 0\right) + 2\max\left(Sce_i, 0\right)\right)} \tag{84}$$

which is also:

$$\Delta t_{stab} < \frac{S_i h_i^{n+1}}{\left(\sum_j abs\left(\Phi_{ij}\right) - 2\min\left(b_i, 0\right) + 2\max\left(Sce_i, 0\right)\right)} \tag{85}$$

This less restrictive form has not been retained, though it gives slightly better results, in order to enable multiple corrections.

Figure 13 shows the result obtained with the rotating cone and parameter $k = 2$. The cone height after one rotation is now 0.4795, a tremendous progress, and the standard deviation is 34.68 10^{-3}.

Figure 14 shows the tracer in the bridge piers test case. Monotonicity is obeyed and the relative error on the mass of tracer is 0.74 10^{-12}. Compared to the PSI scheme, there is no significant improvement on numerical diffusion. The explanation is that we have between the piers a quasi-steady flow, thus including the derivative in time in the corrector step has little effect and we fall back on the PSI scheme.

For the different levels of refinement the cone height after one rotation and the standard deviation are given in the following table (N-PC stands for Predictor-Corrector with N scheme as predictor):

level	N-PC order 1, cone height	N-PC order 1, standard deviation
0	0.4795	34.68 10^{-3}
1	0.7325	13.87 10^{-3}
2	0.8859	7.03 10^{-3}
3	0.9545	4.12 10^{-3}
4	0.9830	1.95 10^{-3}

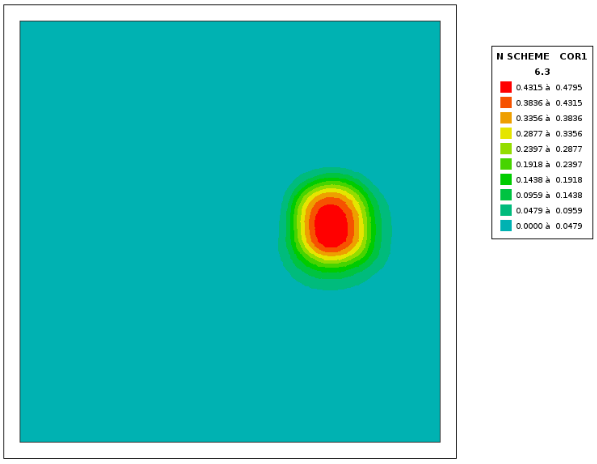

Figure 13: First-order predictor-corrector scheme. Cone after one rotation.

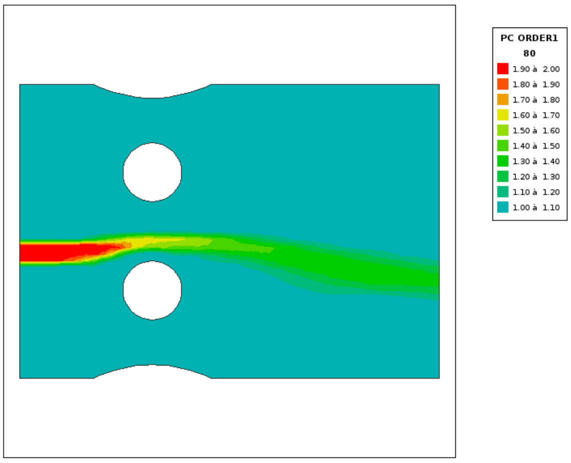

Figure 14: Bridge piers test case. Tracer advected with the first-order predictor-corrector scheme, with the PSI scheme as predictor.

Now the error is divided by around 2 at every new level. This is consistant with a scheme of order 1 in time and order 1 in space.

12.3 Predictor-corrector with PSI scheme

Actually none of the proofs given so far is spoiled if we use the PSI scheme instead of the N scheme in the predictor step. In this case the results are even better, the cone height after one rotation is 0.4986, and the standard deviation becomes 32.69 10^{-3}. Results for the four levels are reported in the table below:

level	cone height, PC order 1	standard deviation
0	0.4986	32.69 10^{-3}
1	0.7498	12.70 10^{-3}
2	0.8969	6.68 10^{-3}
3	0.9602	3.98 10^{-3}
4	0.9857	1.89 10^{-3}

However we could think of using also the PSI scheme in the corrector, to be added with the derivative in time, before the PSI reduction. It could be argued that it would do a double reduction, but the results are consistently slightly better for the cone height, as shown in the table below for 4 levels:

level	cone height, PC order 1	standard deviation
0	0.5079	$32.01\ 10^{-3}$
1	0.7547	$12.50\ 10^{-3}$
2	0.8995	$6.81\ 10^{-3}$
3	0.9614	$4.03\ 10^{-3}$

However the standard deviation is slightly worse for levels 2 and 3, and also with some of the improvements described hereafter, so for simplicity and code optimisation it appeared preferable to keep the PSI scheme only at the predictor step.

12.4 Predictor-corrector with multiple corrections

The question is now: can we do more corrections and use the result of the corrector as a new and more accurate predictor? A key remark is that any function obeying Equation 82 would be suitable to give a stable corrector. As a matter of fact mass conservation does not raise any problem and is guaranteed even if C_i^* is not solution of the predictor. In Equation 63 the same mass depending on C_i^* is withdrawn from both sides, thus C_i^* does not interfere with mass conservation.

We can thus imagine that the result of a previous correction is re-used as a new predictor after being clipped to obey Condition 82. Clipping will not endanger monotonicity nor mass conservation, but only the quality of the result. We see the results obtained wih the rotating cone test in the following table, with the PSI scheme as initial predictor, for the level 0 mesh.

number of corrections	cone height after one rotation	standard deviation
0	0.2137 (=PSI scheme)	$63.30\ 10^{-3}$
1	0.4986	$32.69\ 10^{-3}$
2	0.6562	$20.64\ 10^{-3}$
3	0.7020	$18.72\ 10^{-3}$
4	0.7177	$18.15\ 10^{-3}$
5	0.7249	$17.88\ 10^{-3}$
6	0.7288	$17.72\ 10^{-3}$
7	0.7308	$17.64\ 10^{-3}$
8	0.7318	$17.51\ 10^{-3}$
9	0.7323	$17.50\ 10^{-3}$
10	0.7323	$17.45\ 10^{-3}$

Table 1: effect of the number of corrections with the first-order predictor-corrector scheme

Even a second correction triggers a dramatic improvement. Figure 15 shows the cone after one rotation, in the case with five corrections, with a height of 0.7249. The standard deviation is $17.88\ 10^{-3}$. In terms of error we are thus now better than the method of characteristics.

It seems that we have a convergence after very few iterations of the corrector. Now a new question arises: is there a stability condition that could be applied to the predictor step and would allow an arbitrary number of iterations without limiting C_i^*? Actually it can be shown that n corrections will require $k = n + 1$ in Condition 73. This is very demanding, and our approach consisting in forcing Condition 82 is better. Tests of the

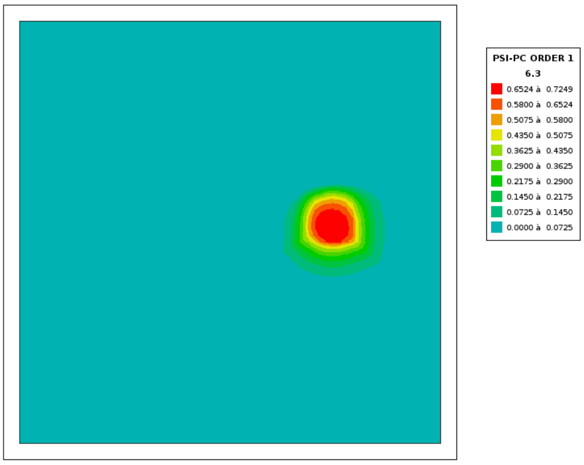

Figure 15: First-order predictor-corrector scheme with 5 corrections. Cone after one rotation.

rotating cone with $k = 3$ do not show any improvement, either on the cone height or on the error.

For the different levels of refinement the cone height after one rotation and the standard deviation are given in the following table, with 5 corrections:

level	cone height, PSI PC order 1 with 5 corrections	standard deviaion
0	0.7249	$17.88\ 10^{-3}$
1	0.9034	$10.53\ 10^{-3}$
2	0.9610	$5.83\ 10^{-3}$
3	0.9814	$3.10\ 10^{-3}$
4	0.9857	$1.89\ 10^{-3}$

Let us now explore the possibility of a second-order in time predictor-corrector.

12.5 Dividing the N time-step by a factor less than 2

So far choosing $k = 2$ seemed a reasonable choice and we have discarded values larger than 2. What about values of k smaller than 2 ? In this case the limitation of the predictor value must be done even in the first correction. Of course $k = 1$ would force the predictor $C^*$to be equal to C^n because of the limitation in Inequality 82. However there could be an optimum between 1 and 2. As a matter of fact, with a predictor-corrector starting with the PSI scheme, with 5 corrections, a minimum of standard deviation was found for $k = 1.55$ in the rotating cone test. Compared to $k = 2$, the cone height changes from 0.7249 to 0.7482 and the standard deviation from $17.88\ 10^{-3}$ to $17.40\ 10^{-3}$. This is not a big difference, but it may give a smaller computer time.

13 Second-order in time predictor-corrector scheme

The predictor is still the same. The general idea for the corrector is to keep an explicit scheme but to tend to semi-implicit tracer fluxes, with a constant θ that will be $1/2$. It is not so simple because of mass-conservation issues! We need to start again nearly from scratch with Equation 25 written:

$$S_i h_i^{n+1} C_i^{n+1} = S_i h_i^n C_i^n + \Delta t \left(Sce_i C_i^{sce} - \sum_j C_{ij} \Phi_{ij} - b_i C_i^{boundary} \right) \tag{86}$$

Then we add $-S_i h_i^{n+1} C_i^*$ on both sides and write C_{ij} in a semi-implicit form $\theta C_{ij}^* + (1-\theta) C_{ij}^n$, it yields:

$$\begin{gathered} S_i h_i^{n+1} C_i^{n+1} - S_i h_i^{n+1} C_i^* = \\ -\left(S_i h_i^{n+1} C_i^* - S_i h_i^n C_i^n\right) + \Delta t \left(Sce_i C_i^{sce} - \textstyle\sum_j \left(\theta C_{ij}^* + (1-\theta) C_{ij}^n\right) \Phi_{ij} - b_i C_i^{boundary} \right) \end{gathered} \tag{87}$$

In the right-hand side, we use the fact that:

$$S_i h_i^n = S_i h_i^{n+1-\theta} - (1-\theta) \Delta t \left(Sce_i - \sum_j \Phi_{ij} - b_i \right) \tag{88}$$

and:

$$S_i h_i^{n+1} = S_i h_i^{n+1-\theta} + \theta \Delta t \left(Sce_i - \sum_j \Phi_{ij} - b_i \right) \tag{89}$$

to get:

$$\begin{gathered} S_i h_i^{n+1} C_i^{n+1} - S_i h_i^{n+1} C_i^* = -\left(S_i h_i^{n+1-\theta} C_i^* - S_i h_i^{n+1-\theta} C_i^n\right) \\ -\Delta t \left(\sum_j \left(\theta C_{ij}^* + (1-\theta) C_{ij}^n - (1-\theta) C_i^n - \theta C_i^*\right) \Phi_{ij} \right) \\ +\Delta t \left(Sce_i \left(C_i^{sce} - \theta C_i^* - (1-\theta) C_i^n\right) - b_i \left(C_i^{boundary} - \theta C_i^* - (1-\theta) C_i^n \right) \right) \end{gathered} \tag{90}$$

We must then choose an upwind form of C_{ij}^* and C_{ij}^n, and decide that on exits and sinks the tracer value is $\theta C^* + (1-\theta) C^n$, which eventually yields:

$$\begin{gathered} S_i h_i^{n+1} C_i^{n+1} - S_i h_i^{n+1} C_i^* = -\left(S_i h_i^{n+1-\theta} C_i^* - S_i h_i^{n+1-\theta} C_i^n\right) \\ -\Delta t \sum_j \min(\Phi_{ij}, 0) \left(\theta \left(C_j^* - C_i^*\right) + (1-\theta) \left(C_j^n - C_i^n\right)\right) \\ +\Delta t \left(\max(Sce_i, 0) \left(C_i^{sce} - \theta C_i^* - (1-\theta) C_i^n\right) - \min(b_i, 0) \left(C_i^{boundary} - \theta C_i^* - (1-\theta) C_i^n \right) \right) \end{gathered} \tag{91}$$

The last step consists in doing a PSI reduction, including the derivative in time, in the right-hand side:

$$\begin{gathered} S_i h_i^{n+1} C_i^{n+1} - S_i h_i^{n+1} C_i^* = \\ \overleftarrow{-\left(S_i h_i^{n+1-\theta} C_i^* - S_i h_i^{n+1-\theta} C_i^n\right) + \Delta t \sum_j \min(\Phi_{ij}, 0) \left(\theta \left(C_j^* - C_i^*\right) + (1-\theta) \left(C_j^n - C_i^n\right)\right)} \end{gathered} \tag{92}$$

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

It thus appears that for mass-conservation reasons and to counter-act the choice of semi-implicit fluxes:

$$\text{The second-order explicit predictor-corrector needs a different derivative in time in the right-hand side} \tag{93}$$

In the mass balance the computation of fluxes at boundaries must be semi-implicit.

13.1 Monotonicity

Now the monotonicity proof. It is rather long and cumbersome and has been put in Annex 1. We just summarize here the important results. We work under the assumption that the predictor has been done with the condition 73 or 74, thus with a parameter k to be chosen. The semi-implicitation θ is another parameter. The extrema to be considered in the maximum principle are now different, since new values of C appear in Formula 92. We must now define:

$$\widetilde{C}_i^{\min}=\min(C_i^*,\ \text{all }C_j^*,\,C_i^n,\text{all }C_j^n,C_i^{boundary},C_i^{sce}) \tag{94}$$

$$\widetilde{C}_i^{\max}=\max(C_i^*,\ \text{all }C_j^*,\,C_i^n,\text{all }C_j^n,C_i^{boundary},C_i^{sce}) \tag{95}$$

According to Annex 1, there are now actually two conditions that must be satisfied by the predictor:

$$C_i^n+\frac{k-1}{2\theta}\left(C_i^n-\widetilde{C}_i^{\max}\right)\le C_i^*\le C_i^n+\frac{k-1}{2\theta}\left(C_i^n-\widetilde{C}_i^{\min}\right) \tag{96}$$

An important finding is that this condition cannot always be satisfied by the predictor, whatever the choice of k and θ, without limiting C_i^*.

The second condition is:

$$C_i^n+\left(\widetilde{C}_i^{\min}-C_i^n\right)\frac{k-1}{k-\theta}\le C_i^*\le C_i^n\ +\left(\widetilde{C}_i^{\max}-C_i^n\right)\frac{k-1}{k-\theta} \tag{97}$$

and is ensured by the PSI predictor for every value of $\theta\ge 0$ as soon as $k\ge 1$. We see that this condition vanishes if $\theta=1$.

$$\text{The second-order explicit predictor-corrector needs two different limitations} \tag{98}$$

If we make the reasonable choice $k=2$ and $\theta=\frac{1}{2}$. It gives:
for all the corrections:

$$2C_i^n-\widetilde{C}_i^{\max}\le C_i^*\le 2C_i^n-\widetilde{C}_i^{\min} \tag{99}$$

from the second correction on:

$$\frac{2\widetilde{C}_i^{\min}}{3}+\frac{C_i^n}{3}\le C_i^*\le\frac{2\widetilde{C}_i^{\max}}{3}+\frac{C_i^n}{3} \tag{100}$$

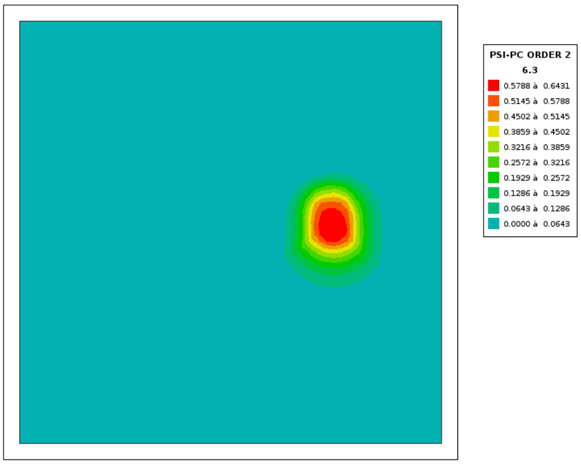

Figure 16: Second-order predictor-corrector scheme with 5 corrections. Cone after one rotation.

There is no hierarchy between these conditions, and depending on C_i^n one or the other may be the stricter one, or one may be a constraint for the minimum and the other for the maximum.

The table below gives the cone height after one rotation, and the standard deviation. The results are unfortunately not better and the standard deviation is larger than with the first order. It is only when refining that we see a clear effect of the higher order in time.

number of corrections	cone height after one rotation	standard deviation
0	0.2137 (PSI scheme)	$63.30\ 10^{-3}$
1	0.4890	$35.35\ 10^{-3}$
2	0.6059	$25.95\ 10^{-3}$
3	0.6326	$24.60\ 10^{-3}$
4	0.6414	$24.30\ 10^{-3}$
5	0.6442	$24.23\ 10^{-3}$
6	0.6449	$24.21\ 10^{-3}$
7	0.6449	$24.22\ 10^{-3}$
8	0.6450	$24.24\ 10^{-3}$
9	0.6454	$24.23\ 10^{-3}$
10	0.6452	$24.27\ 10^{-3}$

Table 3: effect of the number of corrections with the second-order predictor-corrector scheme

Figure 16 shows the cone after one rotation, with 5 corrections.

For the different levels of refinement the cone height after one rotation and the standard deviation are given in the following table, with 5 corrections:

level	cone height, PSI PC order 2 with 5 corrections	standard deviation
0	0.6442	$24.23\ 10^{-3}$
1	0.8429	$8.89\ 10^{-3}$
2	0.9315	$2.95\ 10^{-3}$
3	0.9700	$1.05\ 10^{-3}$
4	0.9873	$0.4219\ 10^{-3}$

We now seem to have more than a division by 2 of error at every new refinement.

14 LIPS: a Locally semi-Implicit Predictor-corrector Scheme

We now want to add a predictor-corrector approach to our locally semi-implicit scheme, and set up, as for the other previous schemes, a procedure with corrections. The goal is both to cope with dry zones and to keep the very low diffusion of explicit schemes. We thus consider that Scheme 33 is our new predictor, solving:

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

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

$$-\Delta t \min(b_i, 0) \left(C_i^{boundary} - (\theta_i C_i^* + (1-\theta_i) C_i^n) \right) \tag{101}$$

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

Still with the parameter k, the stability of a predictor-corrector procedure was so far chosen to be:

$$\Delta t(i) < \frac{1}{k(1-\theta_i)} \frac{S_i h_i^{start}}{\sum_j \max(\Phi_{ij}, 0) + \max(b_i, 0) - \min(Sce_i, 0)} \tag{102}$$

Now the following question: what local θ_i can we choose to have stability whatever the time step given by the user? We add a parameter to the process, with a number of sub-iterations n, a user parameter allowing to tune the CFL number. Every point has a potential explicit time-step suitable for the PSI scheme, equal to:

$$\Delta t_{stab}(i) = \frac{S_i h_i^{start}}{\sum_j \max(\Phi_{ij}, 0) + \max(b_i, 0) - \min(Sce_i, 0)} \tag{103}$$

and we actually want all the points to have the same time step $\Delta t/n$, which gives:

$$\frac{1}{k(1-\theta_i)} \Delta t_{stab}(i) = \frac{\Delta t}{n} \tag{104}$$

which yields, adding the necessary limitation to 0:

$$\theta_i = \max(0, 1 - \frac{n \Delta t_{stab}(i)}{k \Delta t}) \tag{105}$$

This formula will tend to give $\theta_i = 0$ if n is large enough. This may not be what we want, as $\theta_i = 0.5$ would be *a priori* of a higher order in time. So let us suppose that a given θ is requested, and is a data given by the user. In this case we just change the formula into:

$$\theta_i = \max(\theta, 1 - \frac{n \Delta t_{stab}(i)}{k \Delta t}) \tag{106}$$

14.1 Corrector

Let us suppose now that we have an approximation C_i^* of the final concentration, we can write the original derivative in time in the form:

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

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

$$\begin{gathered}
S_i h_i^{n+1-\theta_i} C_i^{n+1} - S_i h_i^{n+1-\theta_i} C_i^* = -\left(S_i h_i^{n+1-\theta_i} C_i^* - S_i h_i^{n+1-\theta_i} C_i^n\right) \\
-\Delta t \sum_j \left(\theta_j C_j^{n+1} - \theta_i C_i^{n+1}\right) \min\left(\Phi_{ij}, 0\right) \\
-\Delta t \sum_j \left(\left(1-\theta_j\right) C_j^n - \left(1-\theta_i\right) C_i^n\right) \min\left(\Phi_{ij}, 0\right) \\
-\Delta t \min\left(b_i, 0\right) \left(C_i^{boundary} - \left(\theta_i C_i^{n+1} + (1-\theta_i) C_i^n\right)\right) \qquad (107) \\
+\Delta t \max\left(Sce_i, 0\right) \left(C_i^{sce} - \left(\theta_i C_i^{n+1} + (1-\theta_i) C_i^n\right)\right)
\end{gathered}$$

We now do a PSI reduction of the sum of the derivative in time and the explicit part of the flux contributions, it gives:

$$\begin{gathered}
S_i h_i^{n+1-\theta_i} C_i^{n+1} - S_i h_i^{n+1-\theta_i} C_i^* = \\
-\Delta t \sum_j \left(\theta_j C_j^{n+1} - \theta_i C_i^{n+1}\right) \min\left(\Phi_{ij}, 0\right) \\
\overleftarrow{-\left(S_i h_i^{n+1-\theta_i} C_i^* - S_i h_i^{n+1-\theta_i} C_i^n\right) - \Delta t \sum_j \left(\left(1-\theta_j\right) C_j^n - \left(1-\theta_i\right) C_i^n\right) \min\left(\Phi_{ij}, 0\right)} \\
\qquad (108) \\
-\Delta t \min\left(b_i, 0\right) \left(C_i^{boundary} - \left(\theta_i C_i^{n+1} + (1-\theta_i) C_i^n\right)\right) \\
\Delta t \left(\max\left(Sce_i, 0\right) \left(C_i^{sce} - \left(\theta_i C_i^{n+1} + (1-\theta_i) C_i^n\right)\right)\right)
\end{gathered}$$

This scheme conserves mass (no mass error has been done during this short derivation from the original scheme 33). We thus just need to check the maximum principle.

14.2 Monotonicity

We now rewrite our corrector step so that only positive coefficients of values of C appear. We also introduce the coefficients f_i and μ_{ij} as before to account for the PSI reduction acting differently on the derivative in time and on the fluxes, it yields:

$$\begin{gathered}
\left(S_i h_i^{n+1-\theta_i} + \theta_i \Delta t \max\left(Sce_i, 0\right) - \theta_i \Delta t \sum_j \min\left(\Phi_{ij}, 0\right) - \theta_i \Delta t \min\left(b_i, 0\right)\right) C_i^{n+1} = \\
\Delta t \left(\max\left(Sce_i, 0\right) C_i^{sce} - \min\left(b_i, 0\right) C_i^{boundary}\right)
\end{gathered}$$

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

$$-\mu_{ij}\Delta t \sum_j \left(1-\theta_j\right) C_j^n \min\left(\Phi_{ij},0\right) \tag{109}$$

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

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

On this form we see that the only risk of negative coefficients lies in the coefficient of C_i^n. The coefficient of C_i^{n+1} is positive and always greater than or equal to $S_i h_i^{n+1-\theta_i}$. Now we see that there is a risk of negative coefficient of C_i^n, unless we consider also the value of C_i^*. As the terms depending on μ_{ij} are negative in the coefficient of C_i^n we remain on the safe side by choosing $\mu_{ij}=1$. The first and important question is the extrema in the maximum principle. In view of Formula 108 we should have now two new extrema:

$$\widehat{C}_i^{\min} = \min(C_i^*,\ \text{all } C_j^{n+1},\ C_i^n, \text{all } C_j^n, C_i^{boundary}, C_i^{sce}) \tag{110}$$

$$\widehat{C}_i^{\max} = \max(C_i^*,\ \text{all } C_j^{n+1},\ C_i^n, \text{all } C_j^n, C_i^{boundary}, C_i^{sce}) \tag{111}$$

but these extrema can hardly be anticipated, since they depend on C^{n+1}. Given the implicit character of the scheme, extrema in a neighbourhood more remote than the immediate neighbours of point i could be involved. We keep the notation $\widehat{C}_i^{\min}$ and $\widehat{C}_i^{\max}$ hereafter, but it will be only estimations, e.g. $\widetilde{C}_i^{\min}$ and $\widetilde{C}_i^{\max}$ could be a good approximation. We could also think of an iterative process, using Formulas 110 and 111 but replacing in them C_j^{n+1} by the best estimation obtained previously.

As before, we now introduce:

$$C_i^* = \widehat{C}_i^{\min} + \alpha\left(\widehat{C}_i^{\max} - \widehat{C}_i^{\min}\right) \tag{112}$$

$$C_i^n = \widehat{C}_i^{\min} + \beta\left(\widehat{C}_i^{\max} - \widehat{C}_i^{\min}\right) \tag{113}$$

We are left with proving that:

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

$$+C_i^n \left(f_i S_i h_i^{n+1-\theta_i} - (1-\theta_i)\Delta t \left[\max\left(Sce_i,0\right) - \min\left(b_i,0\right) - \sum_j \min\left(\Phi_{ij},0\right) \right] \right) = \tag{114}$$

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

with $C_i^{average}$ obeying the maximum principle. We denote:

$$S_i h_i^{n+1-\theta_i} - (1-\theta_i)\Delta t \left(\max\left(Sce_i, 0\right) - \sum_j \min\left(\Phi_{ij}, 0\right) - \min\left(b_i, 0\right) \right) = \gamma \tag{115}$$

This coefficient is positive as soon as $k \geq 1$ in the predictor stability condition. It eventually yields:

$$C_i^* \left(1 - f_i\right) S_i h_i^{n+1-\theta_i} + C_i^n \left(f_i S_i h_i^{n+1-\theta_i} + \gamma - S_i h_i^{n+1-\theta_i} \right) = \gamma C_i^{average} \tag{116}$$

or:

$$\begin{gathered} \left(\gamma - S_i h_i^{n+1-\theta_i}\right) \left(\widehat{C}_i^{\min} + \beta\left(\widehat{C}_i^{\max} - \widehat{C}_i^{\min}\right)\right) + S_i h_i^{n+1-\theta} \left(\widehat{C}_i^{\min} + \alpha\left(\widehat{C}_i^{\max} - \widehat{C}_i^{\min}\right)\right) \\ - \left(f_i S_i h_i^{n+1-\theta_i} \left(\widehat{C}_i^{\min} + \alpha\left(\widehat{C}_i^{\max} - \widehat{C}_i^{\min}\right)\right) - f_i S_i h_i^{n+1-\theta_i} \left(\widehat{C}_i^{\min} + \beta\left(\widehat{C}_i^{\max} - \widehat{C}_i^{\min}\right)\right) \right) = \gamma C_i^{average} \end{gathered} \tag{117}$$

which is:

$$\widehat{C}_i^{\min} + \frac{\left[\beta\left(\gamma - S_i h_i^{n+1-\theta_i}\right) + \alpha\left(1 - f_i\right) S_i h_i^{n+1-\theta_i} + \beta f_i S_i h_i^{n+1-\theta_i}\right]}{\gamma} \left(\widehat{C}_i^{\max} - \widehat{C}_i^{\min}\right) = C_i^{average} \tag{118}$$

We thus need to have:

$$0 < \beta\gamma + \left(\alpha - \beta\right)\left(1 - f_i\right) S_i h_i^{n+1-\theta_i} < \gamma \tag{119}$$

If $\alpha > \beta$ positivity is ensured and then the worst situation happens when $f_i = 0$, in which case we get the condition $\beta\gamma + \left(\alpha - \beta\right) S_i h_i^{n+1-\theta_i} < \gamma$ which also reads:

$$\alpha S_i h_i^{n+1-\theta_i} < \gamma\left(1 - \beta\right) + \beta S_i h_i^{n+1-\theta_i} \tag{120}$$

Our predictor stability condition then gives the property:

$$\gamma > \left(1 - \frac{1}{k}\right) S_i h_i^{n+1-\theta_i} \tag{121}$$

Our most demanding condition for α is then (the smallest γ is to be considered):

$$\alpha < \left(1 - \frac{1}{k}\right) + \frac{\beta}{k} \tag{122}$$

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 γ, is:

$$\frac{\beta}{k} < \alpha \tag{123}$$

We end up with the general condition:

$$\frac{\beta}{k} < \alpha < \left(1 - \frac{1}{k}\right) + \frac{\beta}{k} \tag{124}$$

Which is also:

$$C_i^n + \left(1 - \frac{1}{k}\right)\left(\widehat{C}_i^{\min} - C_i^n\right) < C_i^* < C_i^n + \left(1 - \frac{1}{k}\right)\left(\widehat{C}_i^{\max} - C_i^n\right) \tag{125}$$

Which gives with $k = 2$:

$$C_i^n + \frac{1}{2}\left(\widehat{C}_i^{\min} - C_i^n\right) < C_i^* < C_i^n + \frac{1}{2}\left(\widehat{C}_i^{\max} - C_i^n\right) \tag{126}$$

Now the next question is: is this property ensured by C_i^* when we use a semi-implicit predictor? Actually not, counterexamples have been found, which leads us to the conclusion:

$$\begin{gathered}\text{The LIPS predictor must be limited}\\[1em] \text{also at the first iteration}\end{gathered} \tag{127}$$

14.3 A correct sum of coefficients

Though our results on the LIPS scheme will eventually happen to be correct, there is however a hack in what has been done so far. 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 contain well balanced differences of values of C, every positive coefficient being counterbalanced by the corresponding negative coefficient. 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 reduction in an unbalanced way. In what precedes it is the case with the term $\overleftarrow{-\Delta t \sum_j \left(\left(1-\theta_j\right) C_j^n - \left(1-\theta_i\right) C_i^n\right) \min\left(\Phi_{ij}, 0\right)}$. The balance of the reduced terms $-\left(1-\theta_j\right) C_j^n + \left(1-\theta_i\right) C_i^n$ is ensured by the terms $-\ \theta_j C_j^{n+1} - \theta_i C_i^{n+1}$ which are not reduced, and here is the hack. We are thus doomed to reduce only true differences of C values. In the case of term:

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

a solution consists in not upwinding all the terms, but only those that can be balanced, i.e., denoting:

$$\min\theta(i,j) = \min(1-\theta_j, 1-\theta_i) \tag{128}$$

we replace our term with:

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

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

This can be done at element level when doing the PSI reduction, a part of the original explicit fluxes contribution being set aside and transmitted without reduction.

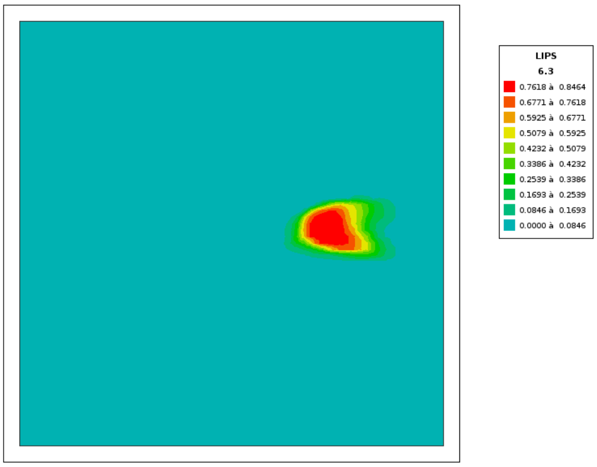

Figure 17: LIPS scheme with 5 corrections and 10 sub-iterations. Cone after one rotation.

14.4 Results

Choosing $k = 2$ and a number of corrections of 5, the height of the rotating cone after 1 rotation, depending on the number of substeps n, gives:

n	cone height	standard deviation
1	0.1017	77.04 10^{-3}
2	0.1460	72.07 10^{-3}
3	0.2043	66.61 10^{-3}
4	0.3199	58.14 10^{-3}
5	0.4814	45.69 10^{-3}
6	0.6365	33.90 10^{-3}
7	0.7554	30.71 10^{-3}
8	0.8207	32.38 10^{-3}
9	0.8369	32.45 10^{-3}
10	0.8464	31.66 10^{-3}
11	0.8455	30.39 10^{-3}
12	0.8409	27.97 10^{-3}
13	0.8380	26.51 10^{-3}
14	0.8318	25.05 10^{-3}
15	0.8259	23.86 10^{-3}
16	0.8188	22.67 10^{-3}
17	0.8119	21.68 10^{-3}
39	0.7249	17.88 10^{-3}

There is an optimum for $n = 10$. The more n increases, the more we tend to the predictor-corrector scheme. With $n = 39$ we get almost the same results. It is due to the fact that from $n = 39$ on, the scheme becomes fully explicit.

Figure 17 shows the result obtained with $n = 10$. The cone height after one rotation is now 0.8464, a tremendous progress, and the standard deviation is 31.66 10^{-3}. It is our best result so far in terms of cone height, but the cone shape is somewhat distorted, indeed the standard deviation still decreases if we increase n.

For the different levels of refinement the cone height after one rotation and the standard deviation are given in the following table, with 5 corrections:

level	Cone height, LIPS with 10 sub-iterations and 5 corrections	Standard deviation
0	0.8464	$31.66\ 10^{-3}$
1	0.9305	$24.10\ 10^{-3}$
2	0.9501	$11.76\ 10^{-3}$
3	0.9736	$7.38\ 10^{-3}$
4	0.9865	$5.28\ 10^{-3}$

The preceding table was established with $k = 2$. At level 0, a maximum of height of 0.8505 is reached with $k = 1.8$ and a minimum of standard deviation of $29.30\ 10^{-3}$ is reached with $k = 1.29$.

14.5 Second order in time?

In what has been done so far a large n will tend to an explicit scheme. We now suppose that we want $\theta_i = 0.5$ in as many points as we can. In this case we can just apply the formula:

$$\theta_i = \max(\frac{1}{2}, 1 - \frac{n\Delta t_{stab}(i)}{k\Delta t}) \tag{129}$$

If we now set the number of corrections to 5 and study the effect of n, still with $k = 2$, we find a convergence to a value that is disappointingly not better than the PSI scheme. The reason of this poor behaviour is that there is no PSI reduction of the implicit part of the fluxes contribution, and this precludes a good second order scheme.

14.6 Behaviour on dry points

We have seen that the diagonal in the linear system given by the LIPS scheme is:

$$S_i h_i^{n+1-\theta_i} + \theta_i \Delta t \max(Sce_i, 0) - \theta_i \Delta t \sum_j \min(\Phi_{ij}, 0) - \theta_i \Delta t \min(b_i, 0) \tag{130}$$

On dry points there is a potential division by zero, for example when we have no source, no boundary terms and no fluxes, and $h_i^{n+1-\theta_i} = 0$. A dry point with $h_i^{n+1-\theta_i} = 0$ will have also $h_i^n = 0$ and $h_i^{n+1} = 0$, since it is an interpolation of these two not negative depths. Consequently, through the continuity equation it will have $\sum_j \Phi_{ij} = 0$, but not necessarily all of them equal to 0, and also $\theta_i = 1$. Then one at least of the terms in the sum $-\theta_i \Delta t \sum_j \min(\Phi_{ij}, 0)$ will give a positive diagonal. We shall thus have a problem only with points which are dry AND remain dry in the time step. In this case the value of the tracer at this point can be left unchanged.

14.7 Optimising the locally implicit scheme

In this chapter we shall try to avoid solving too many linear systems, starting with the predictor step. In this step we solve Equation 101. The terms that create extra-diagonal terms in the final matrix are:

$$-\Delta t \sum_j \theta_j C_j^* \min(\Phi_{ij}, 0)$$

We can imagine that $\theta_j C_j^*$ is replaced by $\theta_j C_j^n$ and the matrix will become a diagonal. It raises no problem in the stability analysis. However the mass conservation will be spoiled because doing this is like considering that a quantity of tracer leaving a point is

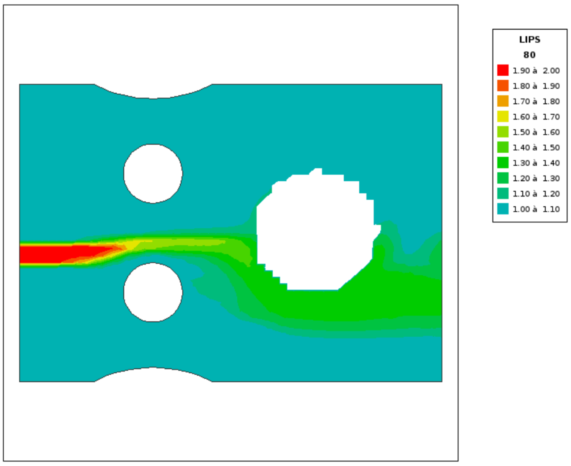

Figure 18: Bridge piers test case with an island treated as a dry zone. Tracer advected with the LIPS scheme.

changed when it arrives to another point. Namely a quantity $\theta_i C_i^* + (1-\theta_i)C_i^n$ leaving a point i arrives as C_i^n to another point. This is actually not a problem since we have seen that the mass of C^* has no influence on the final result, only numerical diffusion may be spoiled, and according to numerical tests it is not. This simplification can be done also in the corrections, except the last correction which needs the true implicit terms for the sake of mass conservation.

It happens that replacing $\theta_j C_j^*$ with $\theta_j C_j^n$ is also what does the Jacobi linear system solver. We conclude that as soon as we are not at the last correction, we could replace the solution of a linear system with the result of a few iterations of the Jacobi method. We can even imagine that "a few iterations" is 1.

Tests on the flow around bridge piers show no notable difference in the results and the mass conservation is not downgraded. More accurate tests with the rotating cone show that the differences are negligible.

$$\text{The predictor and all but the last correction of the LIPS scheme} \\ \text{do not require solving a linear system.} \tag{131}$$

$$\text{One iteration of the Jacobi method is enough}$$

The Jacobi method has another good property: every iteration actually does an interpolation of values of C, it thus well preserves monotonicity. In view of this method it also appears natural to choose $\widetilde{C}_i^{\min}$ and $\widetilde{C}_i^{\max}$ as extrema for the maximum principle, i.e. the same values chosen for the second-order in time predictor-corrector.

Figure 18 shows the result obtained with the bridge piers test case, with an island treated as a dry zone.

15 NERD: N Edge-based Residual Ditributive scheme

This scheme which is unconditionally stable was introduced in 2011 (Reference [13]). It is more diffusive than LIPS but from 2009 up to 2015 it was the only scheme in Telemac able to combine stability on dry zones, mass conservation and monotonicity. Its main idea will also lead to the ERIA scheme. We must first explain how starting from results stemming from finite elements, possibly with negative depths, we build a continuity equation exact at machine accuracy and with positive depths.

15.1 The positive depths algorithm

The problem of negative depths in Telemac-2D and 3D has always been the price to pay to have fast and implicit schemes, whereas explicit techniques such as the finite volume option with kinetic schemes were able to ensure a positive depth, but at a considerably higher computer time, due to much smaller time steps. The solution presented here keeps large time steps, it consists of an iteration procedure and a limitation of the fluxes between points. It is actually a post-treatment which ensures both mass-conservation and positivity of depth. The final continuity equation involves positive depths at the beginning and at the end of the time step, and the compatible fluxes that cause the modifications of the depths. The procedure is summarised hereafter in 3 steps:

- The fluxes between points of the N-scheme are computed.
- Starting from depths at time n, water corresponding to these fluxes are transfered between points, in a loop over all segments, provided that the depths remain positive, otherwise the fluxes are locally and temporarily limited, part of them being kept for a further iteration of the process). This can be repeated until there is no more possible water to transfer.
- The remaining fluxes are left over, they are considered as non physical.

We have already established (Equation 9) that the continuity equation can be put in the form:

$$\frac{S_i\left(h_i^{n+1}-h_i^n\right)}{\Delta t} = Sce_i + \int_\Omega h\overrightarrow{u}.\overrightarrow{\mathrm{grad}}(\Psi_i)\ d\Omega - b_i \tag{132}$$

and that the fluxes between points Φ_{ij} are then deduced from the integrals $\int_\Omega h\overrightarrow{u}.\overrightarrow{\mathrm{grad}}(\Psi_i)\ d\Omega$. However after the finite element treatment which involves the solution of a linear system, the accuracy depends on the solver used. Our transfer of fluxes will give new values h_i^{n+1} and the new continuity equation will be exact at machine accuracy.

We shall now get into the details of the technique.

15.1.1 Transfer and limitation of the internal fluxes

Let us first deal with fluxes between points, regardless of other boundary and source terms which will be addressed later. Starting from h^n we want to construct a new depth at time $n+1$, and the depth "in construction" is denoted here $\widetilde{h}$, and initialised with h^n. We assume that the starting depths h^n are positive. In a loop over all segments, we get every time specific i and j (apices of the segment), and we would like to apply the formulas:

$$\widetilde{h}_i \text{ replaced by } \widetilde{h}_i - \frac{\Delta t}{S_i}\Phi_{ij} \tag{133}$$

$$\widetilde{h}_j \text{ replaced by } \widetilde{h}_j + \frac{\Delta t}{S_j}\Phi_{ij} \tag{134}$$

but there is a risk of negative $\widetilde{h}_i$. If there is a risk, i.e. if $\Phi_{ij} > \frac{S_i\widetilde{h}_i}{\Delta t}$, then the flux is limited by a factor:

$$\theta = \frac{S_i\widetilde{h}_i}{\Phi_{ij}\Delta t} \tag{135}$$

We then do:

$$\widetilde{h}_i \text{ replaced by } \widetilde{h}_i - \theta \frac{\Delta t}{S_i} \Phi_{ij} \tag{136}$$

$$\widetilde{h}_j \text{ replaced by } \widetilde{h}_j + \theta \frac{\Delta t}{S_j} \Phi_{ij} \tag{137}$$

which ensures the conservation of water, and:

$$\Phi_{ij} \text{ is replaced by } (1-\theta)\Phi_{ij}$$

which stores in Φ_{ij} the flux that has not yet been taken into account (it will be used in the next loop over all segments). Then this loop over all segments is repeated with the remaining Φ_{ij}. This is the key point! After a number of iterations, the situation remains unchanged, i.e. a criterion like $\sum abs(\Phi_{ij})$ is no longer decreasing. The remaining Φ_{ij} are then left over as non physical because they would lead to negative depths. The parts of fluxes which have been duly transfered form a perfect continuity equation, with positive depths and fluxes in accordance.

15.1.2 Boundary and source terms

Boundary and source terms are not likely to be interpreted in terms of fluxes between points. Moreover a sink term may lead to negative depths, which brings in fact a specific CFL number for the time step. We apply the following algorithm:

- Step 1: taking into account the source and boundary terms bringing water (the depths are increased).
- Step 2: applying the limitation of internal fluxes described above, this leads to positive depths.
- Step 3: taking into account the source and boundary terms removing water (the depths are decreased).

Step 3 may raise problems, thus a limiting factor of the source terms or boundary terms may be applied also at this level, as data do not allow to keep the depths positive. It is obviously the case when an evaporation is forced on dry land.

15.1.3 Dependency to numbering and parallelism

As described above the algorithm raises problems with parallelism (points on an interface may not see the water coming from segments to which they belong but are in another sub-domain). Moreover it appeared that the algorithm was sensitive to the numbering of segments. As a matter of fact, if two segments take water out of a point, they can be in competition if there is not enough water, and the first segment will be better "served", the second will have its flux limited. The previous algorithm has thus been slightly modified. At every iteration we consider that the mesh is split into single segments, and the tips of the segments are given water for the transfer. After the transfer segments are grouped again and they bring back water to points. This is a split-transfer-merge procedure. It is not sensitive to segment numbering and it can cope with parallelism. For example for a segment along an interface all quantities can be multiplied by $1/2$, or alternatively (current solution) one of the two neighbouring processors can treat the segment, and not the other. The key problem in the process is the distribution of water between tips of segments. A point i that belongs to e.g. 6 segments will appear in all of them and has only $S_i h_i$ of water to share. Giving $S_i h_i/6$ to all occurrences of point i in segments gave a very slow rate of transfer. Actually the segment tips give or receive water, those which receive have no need of initial water. The distribution is thus done proportionally to the demand. In a segment $i-j$, if Φ_{ij} is positive, point i is in need of $\Delta t\ \Phi_{ij}$. If $\Delta t \sum\limits_{k \text{ neighbour of } i} \max(\Phi_{ik},0) > S_i h_i$ all needs cannot be satisfied. Actually in our algorithm every point i in a segment $i-j$ receives:

$$\frac{\max(\Phi_{ij},0)}{\sum\limits_{k \text{ neighbour of } i} \max(\Phi_{ik},0)} S_i h_i$$

If the denominator is 0 we can revert to an equal sharing. Let us see now what happens to the tracers in this process.

15.1.4 Edge-based transfer of tracers

We have seen that the explicit distributive schemes cannot be stable on dry zones. If we exclude sources and boundaries they give new concentrations equal to:

$$C_i^{n+1} = \left(1 + \frac{\Delta t}{h_i^{n+1} S_i} \sum_j \min(\Phi_{ij},0)\right) C_i^n + \frac{\Delta t}{h_i^{n+1} S_i} \sum_j C_j^n \max(\Phi_{ij},0) \tag{138}$$

and it gives no hope if $h_i^{n+1} = 0$. The fundamental reason is that all fluxes to and from a point are considered at the same time. Imagine now that we do it edge by edge, thus considering only two points at a time. Let us number these points 1 and 2 and let us assume that the flux Φ_{12} is positive, i.e. the water goes from point 1 to point 2. The conservative tracer equations of both points read simply (we omit boundary and source terms for simplicity, they will be treated outside the iterative process of transfer):

$$\frac{S_1}{\Delta t}(h_1^{n+1} C_1^{n+1} - h_1^n C_1^n) + \Phi_{12} C_1^n = 0 \tag{139}$$

$$\frac{S_2}{\Delta t}(h_2^{n+1} C_2^{n+1} - h_2^n C_2^n) - \Phi_{12} C_1^n = 0 \tag{140}$$

C_1^n appears in both equations because the flux goes from 1 to 2 (upwind scheme). If we also treat the continuity equation only for this edge (what we did in the positive depths algorithm), we have also:

$$h_1^{n+1} = h_1^n - \frac{\Delta t}{S_1} \Phi_{12} \tag{141}$$

$$h_2^{n+1} = h_2^n + \frac{\Delta t}{S_2} \Phi_{12} \tag{142}$$

It then turns out that equations 139 and 140 become simply:

$$C_1^{n+1} = C_1^n \tag{143}$$

$$C_2^{n+1} = \frac{h_2^n}{h_2^{n+1}} C_2^n + (1 - \frac{h_2^n}{h_2^{n+1}}) C_1^n \tag{144}$$

In this context there is no risk of division by 0 because we started from a positive depth h_2^n which was increased by the positive quantity $\frac{\Delta t}{S_2}\Phi_{12}$. h_2^n / h_2^{n+1} is then in the range $[0,1]$. The positivity and monotonicity of tracers is ensured even on dry zones (in case of zero depth the concentrations remain unchanged). This very simple edge by edge treatment

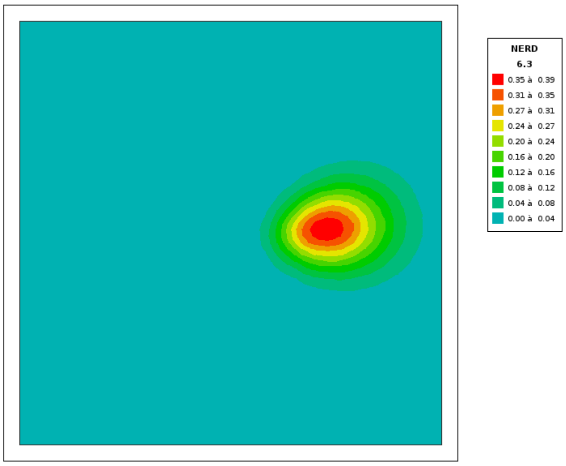

Figure 19: NERD scheme. Cone after one rotation

may be inserted within the previous positive depths algorithm. It is the NERD scheme. We can notice that all stability condition has disappeared. It has migrated in the more or less difficult iterative process of water transfer. With sufficiently deep waters, all the fluxes will be transfered in one iteration. Near dry zones we may have to stop the procedure when a maximum of iteration is reached. However in practice the number of segments that raise problems is a small percentage, and others can be eliminated from the loop on segments: the transfer loops become smaller and smaller.

Figure 19 shows the result obtained with the rotating cone. The cone height after one rotation is 0.3920. It is a surprise since we only used N fluxes. The standard deviation is $47.32\ 10^{-3}$.

For the different levels of refinement the cone height after one rotation and the standard deviation are given in the following table:

level	cone height, NERD scheme	standard deviation
0	0.3920	$47.32\ 10^{-3}$
1	0.5578	$33.13\ 10^{-3}$
2	0.7116	$20.99\ 10^{-3}$
3	0.8287	$12.20\ 10^{-3}$
4	0.9052	$6.66\ 10^{-3}$

Figure 20 shows the results on the bridge piers test case with an island treated as a dry zone. During all the computation the tracer strictly remains in the range [1,2], and the relative mass error is $-0.47\ 10^{-15}$. This is a big progress since we have now an unconditionally stable scheme suited for dry zones, without solving a linear system, but the numerical diffusion is still high, higher than the LIPS scheme, even if it is lower than the N scheme, due to the higher time steps.

16 ERIA: a triangle-based iterative predictor-corrector scheme

The NERD scheme is based on fluxes between points given by the N scheme. As the NERD scheme basically works on isolated segments, there is no way to use the PSI scheme concept and the predictor-corrector approach. We explore here the possibility of a triangle-based iterative scheme. It consists in treating independently every triangle with

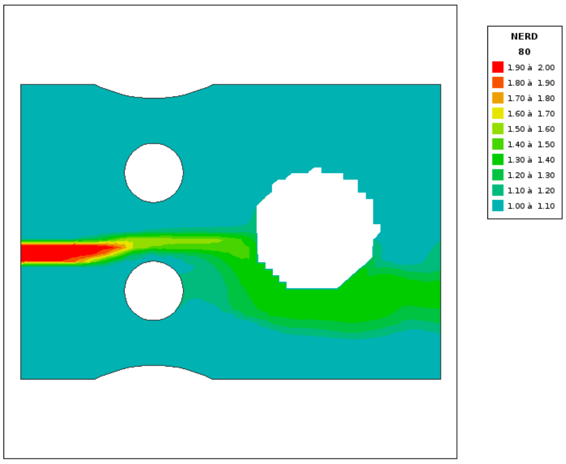

Figure 20: Island treated as a dry zone. Advection of a tracer with the NERD scheme.

its own local fluxes, the quantities of water and tracers carried by points being shared between triangles according to rules that will be detailed. The local fluxes are limited to ensure the positivity of the water mass locally carried by the points. This is done by provisionnally reducing the local fluxes. The part of the initial fluxes which is left-over is kept for the next iteration. The iterations are stopped exactly like in the NERD scheme, when all the fluxes have been transfered, or when nothing can move anymore, or when the fluxes have been sufficiently reduced.

After one iteration the quantities carried by points are assembled, so that a new depth and a new value of tracer can be computed. This keeps the positivity of depth and the monotonicity of tracers if it has been ensured locally on every triangle. Hereafter we thus only study the problem on a single triangle, with fluxes that do not cause negative depths. Boundaries and sources are treated before and after the transfer of internal fluxes, and so are not taken into account here.

This scheme is called ERIA (**E**lement by element **R**esidual distributive **I**terative **A**dvection scheme). *Eria* is a genus of asiatic orchids.

16.1 Predictor step

In the predictor step at element level, we will have initial quantities of water dedicated to every point, denoted $volp(i)$ ("vol" for volume and "p" for predictor). Classical distributive schemes choose simply:

$$volp(i) = \frac{S_T \; h_i^n}{3} \tag{145}$$

where S_T is the area of the triangle and h_i^n is the initial depth of point i, so that the sum of all volumes locally given to point i is the total quantity of water carried by this point, i.e. $S_i \; h_i^n$, where S_i is the integral of the test function of i, also the area associated to this point. We keep this constraint here but the distribution is different. When dealing with an element we want to get final local volumes denoted $V_{i\ local}^{n+1}$ such that:

$$V_{i\ local}^{n+1} = volp(i) - \Delta t \sum_{j\ in\ t} \overline{\Phi}_{ij} \geq 0 \tag{146}$$

where $volp(i)$ is our initial volume that remains to be defined. The fluxes $\overline{\Phi}_{ij}$ are the local fluxes Φ_{ij} (from i to j) given by the N scheme, but limited in a way that will also be defined later. The bar thus means "limited". The notation $\sum_{j\ in\ t}$ means a sum on the two other points of the triangle t that contains i. This can also be written in terms of depth, but if we start from the initial depths and if we transfer all the fluxes of one element it will give a local depth $h^{n+1}_{i\,local}$ that may be different, for the same point, in another element. Namely we have:

$$V^{n+1}_{i\ local} = \frac{S_T\ h^{n+1}_{i\ local}}{3} = \frac{S_T\ h^n_i}{3} - \Delta t \sum_{j\ in\ t} \overline{\Phi}_{ij} \geq 0 \tag{147}$$

The initial volumes $volp(i)$ are chosen following an offer and demand principle, so as to minimise the further reduction of fluxes. Let us first imagine that a classical local volume $S_T h^n_i/3$ has been *a priori* given to point i in a triangle. Sometimes this local volume will not be large enough to keep the depth positive (without reducing the fluxes). Sometimes it will be largely enough, e.g. points that will receive water in the triangle could even be given no initial volume. Namely when point i in an element is such that:

$$\frac{S_T\ h^n_i}{3} - \Delta t \sum_{j\ in\ t} \max(\Phi_{ij}, 0) \geq 0 \tag{148}$$

it can give this positive quantity to its *alter ego* in other elements and keep a positive final local depth $h^{n+1}_{i\ local}$. On the contrary in elements where i is such that:

$$\frac{S_T\ h^n_i}{3} - \Delta t \sum_{j\ in\ t} \max(\Phi_{ij}, 0) < 0 \tag{149}$$

it is in need of the opposite of this negative quantity. We can thus compute a total demand $td(i)$ and a total offer $to(i)$ for every point, by summing on all the neighbouring elements, introducing the notation $\sum_{t \ni i}$ meaning a sum on all triangles t containing a point i:

$$\begin{aligned} to(i) &= \sum_{t \ni i} \max\left(\frac{S_T\ h^n_i}{3} - \Delta t \sum_{j\ in\ t} \max(\Phi_{ij}, 0), 0\right) \\ &= \sum_{t \ni i} \max\left(\frac{S_T\ h^{n+1}_{i\,local}}{3} + \Delta t \sum_{j\ in\ t} \min(\Phi_{ij}, 0), 0\right) \end{aligned} \tag{150}$$

$$\begin{aligned} td(i) &= -\sum_{t \ni i} \min\left(\frac{S_T\ h^n_i}{3} - \Delta t \sum_{j\ in\ t} \max(\Phi_{ij}, 0), 0\right) \\ &= -\sum_{t \ni i} \min\left(\frac{S_T\ h^{n+1}_{i\,local}}{3} + \Delta t \sum_{j\ in\ t} \min(\Phi_{ij}, 0), 0\right) \end{aligned} \tag{151}$$

We can then choose for each occurrence of i the initial volume that it will get, reasoned as a correction of the *a priori* initial value $S_T\ h^n_i/3$:

In elements where i is "donnor":

$$\begin{aligned} volp(i) &= \frac{S_T\ h^n_i}{3} \\ &- \left(\frac{S_T\ h^n_i}{3} - \Delta t \sum_{j\ in\ t} \max(\Phi_{ij}, 0)\right) * \frac{td(i)}{\max(td(i), to(i))} \end{aligned} \tag{152}$$

In elements where i is "receiver":

$$volp(i) = \frac{S_T\ h_i^n}{3}$$

$$-\left(\frac{S_T\ h_i^n}{3} - \Delta t \sum_{j\ in\ t} \max(\Phi_{ij}, 0)\right) * \frac{to(i)}{\max(td(i), to(i))} \tag{153}$$

The formulas ensure that all that is given is received. If demand exceeds offer, all donnors will give what they have to give and it will be shared between receivers, if offer exceeds demand, all receivers will get what they need and the donnors will give only what is necessary.

We have thus optimally distributed the water between triangles, but this is not enough to avoid negative depths and this is why we now limit the fluxes. We now want that the limited fluxes are such that:

$$volp(i) - \Delta t \sum_{j\ in\ t} \max(\overline{\Phi}_{ij}, 0) \geq 0 \tag{154}$$

So we define $\beta(i)$ such that, if:

$$\Delta t \sum_{j\ in\ t} \max(\Phi_{ij}, 0) > volp(i) \tag{155}$$

we have:

$$\beta(i) = \frac{volp(i)}{\Delta t \sum_{j\ in\ t} \max(\Phi_{ij}, 0)} \tag{156}$$

and for all fluxes that leave point i:

$$\overline{\Phi}_{ij} = \min(\beta(i), \beta(j))\Phi_{ij} \tag{157}$$

A key point in the procedure is the fact that the fluxes Φ_{ij} are N or PSI fluxes. It means that in a triangle one of them at least is 0, and that the two others are either converging to a single point (1-target case) or leaving a single point (2-target case). All fluxes leaving a point have thus the same sign, so reducing them independently will reduce the total flux leaving the point. *It would not be the case with fluxes of different signs.* In the case of N or PSI fluxes it is easy also to understand that in $\min(\beta(i), \beta(j))$ one of the β will be equal to 1 if Φ_{ij} is not 0 (because a point only gives or only receives, and a point that receives water has $\beta = 1$), so our reduction is the minimum that can be done. Choosing a constant reduction within a triangle would slow down a lot the process, with situations where a dry point could be able to stop the flux between the two other possibly wet points.

Compared to the other distributive schemes, here the new volume $volp(i)$ replaces $S_T\ h_i^n/3$ in the formulas, e.g. the predictor will locally become:

$$\frac{S_T\ h_{i\ local}^{n+1}}{3} C_{i\ local}^{n+1} = volp(i) C_i^n$$

$$-\Delta t \sum_{j\ in\ t} \min(\overline{\Phi}_{ij}, 0) C_j^n - \Delta t \sum_{j\ in\ t} \max(\overline{\Phi}_{ij}, 0) C_i^n \tag{158}$$

where C_i^n is the original concentration of tracer for point i, $C_{i\ local}^{n+1}$ is the final local concentration of the same point (i.e. obtained without communicating with other elements), and $h_{i\ local}^{n+1}$ is the final local depth of point i, defined by Equation 147.

The predictor equation can be rearranged in the form:

$$\frac{S_T\ h_{i\ local}^{n+1}}{3} C_{i\ local}^{n+1} = \frac{S_T\ h_{i\ local}^{n+1}}{3} C_i^n$$

$$-\Delta t \sum_{j\ in\ t} \min(\overline{\Phi}_{ij}, 0) \left(C_j^n - C_i^n \right) \tag{159}$$

To get the real equation actually solved at predictor level, we still need to add the PSI reduction, denoted with a backward arrow. It is applied to the right-hand side, so that we now write:

$$\frac{S_T\ h_{i\ local}^{n+1}}{3} \left(C_{i\ local}^{n+1} - C_i^n \right) = -\Delta t \overleftarrow{\sum_{j\ in\ t} \min(\overline{\Phi}_{ij}, 0) \left(C_j^n - C_i^n \right)} \tag{160}$$

Equation 158 shows that monotonicity is given by the positivity of the coefficient of C_i^n, which is $volp(i) - \Delta t \sum_{j\ in\ t} \max(\overline{\Phi}_{ij}, 0)$, or $S_T\ h_{i\ local}^{n+1}/3$, and it is exactly the condition 154 that we have secured with the reduction of fluxes. This is also valid for Equation 160 where, compared to Equation 158, the negative component in the coefficient of C_i^n is reduced. It appears thus that the local positivity of volumes is the only condition to stabilise a PSI scheme. Merging all local values, by weigh-averaging all occurences of a point in its different triangles, will give a final mass conservative and monotone result.

It is not necessary to solve Equation 160 at element level. When summed over all elements, the left-hand side can be replaced with $S_i h_i^{n+1} \left(C_i^{n+1} - C_i^n \right)$ to get the final value C_i^{n+1}. Building $C_{i\ local}^{n+1}$ and $h_{i\ local}^{n+1}$ is thus not useful.

As we have limited the fluxes, we must do a book-keeping of all fluxes that still must be transfered, and try to transfer them in successive iterations. At the end of an iteration k, the fluxes that have not been transfered, thus being kept for iteration $k+1$, are denoted Φ_{ij}^{k+1}, they are:

$$\Phi_{ij}^{k+1} = \Phi_{ij}^k - \overline{\Phi}_{ij}^k = (1 - \min(\beta(i), \beta(j)))\, \Phi_{ij}^k \tag{161}$$

Iterations are stopped when all the remaining fluxes are 0, or small enough, or after a maximum number of iterations. This will cause no problem if the fluxes finally transfered are the same than the fluxes transfered for computing the new depths with the continuity equation, what we have called the "positive-depths" algorithm. However this algorithm was so far based on a "segment by segment" transfer which is compatible with the NERD scheme (and is in fact its main idea). If we want the ERIA scheme to be fully compatible with the algorithm doing the correction of depths to get positive values, we must then change this algorithm and organise "triangle by triangle" transfers of water, as described above. This raises no additional difficulty, except that this new algorithm had to be implemented and offered as a new option for the treatment of negative depths (namely option 3, the NERD scheme requiring option 2). NERD and ERIA are thus incompatible.

A first very promising result is that testing what has just been said, by running only the predictor step without further correction, the rotating cone height after 1 rotation is already 0.4603. This is to be compared with the 0.21 of the PSI scheme and the 0.39 of the NERD scheme.

16.2 Corrector step

We now consider that the predictor step has given us, on a given triangle, local values of the predictor, which we denote $C_{i\,local}^*$. When assembled, these local values will give another monotone value $C_{i\,global}^*$. To facilitate the explanations we first study a basic solution that will be monotone but with high numerical diffusion. We shall then present a more complicated version with a very low numerical diffusion. The key difference between them is the value considered for computing the derivative in time.

16.2.1 Basic solution

We take here the local value of the depth for the derivative in time introduced in the corrector right-hand side, i.e. the value obtained after a local transfer of fluxes, without considering the other triangles. We thus write the corrector in the form:

$$\begin{aligned}
\frac{S_T\ h^{n+1}_{i\ local}}{3} C^{n+1}_{i\ local} &= \frac{S_T\ h^{n+1}_{i\,local}}{3}(C^*_i - C^n_i) \\
&+ \frac{S_T\ h^n_i}{3} C^n_i - \Delta t \sum_{j\ in\ t} \overline{\Phi}_{ij} C^n_i \qquad (162) \\
\overleftarrow{+\frac{S_T\ h^{n+1}_{i\,local}}{3}(C^n_i - C^*_i) + \Delta t \sum_{j\ in\ t} \min(\overline{\Phi}_{ij}, 0)\left(C^n_i - C^n_j\right)}
\end{aligned}$$

It is nothing else than a PSI scheme with in the right-hand side the estimated derivative first added (immediately after the sign =), then removed in PSI reduced form (first term under the backward arrow). It simplifies into:

$$\begin{aligned}
\frac{S_T\ h^{n+1}_{i\ local}}{3} C^{n+1}_{i\ local} &= \frac{S_T\ h^{n+1}_{i\,local}}{3} C^*_i \\
\overleftarrow{+\frac{S_T\ h^{n+1}_{i\,local}}{3}(C^n_i - C^*_i) + \Delta t \sum_{j\ in\ t} \min(\overline{\Phi}_{ij}, 0)\left(C^n_i - C^n_j\right)} & \qquad (163)
\end{aligned}$$

The monotonicity condition can be enforced locally by imposing that the coefficient of C^n_i is positive:

$$\frac{S_T\ h^{n+1}_{i\,local}}{3} + \Delta t \sum_{j\ in\ t} \min(\overline{\Phi}_{ij}, 0) \geq 0 \qquad (164)$$

This is, if we look at our definition of $h^{n+1}_{i\,local}$, strictly equivalent to Condition 154. We have thus derived a viable scheme, but unfortunately its numerical diffusion is too high, probably because the local depth is too far from the actual final depth, so upwinding is not well done locally. The results with the rotating cone are the following:

TABLE I: basic solution, effect of corrections on numerical diffusion.

corrections	0	1	2	3	4	5
cone height	0.46	0.30	0.32	0.33	0.34	0.34

The performance is downgraded by the corrections, this is thus very disappointing.

16.2.2 A better solution

In the previous solution, instead of choosing a local volume $S_T\ h^{n+1}_{i\,local}/3$, we would prefer taking $S_T\ h^{n+1}_i/3$, i.e. choosing a volume corresponding to the real final depth of point i, but this would lead to monotonicity problems (this is well exemplified by the rotating cone test case which crashes). What freedom do we have to choose the local volumes? Actually, any kind of volume $volc(i)$ (c added for "corrector") would not spoil mass conservation as soon as we have:

$$\sum_{t\ \ni\ i} volc(i) = S_i\ h^{n+1}_i \qquad (165)$$

at the condition that we write the corrector as:

$$volc(i)\left(C^{n+1}_{i\;local}-C^*_i\right)=$$

$$\overleftarrow{volc(i)(C^n_i-C^*_i)+\Delta t\sum_{j\;in\;t}\min(\overline{\Phi}_{ij},0)\left(C^n_i-C^n_j\right)} \tag{166}$$

i.e. that we use also $volc(i)$ for the derivative in time in the right-hand side. As a matter of fact the sum over all triangles around i will then give:

$$S_i\;h^{n+1}_i C^{n+1}_i = S_i\;h^{n+1}_i C^*_i +$$

$$\sum_{t\,\ni\,i}\overleftarrow{volc(i)(C^n_i-C^*_i)+\Delta t\sum_{j\;in\;t}\min(\overline{\Phi}_{ij},0)\left(C^n_i-C^n_j\right)} \tag{167}$$

where we see that $volc(i)$ just replaces the classical $S_T\;h^{n+1}_i/3$ in the right-hand side. Only the distribution of volumes locally reduced has an influence on the global result. At this level it is interesting to look at the proof of monotonicity of our previous explicit predictor-corrector, if we exclude sources and boundary terms. It was:

$$S_i h^{n+1}_i C^{n+1}_i = S_i h^{n+1}_i C^*_i\;+$$

$$f_i S_i h^{n+1}_i\left(C^n_i-C^*_i\right)+\Delta t\sum_j \mu_j\min(\Phi_{ij},0)\left(C^n_i-C^n_j\right) \tag{168}$$

f_i and μ_j being coefficients in the range [0,1], due to the PSI reduction in various triangles. We have here a very similar problem, but with a fundamental advantage on our side: at element level the PSI reduction represented by the backward arrow would give $f_i=\mu_j$, which fortunately avoids a stricter stability condition. The only problem is, as usual, the positivity of the coefficient of C^n_i, which is locally, before reduction:

$$volc(i)+\Delta t\sum_{j\;in\;t}\min(\overline{\Phi}_{ij},0) \tag{169}$$

The PSI reduction, which is actually a multiplication by a number in the range [0,1], will not change the sign. We thus only need to ensure that:

$$volc(i)+\Delta t\sum_{j\;in\;t}\min(\overline{\Phi}_{ij},0)\geq 0 \tag{170}$$

Condition 170 is close to Condition 154. A striking remark is that the classical predictor-corrector approach would have introduced here a combination of both conditions in the form:

$$volc(i)-\Delta t\sum_{j\;in\;t}\max(\overline{\Phi}_{ij},0)+\Delta t\sum_{j\;in\;t}\min(\overline{\Phi}_{ij},0)\geq 0 \tag{171}$$

It is also:

$$volc(i)-\Delta t\sum_{j\;in\;t}abs(\overline{\Phi}_{ij})\geq 0 \tag{172}$$

We are here less restrictive and it will bring a better behaviour of the scheme for high Courant numbers. We have seen that choosing $volc(i)=S_T\;h^{n+1}_{i\,local}/3$ for the derivative in time leads to monotonicity but behaves poorly. We would like to have instead $volc(i)$ $=S_T\;h^{n+1}_i/3$ but it is potentially unstable. Can we mix both solutions? We can again organise exchanges between triangles. When $h_{i\,local}^{n+1} < h_i^{n+1}$ the point i needs an extra volume $S_T\ \left(h_i^{n+1} - h_{i\,local}^{n+1}\right)/3$ to get the correct derivative in time, without spoiling the local monotonicity. When $h_{i\,local}^{n+1} > h_i^{n+1}$ this is not so obvious, we can only, to avoid negative volumes, go down to a minimum value $h_{i\ \min}^{n+1}$ such that:

$$\frac{S_T\ h_{i\ \min}^{n+1}}{3} = -\Delta t \sum_{j\ in\ t} \min(\overline{\Phi}_{ij}, 0) \tag{173}$$

That is to say we can only give a volume $S_T\ \left(h_{i\,local}^{n+1} - h_{i\ \min}^{n+1}\right)/3$. We have thus for every point again a total offer and a total demand, but with different definition, thus denoted $toc(i)$ and $tdc(i)$ (again c added for "corrector"). If the total offer $toc(i)$ exceeds the total demand $tdc(i)$ we can revert to choosing $S_T\ h_i^{n+1}/3$ everywhere. If not we can share the available extra quantity. The strategy is summarised by using for every point the volume $volc(i)$:

$$volc(i) = \frac{S_T\ h_{i\,local}^{n+1}}{3}$$

$$+\frac{S_T\ \max\left(h_i^{n+1} - h_{i\,local}^{n+1}, 0\right)}{3}\,\frac{\min(toc(i), tdc(i))}{tdc(i)} \tag{174}$$

The total demand is:

$$tdc(i) = \sum_{T\ \ni i} \max\left(\frac{S_T h_i^{n+1}}{3} - \frac{S_T h_{i\,local}^{n+1}}{3}, 0\right) \tag{175}$$

The total offer is:

$$toc(i) = \sum_{T\ \ni i} \frac{S_T\ \left(h_{i\,local}^{n+1} - h_{i\ \min}^{n+1}\right)}{3}$$

$$= \sum_{T\ \ni i} \left[\frac{S_T h_{i\,local}^{n+1}}{3} + \Delta t \sum_{j\ in\ t} \min(\overline{\Phi}_{ij}, 0)\right] \geq 0 \tag{176}$$

This solution leads to the following results with the rotating cone:

corrections	ERIA, cone height	ERIA, standard deviation
0	0.4603	$40.26\ 10^{-3}$
1	0.6982	$18.69\ 10^{-3}$
2	0.7384	$15.58\ 10^{-3}$
3	0.7478	$15.09\ 10^{-3}$
4	0.7516	$15.02\ 10^{-3}$
5	0.7533	$15.01\ 10^{-3}$

Table II: final solution, effect of corrections on numerical diffusion.

The shape of the cone is well preserved (see Figure 21 obtained with 5 corrections). We have a regular convergence, tested up to 12 corrections, where we have a height of 0.756.

Using the PSI fluxes or the N fluxes in the computation of the offer $to(i)$ in the predictor step does not make any difference, so it is simpler to keep the N fluxes.

By construction, the scheme is sensitive to the time step (like NERD and LIPS, but unlike other distributive schemes which organise their own time stepping). The table below compares NERD, ERIA and LIPS at various time steps on the rotating cone test

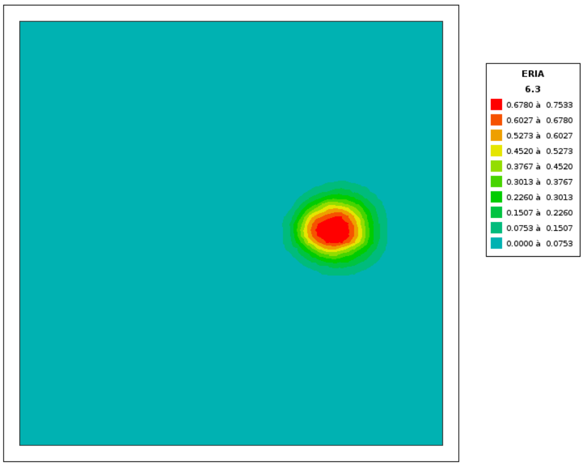

Figure 21: ERIA scheme with 5 corrections. Cone after one rotation

(for the LIPS scheme, no sub-stepping is done). Δt is the basic time step that does a rotation in 32 steps. It is not a convergence study, since the mesh size is unchanged, we only test here the effect of the Courant number. All the tests are done with a fixed number of 5 corrections. The table shows however that NERD and ERIA are at their best with larger Courant numbers (in the table "dev." stands for standard deviation). The maximum height obtained by ERIA with $\Delta t/7$ is shown on Figure 22. Though a bit distorted the shape is better than with the LIPS scheme.

time step	NERD height	NERD dev.	ERIA height	ERIA dev.	LIPS height	LIPS dev.
Δt	0.3920	47.32 10^{-3}	0.7533	15.01 10^{-3}	0.1017	77.05 10^{-3}
$\Delta t/2$	0.3433	50.55 10^{-3}	0.7825	15.51 10^{-3}	0.1460	72.07 10^{-3}
$\Delta t/4$	0.2891	50.21 10^{-3}	0.8004	22.84 10^{-3}	0.3199	58.15 10^{-3}
$\Delta t/6$	0.2282	61.43 10^{-3}	0.8583	31.66 10^{-3}	0.6365	33.90 10^{-3}
$\Delta t/7$	0.2149	63.01 10^{-3}	0.8659	32.52 10^{-3}	0.7554	30.70 10^{-3}
$\Delta t/8$	0.2063	64.04 10^{-3}	0.8479	29.98 10^{-3}	0.8207	32.41 10^{-3}
$\Delta t/16$	0.1830	66.85 10^{-3}	0.8087	19.80 10^{-3}	0.8189	22.66 10^{-3}
$\Delta t/32$	0.1740	67.94 10^{-3}	0.7380	17.13 10^{-3}	0.7427	17.43 10^{-3}

With the original time step and for the different levels of refinement the cone height after one rotation and the standard deviation are given in the following table:

level	cone height, ERIA scheme	standard deviation
0	0.7533	15.01 10^{-3}
1	0.8899	7.64 10^{-3}
2	0.9381	3.12 10^{-3}
3	0.9796	1.40 10^{-3}
4	0.9912	0.734 10^{-3}

Figure 23 shows the result with the bridge piers test case. Monotonicity is obeyed and the relative error on the mass of tracer is 0.38 10^{-15}. It is the best result regarding mass conservation. Again in this case the numerical diffusion is not significantly different from that of the PSI scheme, due to the quasi-steady flow.

Figure 24 is the same test with an extra island, as already explained. There is a blatant progress in terms of numerical diffusion, compared to NERD and LIPS schemes.

Important note:

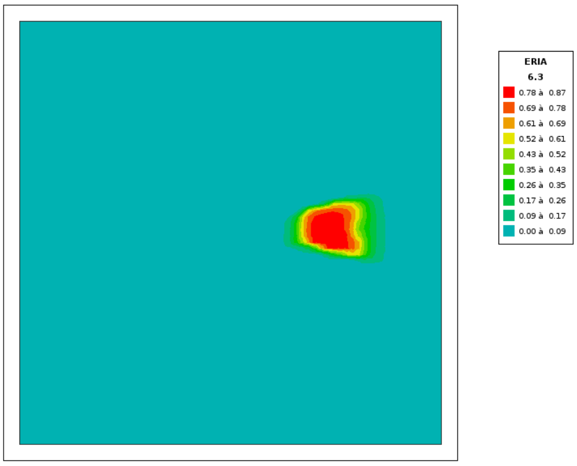

Figure 22: ERIA scheme with 5 corrections. Cone after one rotation, with the time step giving the largest cone height of 0.87.

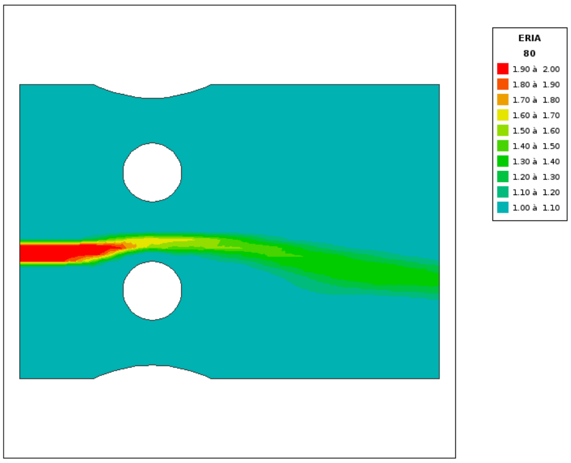

Figure 23: Bridge piers test case. Tracer advected with the ERIA scheme, with one correction.

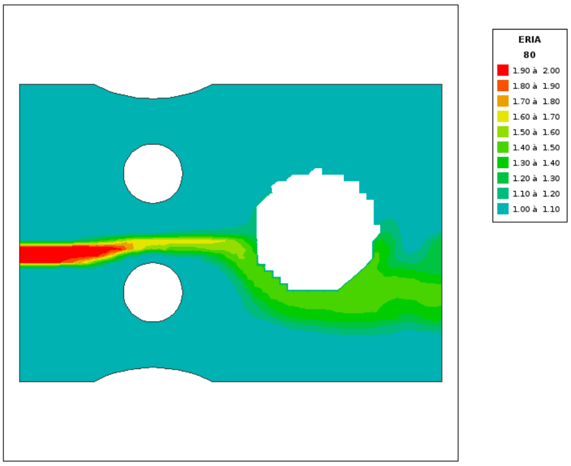

Figure 24: Bridge piers with an island. Tracer advected with the ERIA scheme, with one correction.

To optimise the process of flux transfers in successive iterations, not all the elements are kept. Elements that have transfered all their fluxes are removed from the list. We have thus smaller and smaller loops. However this must be done very carefully: all terms which globally contribute mass must be treated outside the loops. In the case of the corrector it consists in writing it in the form:

$$volc(i)\left(C_{i\ local}^{n+1}-C_i^n\right)=volc(i)\left(C_i^*-C_i^n\right)$$

$$\overleftarrow{-volc(i)(C_i^*-C_i^n)+\Delta t\sum_{j\ in\ t}\min(\overline{\Phi}_{ij},0)\left(C_i^n-C_j^n\right)} \tag{177}$$

so that in the right-hand side the term $volc(i)\left(C_i^*-C_i^n\right)$, being first added and then removed in PSI-reduced form, has no global effect on mass. Assembling all the elements contributions will then give $S_i h_i^{n+1}\left(C_i^{n+1}-C_i^n\right)$.

16.3 Second order in time

16.3.1 Derivation of the scheme

The corrector step is first written in the conservative form:

$$\frac{S_T\ h_{i\ local}^{n+1}}{3}C_{i\ local}^{n+1}=\frac{S_T\ h_i^n}{3}C_i^n$$

$$-\Delta t\sum_{j\ triangle}\min(\overline{\Phi}_{ij},0)\left(\theta_j C_j^*+(1-\theta_j)C_j^n\right)-\Delta t\sum_{j\ triangle}\max(\overline{\Phi}_{ij},0)\left(\theta_i C_i^*+(1-\theta_i)C_i^n\right) \tag{178}$$

where $C_{i\ local}^{n+1}$ will be the new value of tracer at element level and $h_{i\ local}^{n+1}$ is still defined by Equation 147. We open here the possibility of a local θ (at nodal or element level, to be determined later). to anticipate possible exchanges of masses between elements, we write now:

$$volc(i)C_{i\ local}^{n+1}=volp(i)C_i^n$$

$$-\Delta t \sum_{j\ triangle} \min(\overline{\Phi}_{ij},0)\left(\theta_j C_j^* + (1-\theta_j)C_j^n\right) - \Delta t \sum_{j\ triangle} \max(\overline{\Phi}_{ij},0)\left(\theta_i C_i^* + (1-\theta_i)C_i^n\right) \tag{179}$$

with the condition 165 for $volc(i)$ and for $volp(i)$ a similar condition:

$$\sum_{t\ \ni\ i} volp(i) = S_i\ h_i^n \tag{180}$$

The scheme can be put in the form:

$$volc(i) C_{i\ local}^{n+1} = \left(volp(i) - \Delta t \sum_{j\ triangle} \overline{\Phi}_{ij} \right) C_i^n$$

$$-\Delta t \sum_{j\ triangle} \min(\overline{\Phi}_{ij},0)\left(C_j^n - C_i^n\right)$$

$$-\theta_j \Delta t \sum_{j\ triangle} \min(\overline{\Phi}_{ij},0)\left(C_j^* - C_j^n\right) - \theta_i \Delta t \sum_{j\ triangle} \max(\overline{\Phi}_{ij},0)\left(C_i^* - C_i^n\right) \tag{181}$$

but, as we have:

$$\sum_{t\ \ni\ i} \left(volp(i) - \Delta t \sum_{j\ triangle} \overline{\Phi}_{ij} \right) = S_i\ h_i^{n+1} \tag{182}$$

we prefer a slightly different scheme (but still conserving mass) obtained by replacing $volp(i) - \Delta t \sum_{j\ triangle} \overline{\Phi}_{ij}$ with $volc(i)$:

$$volc(i)\left(C_{i\ local}^{n+1} - C_i^n\right) =$$

$$-\Delta t \sum_{j\ triangle} \min(\overline{\Phi}_{ij},0)\left(C_j^n - C_i^n\right)$$

$$-\theta_j \Delta t \sum_{j\ triangle} \min(\overline{\Phi}_{ij},0)\left(C_j^* - C_j^n\right) - \theta_i \Delta t \sum_{j\ triangle} \max(\overline{\Phi}_{ij},0)\left(C_i^* - C_i^n\right) \tag{183}$$

Then the derivative in time is artificially added and removed in the right-hand side, and part of this right-hand side is reduced:

$$volc(i)\left(C_{i\ local}^{n+1} - C_i^n\right) = volc(i)\left(C_i^* - C_i^n\right)$$

$$\overleftarrow{-volc(i)\left(C_i^* - C_i^n\right) - \Delta t \sum_{j\ triangle} \min(\overline{\Phi}_{ij},0)\left(C_j^n - C_i^n\right)} \tag{184}$$

$$\overleftarrow{-\theta_j \Delta t \sum_{j\ triangle} \min(\overline{\Phi}_{ij},0)\left(C_j^* - C_j^n\right) - \theta_i \Delta t \sum_{j\ triangle} \max(\overline{\Phi}_{ij},0)\left(C_i^* - C_i^n\right)}$$

In fact we shall distinguish two kinds of predictor: the one in the derivative in time (kept as C_i^*), and the one in the fluxes (written C_i^{**}), namely we write:

$$volc(i)\left(C_{i\ local}^{n+1}-C_i^n\right)=volc(i)\left(C_i^*-C_i^n\right)$$

$$\overleftarrow{-volc(i)\left(C_i^*-C_i^n\right)-\Delta t\sum_{j\ triangle}\min(\overline{\Phi}_{ij},0)\left(C_j^n-C_i^n\right)} \tag{185}$$

$$\overleftarrow{-\theta_j\Delta t\sum_{j\ triangle}\min(\overline{\Phi}_{ij},0)\left(C_j^{**}-C_j^n\right)-\theta_i\Delta t\sum_{j\ triangle}\max(\overline{\Phi}_{ij},0)\left(C_i^{**}-C_i^n\right)}$$

16.3.2 Proof of monotonicity

If we look at all the coefficients of values of C, the only that could be negative are the coefficients of C_i^* and C_i^n. When the PSI reduction cancels all the terms under the backward arrow we get only $C_{i\ local}^{n+1}=C_i^*$. The other extreme happens when the PSI reduction does nothing. Proving stability in this case will prove it for all cases in between. In this case, we get:

Coefficient of C_i^*:

$$a^*=-\theta_i\Delta t\sum_{j\ triangle}\max(\overline{\Phi}_{ij},0)$$

Coefficient of C_i^n:

$$a^n=volc(i)+\theta_i\Delta t\sum_{j\ triangle}\max(\overline{\Phi}_{ij},0)+\Delta t\sum_{j\ triangle}\min(\overline{\Phi}_{ij},0)$$

$a^*+a^n=volc(i)+\Delta t\sum_{j\ triangle}\min(\overline{\Phi}_{ij},0)$ is positive by construction of $volc(i)$. As already done several times monotonicity will be obtained by imposing that:

$$\left(a^*+a^n\right)C_i^{\min}\leq a^*C_i^{**}+a^nC_i^n\leq\left(a^*+a^n\right)C_i^{\max}$$

The condition is also:

$$\left(a^*+a^n\right)\left(C_i^{\min}-C_i^n\right)+a^*C_i^n\leq a^*C_i^{**}\leq a^*C_i^n+\left(a^*+a^n\right)\left(C_i^{\max}-C_i^n\right)$$

which leads us, as $a^*<0$, to:

$$C_i^n+\frac{volc(i)+\Delta t\sum_{j\ triangle}\min(\overline{\Phi}_{ij},0)}{\theta_i\Delta t\sum_{j\ triangle}\max(\overline{\Phi}_{ij},0)}\left(C_i^n-C_i^{\max}\right)\leq C_i^{**}\leq C_i^n+\frac{volc(i)+\Delta t\sum_{j\ triangle}\min(\overline{\Phi}_{ij},0)}{\theta_i\Delta t\sum_{j\ triangle}\max(\overline{\Phi}_{ij},0)}(C_i^n \tag{186}$$

This requires that the limitation of C_i^{**}, initially equal to C_i^*, be done in the loop on elements, with a different value in every element for a single point i, which is presumably less limiting than a nodal limitation that would have to take into account all neighbouring elements of a point. A similar and equivalent strategy would consist in choosing the local θ_i that relaxes the inequalities 186.

The results are slightly disappointing, they improve only a little on the first order. The rotating cone on the coarser mesh gives a height of 0.7510 and a standard deviation of 14.93 10^{-3}, to be compared with respectively 0.7512 and 15.09 10^{-3}. Similar results, though a bit better, are obtained on the finer mesh : height of 0.9910 and standard deviation of 0.6934 10^{-3}. A possible reason is that with the second order of the predictor-corrector, the division by 2 of the time step (i.e. choosing $k = 2$) had the effect that the predictor was not too much limited. We have here the equivalent of $k = 1$, so little room left for C^* in the proof of monotonicity. However this cannot be the only explanation. It happens that the rotating cone also works if C^* is NOT limited, and in this case the cone height is 0.7371 and the standard deviation becomes 14.54 10^{-3}. Though the shape is slightly better, other errors than error in time obviously interfere, or our approach is not really second order, due to the iteration process within a time step.

17 Adaptation to 3D

Dealing with a 3-dimensional moving grid is formally like dealing with a 2-dimensional grid in depth-averaged context. As a matter of fact the integral of the test-function of a point i is now a volume:

$$V_i = \int_\Omega \Psi_i \, d\Omega \tag{187}$$

This volume is varying in time, as the free surface moves, so we shall have V_i^n and V_i^{n+1}. In 3D V_i will replace $S_i h_i$, integral of the test function multiplied by the depth, which was also a volume. We briefly recall hereafter the gist of the derivation showing that advection in a 2D depth-averaged context and in free surface 3D context are formally alike. The full proofs can be found in Reference [6] and in this paragraph the pages and equation numbers refer to this publication.

The moving mesh raises the problem of relocalisation. As a matter of fact, the new value of a tracer at point i, C_i^{n+1}, does not correspond to the position of point i at time n. The partial derivative in time $\partial C/\partial t$ is no longer an Eulerian derivative. It can be shown (Section 2.2.5 in book) that relocalisation is naturally done if advection is realised in a transformed mesh with fixed coordinates. This is achieved with a generalised sigma transformation, i.e. a sigma transformation in every layer of prisms, the prisms being formed by a superimposition of 2D meshes of triangles. There is one vertical in the mesh per 2D point, and on every vertical we write in every layer:

$$z^* = \frac{z - z_{ip}}{z_{ip+1} - z_{ip}} \tag{188}$$

where z_{ip+1} is the elevation of the upper point on the vertical and z_{ip} is the elevation of the lower point, so that in very layer z^* is in the range [0,1]. $z_{ip+1} - z_{ip}$ is denoted Δz and varies in time, we shall consider thus Δz^n and Δz^{n+1}, linear functions defined everywhere in the mesh. Δz is also $\partial z/\partial z^*$. Horizontal coordinates and horizontal velocities are unchanged. A new vertical velocity is associated to the transformed mesh:

$$W^* = \frac{dz^*}{dt} \tag{189}$$

and the velocity in the transformed mesh, denoted $\overrightarrow{U}^*$ has for components U, V and W^*.

In this fixed mesh the continuity equation becomes (Equation 5.52 page 149 in book):

$$\int_{\Omega^*} \left(\Delta z^{n+1} - \Delta z^n\right) \Psi_i^* d\Omega^* = \Delta t \int_{\Omega^*} \Delta z \overrightarrow{U}^* . \overrightarrow{grad}(\Psi_i^*) \, d\Omega^* - \Delta t \int_{\Gamma^*} \Delta z \overrightarrow{U}^* . \vec{n} \, \Psi_i^* d\Gamma^* \tag{190}$$

where all exponents $*$ refer to the transformed mesh. In this equation $\int_{\Omega^*} \Delta z^n \; \Psi_i^* d\Omega^*$ is the volume associated to point i at time n, denoted V_i^n and in the same way we define V_i^{n+1}. We have in fact (see Section 2.2.5 in book):

$$V_i = \int_{\Omega^*} \Delta z \; \Psi_i^* d\Omega^* = \int_{\Omega} \Psi_i \; d\Omega \tag{191}$$

The boundary term can be denoted b_i as was done in 2D, which leads to, if we add also sources:

$$\frac{V_i^{n+1} - V_i^n}{\Delta t} = \int_{\Omega^*} \Delta z \overrightarrow{U}^* . \overrightarrow{grad}(\Psi_i^*) \, d\Omega^* + Sce_i - b_i \tag{192}$$

This equation is formally like Equation 8, and in fact, when all 3D continuity equations are summed over a vertical, it will give the 2D continuity equation. The terms:

$$-\int_{\Omega^*} \Delta z \overrightarrow{U}^* . \overrightarrow{grad}(\Psi_i^*) \, d\Omega^* \tag{193}$$

are the fluxes of water leaving point i, and we can write as in 2D, V_i replacing $S_i h_i$ as announced:

$$\frac{V_i^{n+1} - V_i^n}{\Delta t} = Sce_i - \sum_j \Phi_{ij} - b_i \tag{194}$$

The conservative tracer equation will be written, like Equation 25 which can be understood as a mere tracer mass-balance:

$$V_i^{n+1} C_i^{n+1} - V_i^n C_i^n = \Delta t \left(Sce_i C_i^{sce} - \sum_j C_{ij} \Phi_{ij} - b_i C_i^{boundary} \right) \tag{195}$$

and from now on it appears obvious that all the derivations done in depth-averaged context can be re-used by changing $S_i h_i$ into V_i.

The only question is: how to find Φ_{ij}, the fluxes between points, once the integrals 193 are known. With triangles we had 3 unknowns and 3 equations, but the fact that all fluxes leaving points at element level summed to zero gave a degree of freedom that was used to minimise numerical diffusion. If we consider that we have now potentially 15 fluxes between points in a prism with 6 points, we have 15 unknowns and 6 equations minus one (the 6 fluxes leaving points summing to zero), which makes ten degrees of freedom!! This will require somewhat heuristic approaches.

17.1 9-flux scheme

We recall in Figure 25 the numbering of points in the prism. We also see 9 segments which will carry fluxes, 3 horizontal at the bottom, 3 horizontal at the top, and 3 vertical. Six other segments are called the "crossed" segments, namely the segments 1-5, 2-4, 2-6, 3-5, 3-4, 1-6. Our first scheme will use only the 9 fluxes of the figure. This is still too many for 6 equations, but the fluxes leaving points can be decomposed into horizontal and vertical components, at element level (hence prism P^* instead of Ω^*):

$$-\int_{P^*} \Delta z \overrightarrow{U}^* . \overrightarrow{grad}(\Psi_i^*) \, dP^* = \Phi_i = a_i + c_i \tag{196}$$

with:

$$a_i = -\int_{P^*} \Delta z \left(U^* \frac{\partial \Psi_i^*}{\partial x} + V^* \frac{\partial \Psi_i^*}{\partial y} \right) dP^* \tag{197}$$

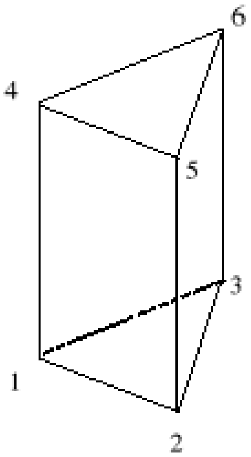

Figure 25: Local numbers of points in a prism.

and:

$$c_i = -\int_{P^*} \Delta z W^* \frac{\partial \Psi_i^*}{\partial z^*} \, dP^* \tag{198}$$

Once all the coefficients a_i and c_i are computed, we decide that the non assembled fluxes in the prism are the following:

For the horizontal segments:

$$\Phi_{ij}^{el} = \max(\min(a_1, -a_2), 0) - \max(\min(a_2, -a_1), 0) \tag{199}$$

i.e. like a N scheme restricted to the bottom triangle, or to the top triangle.

For the vertical segments:

$$\Phi_{i\ i+3}^{el} = \max(b_i, 0) - \max(b_{i+3}, 0) = -\Phi_{i+3\ i}^{el} \tag{200}$$

We recall (book, top of page 192) that:

$$a_1 + a_2 + a_3 = 0 \quad a_4 + a_5 + a_6 = 0 \tag{201}$$

$$b_1 + b_4 = 0 \quad b_2 + b_5 = 0 \quad b_3 + b_6 = 0 \tag{202}$$

Once the fluxes between points are known all the schemes previously described can be derived exactly in the same way, except when we used the fact that we only have a 1-target and a 2-target case. The 9 fluxes can be used as we did for the N scheme, it has been called in Telemac the Leo Postma scheme. In fact these 9 fluxes have long been used as a first step towards the N scheme, and it was Leo Postma who pointed out that he directly used these fluxes to build a finite volume scheme. The drawback is that the scheme does not minimise the numerical diffusion. Suppose that we have $\Phi_{12}^{el} = \Phi_{25}^{el}$ and the others are zero. These two fluxes could be replaced with a single crossed segment flux Φ_{15}^{el}, of the same value, which would cause less numerical diffusion, because the tracer value of point 2 will not interfere. This is called a bypass process, and it is how is built what we shall call the N scheme for prisms.

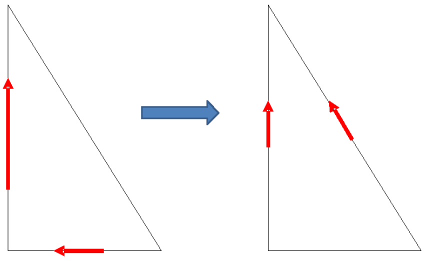

Figure 26: Principle of a bypass of fluxes

17.2 N and PSI schemes

The total contribution of fluxes distributed in a prism is:

$$\Phi_P = \sum_{i=1}^{6}\sum_{j=1}^{6} \min(\Phi_{ij}^{el},0)\left(C_i^n - C_j^n\right) \tag{203}$$

This total contribution is not changed by a bypass, provided that no sign is changed by the process. As a matter of fact if we do a bypass of a value α:

Φ_{ik}^{el} replaced with $\Phi_{ik}^{el}+\alpha$ Φ_{ij}^{el} replaced with $\Phi_{ij}^{el}-\alpha$ Φ_{jk}^{el} replaced with $\Phi_{jk}^{el}-\alpha$

we can check that the coefficients of C_i^n, C_j^n and C_k^n in Φ_P are unchanged. In the 2 coefficients where i appears, it is as the starting point, so these coefficients will be added and their sum is unchanged. It is the same case for point k. On the contrary point j appears as starting point in one case and ending point in the other, the combination $\Phi_{ij}^{el}-\Phi_{jk}^{el}$ will appear in its coefficient and this quantity is also unchanged. Note that the bypass process can be performed considering only the positive fluxes. First the negative fluxes are "forgotten", then positive fluxes are reduced by bypasses, then the negative fluxes are retrieved by doing $\Phi_{ij}^{el}=-\Phi_{ji}^{el}$ when Φ_{ji}^{el} is positive. An example of bypass is shown on Figure 26.

When no more bypass is possible, a point has only fluxes arriving to it or only fluxes leaving it. We have then the equivalent of the N scheme, with points upstream and points downstream clearly defined.

17.3 Differences with 2D schemes

Once the fluxes are known, the derivation and implementation of the schemes is the same than in 2D, except a few details. The problem of fluxes on the same segment but with different directions can no longer be treated in the same way. As a matter of fact, when the assembled value Φ_{ij} is only a sum of two numbers, one of different sign and the other of same sign and larger absolute value than Φ_{ij}, it is easy to cancel the first and to replace the other with Φ_{ij}. With prisms, an assembled flux may be a sum of many more values. This is why a different approach has been chosen, Φ_{ij} is distributed between elements containing the segment, with coefficients in the range [0,1] which sum to 1. The solution is the following: in every element, Φ_{ij} is multiplied by the surface of the triangle on which the prism is based, and divided by the sum of all such surfaces found by the segment in other elements. This sum of surfaces can be done before in an assembly loop.

This solution would probably have to be reviewed if some layers of elements are very thin compared to others (dealing with volumes would then be better but computing volumes is more expensive in time...). We have here a new scheme which is not exactly what has been done in 2D, but is however very close. If we compare 2D and 3D on the rotating cone test, the height after one rotation (level 0, with ten sub-iterations and 5 corrections) is 0.8464 in 2D and 0.8377 in 3D, i.e. a 1% difference.

18 Annex 1: monotonicity of the second-order predictor corrector

Before working on the proof, we need to establish a property on the depth of a point during a time step.

18.1 Property of the depth

The derivation here assumes that the continuity equation is obeyed, it is repeated here for convenience:

$$h_i^n = h_i^{n+1} + \frac{\Delta t}{S_i}\left(-Sce_i + \sum_j \Phi_{ij} + b_i\right) \tag{204}$$

We also assume that the time step is limited by the equivalent conditions 73 and 74:

Whatever the time-step chosen the continuity equation tells us that:

$$h_i^{n+1} + \frac{\Delta t}{S_i}\left(-\max(Sce_i,0) + \min(\sum_j \Phi_{ij},0) + \min(b_i,0)\right) < h_i^n \tag{205}$$

and:

$$h_i^n < h_i^{n+1} + \frac{\Delta t}{S_i}\left(-\min(Sce_i,0) + \max(\sum_j \Phi_{ij},0) + \max(b_i,0)\right) \tag{206}$$

Which gives us immediately:

$$h_i^{n+1}\left(1-\frac{1}{k}\right) \le h_i^n$$

On the other hand (stability conditions with coefficient k):

$$h_i^n \le h_i^{n+1} + \frac{h_i^{n+1}}{k}\frac{\left(-\min(Sce_i,0) + \max(\sum_j \Phi_{ij},0) + \max(b_i,0)\right)}{\left(-\sum_j \min\left(\Phi_{ij}^N,0\right) + \max\left(Sce_i,0\right) - \min\left(b_i,0\right)\right)}$$

or:

$$h_i^n \le h_i^{n+1}\left(1 + \frac{1}{k}\frac{\left(-\min(Sce_i,0) + \max(\sum_j \Phi_{ij},0) + \max(b_i,0)\right)}{\left(-\sum_j \min\left(\Phi_{ij}^N,0\right) + \max\left(Sce_i,0\right) - \min\left(b_i,0\right)\right)}\right)$$

So:

$$h_i^{n+1}\left(1-\frac{1}{k}\right)\le h_i^n\le h_i^{n+1}\left(1+\frac{1}{k}\frac{\Big(-\min(Sce_i,0)+\max(\sum_j \Phi_{ij},0)+\max(b_i,0)\Big)}{\Big(-\sum_j \min\left(\Phi_{ij}^N,0\right)+\max\left(Sce_i,0\right)-\min\left(b_i,0\right)\Big)}\right)$$

As we have (still the continuity equation):

$$h_i^n=h_i^{n+1}+\frac{\Delta t}{S_i}\left(-\min(Sce_i,0)+\sum_j\max(\Phi_{ij},0)+\max(b_i,0)\right)$$
$$-\frac{\Delta t}{S_i}\left(\max(Sce_i,0)-\sum_j\min(\Phi_{ij},0)-\min(b_i,0)\right)\qquad(207)$$

this can be simplified into:

$$h_i^{n+1}\left(1-\frac{1}{k}\right)\le h_i^n\le h_i^{n+1}\left(1+\frac{1}{k}\right)\qquad(208)$$

As a matter of fact:
If $h_i^n\le h_i^{n+1}$it is true that $h_i^n\le h_i^{n+1}\left(1+\frac{1}{k}\right)$
If $h_i^n\ge h_i^{n+1}$ then:

$$\max(Sce_i,0)-\sum_j\min(\Phi_{ij},0)-\min(b_i,0)>-\min(Sce_i,0)+\sum_j\max(\Phi_{ij},0)+\max(b_i,0)\qquad(209)$$

then we have also $h_i^n\le h_i^{n+1}\left(1+\frac{1}{k}\right)$.
As we have: $h_i^{n+1-\theta}=(1-\theta)\,h_i^{n+1}+\theta h_i^n$, we can also deduce that:

$$h_i^{n+1}\left(1-\frac{\theta}{k}\right)\le h_i^{n+1-\theta}\le h_i^{n+1}\left(1+\frac{\theta}{k}\right)\qquad(210)$$

18.2 Proof of monotonicity

We write the corrector in the following way, as already done:

$$S_ih_i^{n+1}C_i^{n+1}=S_ih_i^{n+1}C_i^*-f_iS_ih_i^{n+1-\theta}\left(C_i^*-C_i^n\right)$$

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

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

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

We do not consider the predictor, so that the choice of C_i^* remains free to enable multiple corrections. 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 C in the right hand side. Only the coefficients of C_i^* and C_i^n are questionable. They are:

Coefficient of C_i^*:

$$S_i h_i^{n+1} - f_i S_i h_i^{n+1-\theta} + \theta \Delta t \sum_j \mu_j \min(\Phi_{ij}, 0) - \theta \Delta t \left(\max(Sce_i, 0) \;\; - \min(b_i, 0)\right) = a^*$$

Coefficient of C_i^n:

$$f_i S_i h_i^{n+1-\theta} + (1-\theta)\Delta t \sum_j \mu_j \min(\Phi_{ij}, 0) - (1-\theta)\,\Delta t \left(\max(Sce_i, 0) \;\; - \min(b_i, 0)\right) = a^n$$

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:

$$a^* + a^n = S_i h_i^{n+1} + \Delta t \sum_j \mu_j \min(\Phi_{ij}, 0) + \Delta t \left[-\max(Sce_i, 0) \;\; + \min(b_i, 0)\right]$$

As a matter of fact, we can take $\mu_j = 1$ (worst case), and we fall back to the classical stability condition of the N and PSI scheme.

We write:

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

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

with α and β in the range [0,1], we want to find at which condition we would have:

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

with $C_i^{average}$ obeying the maximum principle, i.e. $C_i^{average} = C^{\min} + \gamma \left(C^{\max} - C^{\min}\right)$, and γ in the range [0,1]. We get:

$$\gamma = \frac{\alpha a^* + \beta a^n}{a^* + a^n}$$

We must thus ensure that:

$$0 \le \alpha a^* + \beta a^n \le a^* + a^n$$

γ will be positive if: $\alpha a^* + \beta a^n \ge 0$

γ will be less than 1 if: $\alpha a^* + \beta a^n \le a^* + a^n$, i.e. if $(1-\alpha)\, a^* + (1-\beta)\, a^n \ge 0$.

So we have to find a condition on C_i^*, i.e. on α depending on β and then we shall have the same condition for $(1-\alpha)$ depending on $(1-\beta)$. It is the same problem. Only the positivity of γ will then be studied. We are sure that γ will be positive if:

$$\alpha S_i h_i^{n+1} + (\beta - \alpha)\, f_i S_i h_i^{n+1-\theta}$$

$$+ \left[\alpha\theta + \beta\,(1-\theta)\right] \left(\Delta t \sum_j \min(\Phi_{ij}, 0) - \Delta t \left(\max(Sce_i, 0) \;\; - \min(b_i, 0)\right)\right) \ge 0$$

We now assume that the time step was chosen with the condition:

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

This is the classical condition for the N scheme, divided by k. The positivity will be thus ensured if:

$$\alpha S_i h_i^{n+1} + (\beta - \alpha) f_i S_i h_i^{n+1-\theta} \geq [\alpha\theta + \beta(1-\theta)] \frac{S_i h_i^{n+1}}{k}$$

If $\alpha \leq \beta$ the worst case happens when $f_i = 0$ and we must have:

$$\alpha \geq \frac{1-\theta}{k-\theta}\beta$$

If $\alpha \geq \beta$ the worst case happens when $f_i = 1$ and we must have:

$$\alpha S_i h_i^{n+1} \geq [\alpha\theta + \beta(1-\theta)] \frac{S_i h_i^{n+1}}{k} + (\alpha - \beta) S_i h_i^{n+1-\theta}$$

We can use the property (see Inequality 210 in Annex 1):

$$h_i^{n+1}\left(1 - \frac{\theta}{k}\right) \leq h_i^{n+1-\theta} \leq h_i^{n+1}\left(1 + \frac{\theta}{k}\right)$$

and we get a stronger condition if we replace $h_i^{n+1-\theta}$ by $h_i^{n+1}\left(1 + \frac{\theta}{k}\right)$:

$$\alpha \geq [\alpha\theta + \beta(1-\theta)] \frac{1}{k} + (\alpha - \beta)\left(1 + \frac{\theta}{k}\right)$$

which is:

$$\alpha \leq \beta \frac{(k + 2\theta - 1)}{2\theta}$$

and we arrive at:

$$\beta\left(1 - \frac{(k-1)}{k-\theta}\right) \leq \alpha \leq \beta\left(1 + \frac{(k-1)}{2\theta}\right)$$

The condition for $\gamma \leq 1$ will give in the same way:

$$(1-\beta)\left(1 - \frac{(k-1)}{k-\theta}\right) \leq 1 - \alpha \leq (1-\beta)\left(1 + \frac{(k-1)}{2\theta}\right)$$

or:

$$1 + (\beta - 1)\left(1 + \frac{(k-1)}{2\theta}\right) \leq \alpha \leq 1 + \ (\beta - 1)\left(1 - \frac{(k-1)}{k-\theta}\right)$$

We arrive thus at two conditions:

$$\beta\left(1 - \frac{(k-1)}{k-\theta}\right) \leq \alpha \leq \beta\left(1 + \frac{(k-1)}{2\theta}\right)$$

$$1 + (\beta - 1)\left(1 + \frac{(k-1)}{2\theta}\right) \leq \alpha \leq 1 + \ (\beta - 1)\left(1 - \frac{(k-1)}{k-\theta}\right)$$

We see that if k tends to infinity, the choice of α becomes wider and tends to: $0 \leq \alpha \leq 1$. Except if $\beta = 0$ which will be the problem, as shown hereafter. We know from the predictor step that:

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

which corresponds to:

$$\left(1-\frac{1}{k}\right)\beta<\alpha<\frac{1}{k}+\left(1-\frac{1}{k}\right)\beta$$

If k tends to infinity, the range narrows to β.

To find a value of k that enables the stability of our semi-implicit scheme, we must find k such that, whatever β in the range [0,1]:

$$\frac{1}{k}+\left(1-\frac{1}{k}\right)\beta\leq\beta\left(1+\frac{(k-1)}{2\theta}\right)$$

$$\frac{1}{k}+\left(1-\frac{1}{k}\right)\beta\leq 1+\ (\beta-1)\left(1-\frac{(k-1)}{k-\theta}\right)$$

$$\left(1-\frac{1}{k}\right)\beta\geq\ \beta\left(1-\frac{(k-1)}{k-\theta}\right)$$

$$\left(1-\frac{1}{k}\right)\beta\geq 1+(\beta-1)\left(1+\frac{(k-1)}{2\theta}\right)$$

As a matter of fact the stability condition used for the predictor must be stricter than 4 limitations necessary for the predictor.

The condition number 1 is:

$$\frac{1}{k}\leq\frac{\beta\,(k-1)}{2\theta}+\frac{\beta}{k}$$

It cannot be satisfied with $\beta=0$!

The condition number 2 is:

$$\frac{1}{k}\leq\frac{(k-1)}{k-\theta}$$

which imposes the condition: $k\geq 1+\sqrt{1-\theta}$, which is always lower than 2, it is not really a constraint.

The condition number 3 is equivalent to condition number 2.

The condition number 4 is:

$$\frac{\beta}{k}\leq(1-\beta)\,\frac{(k-1)}{2\theta}$$

It cannot be satisfied with $\beta=1$!

Conclusion: there is no stability condition possible on the predictor that ensures the stability of the corrector. The reason is that when C_i^n is close to $C^{\min}$ or $C^{\max}$, C_i^* needs to be so close to C_i^n that no stability condition can achieve it. Thus only a correction of C_i^* can be envisaged. It will simply impose that $C_i^*=C^{\min}$ when $C_i^n=C^{\min}$, and $C_i^*=C^{\max}$ when $C_i^n=C^{\max}$.

We must now translate our conditions into limitations of C_i^*:

$$\beta\left(1-\frac{(k-1)}{k-\theta}\right)\leq\alpha\leq\beta\left(1+\frac{(k-1)}{2\theta}\right)$$

$$1+(\beta-1)\left(1+\frac{(k-1)}{2\theta}\right)\leq\alpha\leq 1+\ (\beta-1)\left(1-\frac{(k-1)}{k-\theta}\right)$$

We can swap the conditions to get:

$$1+(\beta-1)\left(1+\frac{(k-1)}{2\theta}\right)\leq\alpha\leq\beta\left(1+\frac{(k-1)}{2\theta}\right)$$

$$\beta\left(1-\frac{(k-1)}{k-\theta}\right)\leq\alpha\leq 1+\ (\beta-1)\left(1-\frac{(k-1)}{k-\theta}\right)$$

Now written in the form:

$$1+(\beta-1)\,\frac{(k-1+2\theta)}{2\theta}\leq\alpha\leq\beta\frac{(k-1+2\theta)}{2\theta}$$

$$\beta\left(\frac{1-\theta}{k-\theta}\right)\leq\alpha\leq 1+\ (\beta-1)\left(\frac{1-\theta}{k-\theta}\right)$$

The first one leads to:

$$C_i^n+\frac{k-1}{2\theta}\left(C_i^n-C^{\max}\right)\leq C_i^*\leq C_i^n+\frac{k-1}{2\theta}\left(C_i^n-C^{\min}\right)$$

It is this condition that cannot always be satisfied, whatever the choice of k and θ, without limiting C_i^*.

The second one leads to:

$$C_i^n\left(\frac{1-\theta}{k-\theta}\right)+C^{\min}\left(1-\frac{1-\theta}{k-\theta}\right)\leq C_i^*\leq C_i^n\left(\frac{1-\theta}{k-\theta}\right)\ +C^{\max}\left(1-\frac{1-\theta}{k-\theta}\right)$$

and is ensured by the PSI predictor for every value of $\theta\geq 0$ as soon as $k\geq 1$. It can also be written:

$$C_i^n+\left(C^{\min}-C_i^n\right)\frac{k-1}{k-\theta}\leq C_i^*\leq C_i^n\ +\left(C^{\max}-C_i^n\right)\frac{k-1}{k-\theta}$$

where we see that this condition vanishes if $\theta=1$.

If we make the reasonable choice $k=2$ and $\theta=\frac{1}{2}$. It gives:

$$\frac{\beta}{3}\leq\alpha\leq 2\beta$$

$$2\beta-1\leq\alpha\leq 1+\ \frac{(\beta-1)}{3}$$

Which for symmetry reason we rather combine in the form:

$$2\beta-1\leq\alpha\leq 2\beta$$

$$\frac{\beta}{3}\leq\alpha\leq\frac{2}{3}+\ \frac{\beta}{3}$$

This is equivalent to:

$$2C_i^n-C^{\max}\leq C_i^*\leq 2C_i^n-C^{\min}$$

$$\frac{2C^{\min}}{3}+\frac{C_i^n}{3}\leq C_i^*\leq\frac{2C^{\max}}{3}+\frac{C_i^n}{3}$$

References

[1] GODUNOV S.: A Difference Scheme for Numerical Solution of Hydrodynamics Equations. Math. Sbornik, 47, 271-306. 1959.

[2] ROE P.L.: Linear advection schemes on triangular meshes. Technical report coa 8720. Cranfield Institute of Technology. 1987.

[3] Struijs R.: A Multi-Dimensional Upwind Discretization Method for the Euler Equations on Unstructured Grids. PhD thesis, University of Delft, Netherlands, 1994.

[4] POSTMA L., HERVOUET J.-M.: Compatibility between finite volmes and finite elements in solutions of Shallow Water and Navier−Stokes equation. International Journal for Numerical Methods in Fluids, 2002.

[5] 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.

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

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

[8] SLINGERLAND P.: An accurate and robust finite volume method for the advection diffusion equation. Thesis for a Master of Science in applied mathematics. Delft Unversity of Technology. 2007

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

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

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

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

[13] HERVOUET J.-M., RAZAFINDRAKOTO E., VILLARET C.: Dealing with dry zones in free surface flows: a new class of advection schemes. Proceedings of the AIRH congress, Brisbane, June 2011.

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

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

[16] RICCHIUTO M, An explicit residual based approach for shallow water flows. Journal of Computational Physics. 280:306:344. 2015.

[17] HERVOUET J.-M.: The weak form of the method of characteristics, an amazing advection scheme. Proceedings of the 20th Telemac User Club, BAW, Karlsruhe, Germany, 16-18 October 2013.

[18] HERVOUET J.-M., PAVAN S., ATA R.: Ongoing research on advection schemes. Proceedings of the 21^{st} Telemac User Club, Artelia,Grenoble, France, 15-17 October 2014.

[19] HERVOUET J.-M., PAVAN S., ATA R.: Distributive advection schemes and dry zones, new solutions. Proceedings of the 22^{nd} Telemac User Club, STFC Daresbury, UK, 13-16 October 2015.

[20] PAVAN S.: New advection schemes for free surface flows. Thèse présentée pour l'obtention du grade de docteur de l'Université Paris-Est. 15 February 2016.

[21] PAVAN S., ATA R, HERVOUET J-M: Finite volume schemes and residual distribution schemes for pollutant transport on unstructured grids. Environ. Earth. Sci. 74:7337?7356, 2015

[22] PAVAN S., HERVOUET J-M, RICCHIUTO M, ATA R, : A second order residual based predictor?corrector approach for time dependent pollutant transport, J. Comput. Phys. 318:122-141, 2016

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

RESEARCH CENTRE
BORDEAUX – SUD-OUEST

351, Cours de la Libération
Bâtiment A 29
33405 Talence Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399