Ecole Prédoctorale de Physique des Houches

26 août - 6 septembre 2002

Numerical methods for air quality modelling with application to one model (CHIMERE).

J. Roux
Laboratoire de Météorologie Dynamique, Ecole Polytechnique, 91128 Palaiseau Cedex
e-mail : jroux@lmd.polytechnique.fr
T. 01 69 33 36 13

1 Introduction

In this paper about air pollution modelling we emphasize mainly for models on air quality forecasting which is a challenging scientific problem, which has recently undergone considerable pressure in many industrialised countries due to the increasing consciousness of the effect, on health and environment, of the emissions of pollutants, especially volatile organic compounds (VOC's) and nitrogen oxides (NO_x) into the urban atmospheres. The environmental benefits of reliable air quality forecasts are obvious : populations can be more efficiently protected by means of information or real-time emission abatement strategies. The scientific aspects of air-quality forecasting are also of major scientific interest, since prediction lies at the end of the understanding process. Here, in this presentation we insist about models based on the physical equations (often called "deterministic models") driving the chimistry and the transport of pollutants. These models require numerous accurate input data (emissions, meteorology, input cover) which are difficult to collect in real time. These latter problems should become less and less significant with time, due to both increase in computer capabilities and to improvements in the data bases required for the deterministic approach of forecasting. Evidently the methods presented here are generic and valid for any problem related to air quality modelling.

Air pollution modelling begins always by modelling atmospheric chemical kinetics systems which describe chemical reactions between trace gases, such as ozone, nitrogen oxides, methane, hydrocarbons, etc. One often studies ozone levels in the lower atmosphere, as ozone is dangerous for humans and animals during short term smog episodes and can damage crops when over longer seasonals periods levels are too high. Ozone is also a greenhouse gas, similar as methane, carbon dioxide and other species. Air pollution models are therefore also used in connection with climate studies. Ozone itself is not emitted but formed in very many different reactions. A nice introduction to the field of atmospheric chemistry can be found in Graedel and Crutzen [17]. First in this paper we present in the next section a very simple illustration of chemical mechanism for any tropospheric air pollution model.

Air pollution modelling is based on the assumption of no reciprocal effect of the chemical species on flow fields (wind velocity, turbulent diffusivity, temperature). After having preprocessed the flow fields by meteorological computations or parametrizations, ones solves thereafter a reaction-advection-diffusion PDE (Partial Differential Equation), the mass continuity equation, for the concentrations of chemical species :

$$\frac{\partial \vec{C}}{\partial t} + div(\vec{V}\vec{C}) = div(\mathbf{K}\vec{C}) + P(\vec{C},t) - L(\vec{C},t). \tag{1.1}$$

In this equation, characteristic for the Eulerian approach, $\vec{C}$ is a vector containing all the concentrations of all model species depending of x and t, $\vec{V}$ is the three dimensional wind vector, $\mathbf{K}$ is the turbulent diffusive matrix, and P and L represent production and loss terms due to chemical reactions, emissions and depositions. The wind field $\vec{V}$ and the diffusion coefficient matrix $\mathbf{K}$ are given, so that the problem is linear with respect to the transport part ; $\vec{C}$ represents a vector of m concentrations of trace gases like ozone, methane, nitrogen oxides, hydrocarbons, radicals, etc. The chemical reactions between these species are of first of second order, so that the contribution of the chemical mechanism in terms P and L is quadratically non-linear (it is by this mechanism that P and L depend on, for a specific concentration specie c, the concentrations of other constituants). Without the chemical mechanism the m equations (1.1) are uncoupled.

Solution methods often used in air pollution modelling follow the method of lines approach. That's means the PDE system with its boundary conditions is first discretized in space on a three-dimensional Eulerian grid, resulting in a huge continuous-time integration Ordinary Differential

Equation (ODE) system. On the discretized space grid the system of ODE's is then solved in time by a numerical integration technique. That is this approach than we principally propose in this paper.

The major transport mechanism in air pollution models is advective transport in a wind field. Hence the quality of the advection computation is crucial. In section 3 we consider, for simplicity of presentation, the modelization of the scalar advection problem in one space dimension. First we present a survey of finite difference methods with an introduction of concepts, definitions and tools for the field of the theory, then we propose several finite volume methods in the Godunov framework, as pioneered by Godunov, van Leer and others (see for instance [23]). The modelling of the 3D advection problem is presented in the section 4 with a short description of the splitting methods.

The wide range of the chemical timescale ($10^{-9}s$ for excited oxygen radical O^{1D} to be compared with several months for methane) induces the well-known stiffness of the resulting equations as seen in section 2. The time integration of such PDE's (and the resulting discretized ODE's) has then to be performed carefully if a (rather) low CPU cost is required. Furthermore implicit numerical schemes are highly recommended even if they are however associated with CPU costs which can remain large (due to inversions of matrices whose dimension is large : currently the number m of constituants is about one hundred). Due to the dimension and stiffness of the ODE's system together which the great number of cells in 3D, ODE integrators based on standard numerical routines are not feasible in a operational way.

This has led to the use of special time integration techniques dedicated to this problem.

We present in section 5 a survey of numerical integration techniques for ODEs, introducing the usual definitions and notions (the essential notion of stiffness). We show that implicit methods are convenient for stiff problems (the chemical kinetic system for pollution is stiff). We begin by standard onestep methods : the explicit Euler method (the forward Euler method) and the implicit Euler method (the backward Euler method) by which we introduce consistency (and order), stability (in particular the important A-stability definition) and convergence notions. We carry on multisteps methods with order greater than one : Adams methods and Backward Differentiation Formula (BDF method). The more robust method is the BDF2 method : it is the optimal order for multisteps methods (the order two) which hold the convenient A-stability property.

Due to the production-loss form of the ODE's (taking into account the discretization of the PDE's) descended from the chemical mechanism, specific methods have been dedicated. The well-known QSSA method, long ago already advocated by Hessvedt and co-workers [21], is a traditionnal tool for modelling the chemistry, but its use ask sometimes a know-how not very easy to learn and justify ; in order to be efficient the scheme must be heavily tuned by "lumping". An other way is to use a BDF method (implicit by construction and well-suited for stiff problems) and adapt it to the special form of the chemistry modelling, for instance the BDF2 method can be write in a particular form ; the resulting implicit scheme can be solved by a Gauss-Seidel process. These algorithm (BDF2 applied to production-loss form, with resolution of the implicit scheme by Gauss-Seidel), is proposed by Verwer [38] like the TWOSTEP method. This TWOSTEP method is more or less comparable to the QSSA method. The TWOSTEP and QSSA methods are presented in the subsection 5.3.

Finally in a last section we present one example of the modelling of the air pollution problem : the CHIMERE software. By that we precise one possible chemical mechanism, together with the emissions and dry depositions modelling. The advection, diffusion mechanisms and boundary layer (where essentially the pollution exits, furthermore it is the domain of interest for health and environment) modelling are also described with any details. Taking into account the advection and diffusion operators together with the chemical mechanisms (included emissions and deposition), the resulting ODE's system is solved by TWOSTEP.

Evidently a good approach is the search for efficient numerical solvers with a good compromise between accuracy and CPU requirements. We hope that this presentation give any possibilities for that.

2 A fundamental chemical kinetics system

Consider the following reactions which are basic to any tropospheric air pollution model :

$$\begin{aligned} NO_2 + h\nu &\xrightarrow{k_1} NO + O(^{3P}) \\ O(^{3P}) + O_2 &\xrightarrow{k_2} O_3 \\ NO + O_3 &\xrightarrow{k_3} O_2 + NO_2 \end{aligned} \tag{2.1}$$

In this system $O(^{3P})$ is the oxygen atom in its fundamental state. The second equation is the single way for formation of ozone. The first equation is the principal source for $O(^{3P})$ by photolyze of nitrogen dioxide (NO_2) in the troposphere. Finally the ozone formation is balanced simultaneously by its reaction with nitrogen monoxyde (NO) with formation of NO_2. We note that the ozone formation is associated to its principal trap NO. In normal situation NO, NO_2 and $O(^{3P})$ are in equilibrium and the ozone concentration is moderated. The formation of high ozone concentrations require therefore a mecanism which consume quickly NO and regenerate NO_2. In pollutant environment this is realized by coupling $NO|NO_2$ cycle with weathering hydrocarbons (HC) : attacked by a hydroxyl radical (OH) the HC undergoes at alkyl radical (R) which, by a sequence of quick reactions with air-oxygen, produce peroxy radical (RO_2) and hydroperoxyl radical (HO_2) whom, in turn, quickly oxydate NO to NO_2 with radical OH regeneration. These oxydation reaction from NO to NO_2 short-circuit those for NO by O_3 (see (2.1)) and ozone can accumulate it.

The motor of this complex chemistry is therefore the photolysis of any species (O_3, NO_2, etc.). By which it is during intensive solar irradiation that high levels of ozone can be met. But this condition is not sufficient : it is also necessary that the precursors sources let be sufficiently important, especially in HC and that the atmospheric dynamic favor the accumulation for precursors and products with a weak dispersion and dilution. Therefore it is during anticyclonique period that pollution episode are the more frequent.

An other good reference for the chemistry of the troposphere can be found in Seinfeld and Pandis [29].

It is useful for illustration to show the associated ODE system. Putting $c_1 = [O(^{3P})]$, $c_2 = [NO]$, $c_3 = [NO_2]$, $c_4 = [O_3]$, the dynamic of the species in (2.1) is governed by :

$$\begin{aligned} \dot{c_1} &= k_1c_3 - k_2c_1 \\ \dot{c_2} &= k_1c_3 - k_3c_2c_4 \\ \dot{c_3} &= k_3c_2c_4 - k_1c_3 \\ \dot{c_4} &= k_2c_1 - k_3c_2c_4 \end{aligned}$$

The unit for time is seconds and for concentrations number of molecules per cm^3. Note that oxygen is taken constant, which is a natural assumption. Typically, the first of the reactions is photochemical, giving rise to rapid changes in concentrations values in sunset and sunrise.

3 Modelling of linear advection in 1D.

For simplicity of presentation we start with the scalar advection problem in one space dimension. We consider the transport of one pollutant, for instance the ozone. We note $a = [O_3]$ the concentration of ozone. The advection in one space dimension of this pollutant, during the interval [0,T[is given by the following **linear** hyperbolic equation (wrote in the **conservative** form), called also transport equation, to which we add an initial concentration ; the overall cauchy problem is consequently the following one.

Given a field of initial concentrations a_0, a velocity wind u and a time $T > 0$, we want to calculate $a(x,t)$ such that :

$$\left\{\begin{array}{rclcl} \frac{\partial a}{\partial t} + \frac{\partial (au)}{\partial x} & = & 0 & \forall (x,t) & \in \mathbb{R} \times [0,T[\\ a(x,0) & = & a_0(x) & \forall x & \in \mathbb{R} \end{array}\right. \tag{3.1}$$

Evidently the pratical resolution of this equation is made on a finite domain of $\mathbb{R}$. The fact that (3.1) is linear in space implies that the transport does not involve more irregularity that the initial condition. In the following we present the two numerical methods which are using to solve (3.1) : the finite difference method and the finite volume method.

3.1 Finite difference methods

3.1.1 Generalities

We present rapidly the concepts of the finite difference method. To simplify this presentation we consider in all this section 3.1 that **u is a constant given data**, then the partial derivative equation in (3.1) becomes

$$\frac{\partial a}{\partial t} + u\frac{\partial a}{\partial x} = 0 \qquad \forall (x,t) \in \mathbb{R} \times [0,T[\tag{3.2}$$

we suppose also that the initial data a_0 is sufficiently regular and that a_0 is periodic with period L. This periodicity hypothesis is not crucial but it allows to limit in practice the numerical calculation of (3.1) to a bounded interval. We know that the exact solution of the problem (3.1) is given by $a(x,t) = a_0(x - ut)$ (given by the characteristic method (see one next paragraph) in this linear case).

Definition of a regular mesh. We introduce now a regular mesh of the domain [0,L] x $\mathbb{R}^+$ with steps Δt and Δx respectivly in space and time (see figure 1), with $\Delta x = \frac{L}{M}$, the integer M defines the number of nodes of the mesh in the interval]0,L[. For $j \in \{1, 2, ..., M\}$ and $n \in \mathcal{N}$, we note $x_j = j\Delta x$, $t^n = n\Delta t$ and a_i^n an approximation of the true solution $a(x_i, t^n)$ at point x_i and at time t^n with similar approximations for all others index i and exponent n.

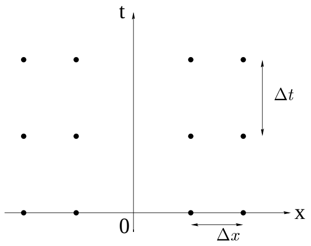


FIG. 1 – Discrete network of the points $(j\Delta x, n\Delta t)$ ($j \in \mathcal{Z}$ and $n \in \mathcal{N}$) for finite difference approximation.

Characteristic method. For the advection equation (3.1) the characteristic method shows us that the solution is constant on all straight line $x - ut = Constant$, because the exact solution is given by $a(x,t) = a_0(x - ut)$. The solution at any point $(\bar{x}, \bar{t})$ depends on the initial data a_0 only at a single point, namely the point $\bar{x_0} = \bar{x} - u\bar{t}$ such that the point $(\bar{x}, \bar{t})$ lies on the characteristic from $\bar{x_0}$. We could change the initial data at any point other than $\bar{x_0}$ without affecting the solution $a(\bar{x}, \bar{t})$. The set $\mathcal{D}(\bar{x}, \bar{t}) = \{\bar{x_0}\}$ is called the **domain of dependance** of the point $(\bar{x}, \bar{t})$. Here this domain consists of the single point $\bar{x_0}$.

In particular we have for the exact solution of (3.1)

$$a(x_i, t^{n+1}) = a(x_i - u\Delta t, t^n) \tag{3.3}$$

for all $i \in \mathcal{Z}$ and $n \in \mathcal{N}$.

We are going to see that (3.3) allows to build any numerical method. First define the *Courant number*

$$\nu = u\frac{\Delta t}{\Delta x} \tag{3.4}$$

Consider the characteristic passing through the point $(x_i,\ t^{n+1})$. The integer part of $|\nu|$ is equal to the number of meshes this characteristic meet between times t^n and t^{n+1}.

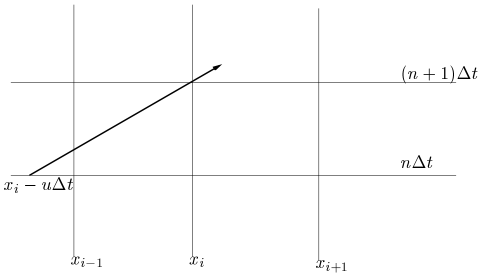


FIG. 2 – Characteristic passing through the point $(x_i,\ t^{n+1})$.

More precisely, if ν is an integer, the point $(x_i - u\Delta t, t^n)$ is a mesh point : then by noting $\nu = k \in \mathbb{Z}$, one has $x_i - u\Delta t = x_{i-k}$ and the exact solution of (3.1) satisfies

$$a(x_i, t^{n+1}) = a(x_{i-k}, t^n), \forall i \in \mathbb{Z}, \forall n \in \mathbb{N} \tag{3.5}$$

and the numerical scheme $a_i^{n+1} = a_{i-k}^n$ is then exact !

More generally when the Courant number ν is not an integer, let be k the single relative integer such that $k \leq \nu \leq k+1$. The point $x_i - u\Delta t$ is then located in the $[x_{i-k-1}, x_{i-k}]$ interval ; more precisely, one checks easily that one has :

$$x_i - u\Delta t = (1 - (\nu - k))x_{i-k} + (\nu - k)x_{i-k-1} \tag{3.6}$$

We can then compute an approached value of $a(x_i - u\Delta t, t^n)$ by a linear interpolation in the $[x_{i-k-1}, x_{i-k}]$ interval :

$$a(x_i - u\Delta t, t^n) \approx (1 - (\nu - k))a(x_{i-k}, t^n) + (\nu - k)a(x_{i-k-1}, t^n) \tag{3.7}$$

Thanks to (3.3), the approached relation (3.7) suggests to consider the numerical scheme

$$a_i^{n+1} = (1 - (\nu - k))a_{i-k}^n + (\nu - k)a_{i-k-1}^n \tag{3.8}$$

for $k \leq \nu = u\frac{\Delta t}{\Delta x} < k + 1, k \in \mathbb{Z}$, (3.8) is a linear interpolation formula.

We call (3.8) the *characteristic scheme with linear interpolation on a mesh.*

Now suppose $u \geq 0$ and $\nu < 1$, by the previous definition of k, $k = 0$. Then the equation (3.8) becomes :

$$a_i^{n+1} = (1-\nu)a_i^n + \nu a_{i-1}^n \tag{3.9}$$

by definition of ν. By that we obtain, taking into account the definition of ν, a very simple finite difference scheme called the *left one-sided scheme* (see figure 3) :

$$a_i^{n+1} = a_i^n - u\frac{\Delta t}{\Delta x}\left(a_i^n - a_{i-1}^n\right) \tag{3.10}$$

N.B. : In fact it is possible to exhibit naively a formula like (3.10). Perhaps a condition like $u \geq 0$, by physical consideration seems a natural one. But it is not so natural to imagine a condition like $\nu < 1$. It is not necessary to suppose a priori this condition for write (3.10), and we can thing that this condition is superfluous. In fact it is necessary because it is under this condition that this scheme is efficient (see sous-paragraph : The CFL condition).•

This method is called **explicit** : we obtain directly the discrete solution without inversing, for instance, a matrix. We could think that it is useless to develop a numerical method for a problem whose solution is known but it is particularly useful to appreciate the numerical solution in respect to the true solution and to prove convergence results before any generalization of this method for more complex problems whose the exact solution is unknown.

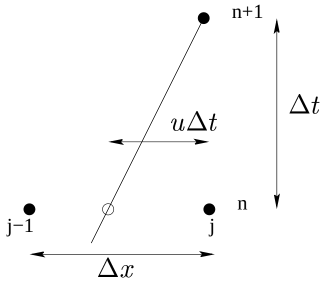


FIG. 3 – Left one-sided scheme. One interpolate the value, at the feet of the characteristic, with help of know values at the two neighbouring points of the grid.

.

The CFL condition. Courant, Friedrichs and Lewy recognized, in 1928, that a necessary stability condition for any numerical method is that the domain of dependance of the finite difference method should include the domain of dependance of the advection equation (it is true for other certain Partial Differential Equations (PDEs)), at least in the limit as $\Delta t, \Delta x \to 0$. This condition is known as the CFL condition after Courant, Friedrichs and Lewy.

For the left one-sided scheme with $0 < \nu < 1$ we know the domain of dependance. The value a_i^n depends on values a_i^{n-1} and a_{i-1}^{n-1} at time t^{n-1}. These values depend on a_i^{n-2}, a_{i-1}^{n-2} and a_{i-2}^{n-2} at time t^{n-2}. Continuing back to $t = 0$ we see that the solution at t^n depends only on initial data at the points x_{i+q} for $q = -n, ..., 0$, so that it is possible to exhibit (see figure 4) the numerical dependance cone for the left on-sided scheme (3.10), under the condition

$$0 \leq u\frac{\Delta t}{\Delta x} \leq 1. \tag{3.11}$$

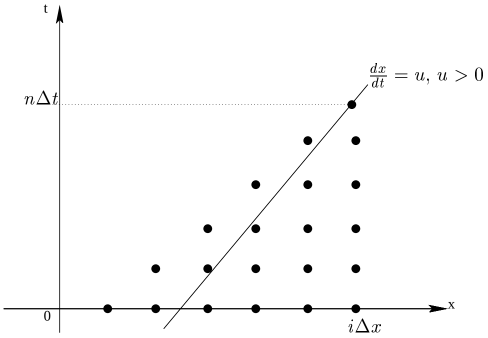


FIG. 4 – Numerical dependance cone for the left one-sided scheme (3.10). The stability condition $0 \leq u\frac{\Delta t}{\Delta x} \leq 1$ means that the characteristic line passing through $(i\Delta x,\ n\Delta t)$ is **inside this numerical dependance cone**.

The CFL condition is only a necessary condition for stability, not sufficient (see subsection 7.0.2 for the unstability of the centered scheme). We shall see also in subsection 7.0.2 that the CFL condition (3.11) is sufficient for stability for the left one-sided scheme.

Discrete maximum principle. First note that, as the exact solution of the problem (3.1) is given by $a(x,t) = a_0(x - ut)$, it is obvious that this solution satisfies the following maximum principle :

$$\inf_{\xi \in \mathbb{R}} a_0(\xi) \leq a(x,t) \leq \sup_{\xi \in \mathbb{R}} a_0(\xi), \qquad \forall (x,t) \in \mathbb{R} \times \mathbb{R}_+.$$

It is desirable that the numerical solution verifies similar discrete inequalities.

For instance, it is easy to prove that the left one sided scheme (3.10) satisfies these inequalities ; it is true under the condition (3.11).

We know (see (3.9)) that the left one-sided scheme can be wrote under the form

$$a_i^{n+1} = (1 - \nu)a_i^n + \nu a_{i-1}^n.$$

Under the condition (3.11), the coefficients $(1 - \nu)$ and ν are positive, which implies easily the following inequalities :

$$\min(a_i^n, a_{i-1}^n) \leq a_i^{n+1} \leq \max(a_i^n, a_{i-1}^n), \tag{3.12}$$

and, by an easy recurrence, we deduce :

$$\inf_{i \in \mathbb{Z}} a_i^0 \leq a_i^n \leq \sup_{i \in \mathbb{Z}} a_i^0, \qquad \forall i \in \mathbb{Z}, \forall n \in \mathbb{N}. \tag{3.13}$$

The discrete maximum principle is well verified.

3.1.2 Other finite difference methods

Suppose now $u \leq 0$ and $\nu > -1$, that means $-1 < \nu < 0$ and by the previous definition in paragraph 3.1.1 of k, $k = -1$. Then the equation (3.8) becomes :

$$a_i^{n+1} = (1 + \nu)a_i^n - \nu a_{i+1}^n, \tag{3.14}$$

and we obtain, by definition of ν the *right one-sided* method

$$a_i^{n+1} = a_i^n - u\frac{\Delta t}{\Delta x}\left(a_{i+1}^n - a_i^n\right) \tag{3.15}$$

Remark 1. *: One call upwind scheme or one-sided scheme the left one-sided scheme (3.10) if* $u \geq 0$ *or the right one-sided scheme (3.15) if* $u \leq 0$. *The one-sided scheme is wrote :*

$$\frac{a_i^{n+1} - a_i^n}{\Delta t} = \begin{cases} -u\frac{a_i^n - a_{i-1}^n}{\Delta x} & if \quad u \geq 0 \\ -u\frac{a_{i+1}^n - a_i^n}{\Delta x} & if \quad u \leq 0 \end{cases} \tag{3.16}$$

We can write this scheme on the following form :

$$\frac{a_i^{n+1} - a_i^n}{\Delta t} + \max(u,0)\frac{a_i^n - a_{i-1}^n}{\Delta x} + \min(u,0)\frac{a_{i+1}^n - a_i^n}{\Delta x} = 0.\bullet \tag{3.17}$$

By taking half the sum of (3.10) and (3.15) we obtain the *centered* scheme :

$$a_i^{n+1} = a_i^n - u\frac{\Delta t}{\Delta x}\left(\frac{a_{i+1}^n - a_{i-1}^n}{2}\right). \tag{3.18}$$

Unfortunately, despite the quite natural derivation of this method, it suffers surprisingly from severe stability problems and is useless in practice. This scheme does not verify the discrete maximum principle (describe in the previous sub-section 3.1.1) and the l^2 stability (see the sub-section 7.0.2) and does not converge in the associate norm. For these reasons we does not specify for (3.18) the efficient sign of u.

Another classical scheme is obtained by replacing in (3.18) a_i^n by half the sum of a_{i+1}^n and a_{i-1}^n. Hence we obtain the *Lax-Friedrichs* scheme :

$$a_i^{n+1} = \frac{a_{i+1}^n + a_{i-1}^n}{2} - u\frac{\Delta t}{2\Delta x}\left(a_{i+1}^n - a_{i-1}^n\right). \tag{3.19}$$

We do now the classical notations $a_t(= \frac{\partial a(x,t)}{\partial t})$, $a_{tt}(= \frac{\partial^2 a(x,t)}{\partial t^2})$, $a_{xx}(= \frac{\partial^2 a(x,t)}{\partial x^2})$. Starting from a limited Taylor series expansion :

$$a(x, t + \Delta t) = a(x,t) + \Delta t a_t(x,t) + \frac{1}{2}\Delta t^2 a_{tt}(x,t) + \cdots, \tag{3.20}$$

we do the observation that from $a_t = -ua_x$, we can compute :

$$a_{tt} = -ua_{xt} = -ua_{tx} = -u(-ua_x)_x = u^2 a_{xx}, \tag{3.21}$$

so that (3.20) becomes :

$$a(x, t + \Delta t) = a(x,t) - u\Delta t a_x(x,t) + \frac{1}{2}\Delta t^2 u^2 a_{xx}(x,t) + \cdots, \tag{3.22}$$

approximating the derivative a_t and a_{xx} by centered difference, and retaining only the first three terms of (3.22) we obtain the following *Lax-Wendroff* scheme :

$$a_i^{n+1} = a_i^n - u\frac{\Delta t}{2\Delta x}\left(a_{i+1}^n - a_{i-1}^n\right) - u^2\frac{\Delta t^2}{2}\frac{\left(a_{i+1}^n - 2a_i^n + a_{i-1}^n\right)}{\Delta x^2}. \tag{3.23}$$

We write also a three level scheme in time, the *leap-frog* scheme

$$a_i^{n+1} = a_i^{n-1} - u\frac{2\Delta t}{2\Delta x}\left(a_{i+1}^n - a_{i-1}^n\right). \tag{3.24}$$

This scheme is found from Taylor expansions up to the order two for the partial derivatives a_t and $a_x(=\frac{\partial a(x,t)}{\partial x})$. From these expansions we can write :

$$\frac{a_i^{n+1}-a_i^{n-1}}{2\Delta t}+u\frac{a_{i+1}^{n}-a_{i-1}^{n}}{2\Delta x}=0, \tag{3.25}$$

and (3.24) is an other writing of (3.25). The word "leap" says that we calculate a_i^{n+1} from a_i^{n-1} by jumping the values at time t^n. Methods involving more than two levels require more storage and require special startup procedures, since initially only one level of data is known.

All the previous schemes are explicit. Then we shall end this enumeration by writing two **implicit** schemes, the *centered implicit* scheme

$$a_i^{n+1}=a_i^n-u\frac{\Delta t}{2\Delta x}\left(a_{i+1}^{n+1}-a_{i-1}^{n+1}\right), \tag{3.26}$$

and the *totally implicit centered* scheme

$$a_i^{n+1}=a_i^n-u\frac{\Delta t}{2}\left(\frac{a_{i+1}^{n+1}-a_{i-1}^{n+1}}{2\Delta x}+\frac{a_{i+1}^{n}-a_{i-1}^{n}}{2\Delta x}\right). \tag{3.27}$$

For the implicit schemes the solution is obtained by the inversion of an evident matrix.

Implicit methods are rarely used for time-dependent hyperbolic problem.

Remark 2. *: All this schemes are composed of M equations (for the M unknowns* a_i^{n+1} *with* $1\le i\le M$*). In all the cases the periodicity hypotesis means that* a_i^n *and* a_i^{n+1} *are defined modulo M, that is* $a_0^n=a_M^n$, $a_1^n=a_{M+1}^n$, $a_0^{n+1}=a_M^{n+1}$ *and* $a_1^{n+1}=a_{M+1}^{n+1}$.•

3.1.3 Consistency and precision : local truncation error

Taking for instance the left one-sided scheme, we can state the following definition and lemma

Definition 1. *One says that the scheme*

$$\frac{a_i^{n+1}-a_i^{n}}{\Delta t}+u\frac{a_{i}^{n}-a_{i-1}^{n}}{\Delta x}, \tag{3.28}$$

define a consistent approximation for the partial derivative operator

$$\frac{\partial}{\partial t}+u\frac{\partial}{\partial x},$$

if, for all function $v=v(x,t)$ *sufficiently regular, the difference*

$$\left(\frac{a(x,t+\Delta t)-a(x,t)}{\Delta t}+u\frac{a(x,t)-a(x-\Delta x,t)}{\Delta x}\right)-(a_t+ua_x)(x,t), \tag{3.29}$$

tends to zero when Δx *and* Δt *tend independtly to zero.*

The difference (3.29) is called the ***local truncation error*** *(for the function a).*

Furthemore one says that the scheme (3.28) define a precise approximation to the order p in space and to the order q in time for the operator

$$\frac{\partial}{\partial t}+u\frac{\partial}{\partial x},$$

if and only if, for all regular function a, the local truncation error (3.29) tend to zero like $O(\Delta x^p+\Delta t^q)$ *when* Δx *and* Δt *tend independtly to zero.*•

Lemma 1. *The left one-sided scheme (3.10) is consistant, precise to the order one in space and time, for the advection equation approximation* $(a_t + ua_x)(x,t)$.

Proof It suffices to write Taylor expansions. For all function a in the class $\mathcal{C}^2$, we easily obtain

$$\begin{aligned} \frac{a(x,t+\Delta t)-a(x,t)}{\Delta t} &+ u\frac{a(x,t)-a(x-\Delta x,t)}{\Delta x} - (a_t + ua_x)(x,t) \\ &= \frac{\Delta t}{2}a_{tt}(x,t+\theta_t) - u\frac{\Delta x}{2}a_{xx}(x-\theta_x,t), \end{aligned} \tag{3.30}$$

with $0 \leq \theta_t \leq \Delta t, 0 \leq \theta_x \leq \Delta x$. Q.E.D. •

Often the complete proof of these truncation errors is omitted.

One check that for all the previous schemes the truncation error is given by the following array :

Schemes	Truncation Error
Left One Sided (3.10)	$O(\Delta x + \Delta t)$
Right One Sided (3.15)	$O(\Delta x + \Delta t)$
Centered (3.18)	$O(\Delta x^2 + \Delta t)$
Lax-Friedrichs (3.19)	$O(\frac{\Delta x^2}{\Delta t} + \Delta t)$
Lax-Wendroff (3.23)	$O(\Delta x^2 + \Delta t^2)$
Leap-Frog (3.24)	$O(\Delta x^2 + \Delta t^2)$
Centered Implicit (3.26)	$O(\Delta x^2 + \Delta t)$
Totally Centered Implicit (3.27)	$O(\Delta x^2 + \Delta t^2)$

Like in the previous array, the truncation error is written with the notation O. Hence it is useful to justify it.

Definition 2. *Let be two functions f and g that we are studying, for instance, when $x \longrightarrow x_0$. We write :*

$$f = O(g) \quad (when \quad x \longrightarrow x_0), \tag{3.31}$$

if it exists a function h, defined for $|x - x_0|$ sufficiently small, such that :

$$f(x) = g(x)h(x) \quad with \quad h \quad bounded. \tag{3.32}$$

If $g(x) \neq 0$ for $|x - x_0|$ sufficiently small, the previous definition means that $\frac{f}{g}$ is bounded for $|x - x_0|$ sufficiently small.•

3.1.4 Fourier analysis. Stability l^2 and convergence in this norm.

The problem is now to prove the convergence of the previous schemes. Generally to prove the convergence of a scheme it is useful to use the fundamental Lax theorem (see for instance [23] p.107). This fundamental convergence theorem for linear difference methods says that for a *consistent* linear method, *stability* is necessary and sufficient for *convergence.* Noting

$$e^n = \left(e_j^n\right)_{1\leq j\leq M} \qquad \text{and} \qquad e_j^n = a_j^n - a(x_j, t^n).$$

We can first define the l^∞-norm by

$$\| e^n \|_\infty = \sup_j |e_j^n|.$$

Ideally one might hope for convergence in this norm but it is unrealistic in approximating discontinuous solutions. The pointwise error in the neighbooring of a discontinuity typically does not go to zero uniformly as the grid is refined even though the numerical results may be perfectly satisfactory. So a more popular norm is the l^2-norm define by

$$\| e^n \|_2 = (\Delta x \sum_{j=1}^{M} |e_j^n|^2)^{\frac{1}{2}}.$$

The convergence property means that, in a certain norm (generally in the l^2-norm), $\| e^n \|$ tends towards zero when $\Delta x, \Delta t \to 0$. For instance we have the precise following result :

Proposition 1. *When $0 \leq \nu \leq 1$, the left one-sided scheme (3.10) converge in the l^2 norm : if the solution of (3.1) is sufficiently regular, then for all $T > 0$:*

$$\lim_{\Delta x, \Delta t \to 0} \left(\sup_{t^n \leq T} \| e^n \|_2 \right) = 0.\bullet \tag{3.33}$$

The following board shows the l^2-stability for the two-level (in time) schemes.

Schemes	l^2- stability
Left One sided (3.10)	Yes, only if $u \geq 0$ and $0 \leq \nu \leq 1$
Centered (3.18)	No, whatever the ν value , except $\nu = 0$
Lax-Friedrichs (3.19)	Yes, under the condition $\lvert\nu\rvert \leq 1$
Lax-Wendroff (3.23)	Yes, under the condition $\lvert\nu\rvert \leq 1$
Centered Implicit (3.26)	Yes, whatever the ν value

The Lax theorem, with the consistency and the l^2-stability properties, gives the possible convergence of any scheme.

3.2 Finite volume method.

For the pollution problem, the datas are given in surface and/or volume averages. It is appropiate, at least for this reason, to use finite volume method. In the finite difference method the number of knots should be too big, with a location of these incompatible with the datas. In practice the speed u **is not constant** and we ought to consider now the conservative formulation (3.1). The aim is to generalize the upwind scheme and other schemes by the **godunov** method which is the right way to proceed. Indeed we know that the use of conservative numerical methods preserve us against the convergence of the method towards a non-solution of (3.1) (in the nonlinear case). The finite volume method that we are using is a particular case of conservative methods. Naturally in this framework we can define methods with different orders in space. We know that first-order methods have the drawback to be diffusive. The use of greatest order methods restore less diffusive schemes but produces eventually spurious oscillations in the vicinity of pollution propagation front ; we say that these methods are dispersive. Fortunately these oscillations can be controlled by the use of **slopes limitors**.

The finite volume method is the framework of the original Godunov scheme and this scheme is the illustration of many others schemes, like MUSCL (see the subsection 3.3.2) or PPM (Piecewise Parabolic Method) (see the subsection 3.3.3), named by extension of Godunov-type. MUSCL and PPM are methods of order greater than one, for which the slope limitor technique, running with any improvements in their employment, secures the control of the spurious oscillations making the scheme **TVD** (Total Variation Diminushing)(see [23] p. 131, for definition). This important property is at least necessary for proof the convergence (by the Lax-Wendroff theorem, see for instance LeVeque [23] p.130) of the numerical solution to the solution of the conservation law (3.1) (see also the following remark 5).

3.2.1 Formulation of the finite volume method.

Rapidly we describe **the finite volume method** (see [23]) applied to the conservative form of the transport equation in one dimension in space. First we divise the spatial domain in *cells* called *finite or controls volumes*, this corresponds in one dimension to a partition of $[0, L] \subset \mathbb{R}$ by intervals, on the other hand one built discrete equations from the *integral form* ot the equation.

Let be $\left([x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}[\right)_{i \in \mathbb{Z}}$ a partition of $[0, L]$, and $\left([t^n, t^{n+1}[\right)_{n \in \mathbb{N}}$ a regular partition of $\mathbb{R}^+$. We define steps of time and space respectively written $\Delta t = t^{n+1} - t^n$ and $\Delta x_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$, the middle points $x_i = \frac{1}{2}(x_{i-\frac{1}{2}} + x_{i+\frac{1}{2}})$ and the control volume $\Omega_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}[$.

By integrating (3.1) on $\Omega_i \times]t^n, t^{n+1}[$, by Fubini, by an easy calculation we obtain the **integral form of the conservation law** :

$$\bar{a}_i^{n+1} = \bar{a}_i^n - \frac{\Delta t}{\Delta x_i} \left(F^n_{i+\frac{1}{2}} - F^n_{i-\frac{1}{2}} \right), \tag{3.34}$$

where we define the **exact flux** by :

$$F^n_{i+\frac{1}{2}} = \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} (a\,u)(x_{i+\frac{1}{2}}, t)\, dt, \tag{3.35}$$

and the average values of the exact solution, at the time t^n, on each cell by :

$$\bar{a}_i^n = \frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} a(x, t^n)\, dx. \tag{3.36}$$

The finite volume method consists in calculating approximations a_i^n of the $\bar{a}_i^n$ for any time t^n and any control interval Ω_i. Fluxes F are practically unknown and we ought necessarily to do an approximation such as (3.37) of it. The approximated fluxes depend only on the approximate solution $A^n = (a_i^n)_{i\, \in \mathcal{Z}}$. We define by A^n the set of the discrete average values of the solution calculated at time t^n. Thus we should define the **numerical flux functions** $\tilde{F}$ such that :

$$\tilde{F}_{i+\frac{1}{2}}(A^n) \approx F^n_{i+\frac{1}{2}}. \tag{3.37}$$

Finally to calculate the discrete average values (on each cell at time t^{n+1}) we write :

$$a_i^{n+1} = a_i^n - \frac{\Delta t}{\Delta x_i} \left(\tilde{F}_{i+\frac{1}{2}}(A^n) - \tilde{F}_{i-\frac{1}{2}}(A^n) \right). \tag{3.38}$$

This equation is the **finite volume formulation** of the problem and belongs to the general framework of **conservative schemes**. In fact the conservative property means that the mass of specie A (with concentration a) is preserved (when time t move from t^n to t^{n+1}) only according to the boundaries conditions.

Remark 3. *: It is necessary to keep in mind that a_i^n are now* ***box average*** *approximations of the field $a(\cdot, t)$ and not point values of this field. It is coherent with the fact that the initial condition is only given in average form by measurements in the physical ground.*•

3.2.2 Concepts for the finite volume method.

In order to prove the convergence of this class of methods, the definitions of new concepts are necessary : principally the conservative and consistency notions for a finite volume scheme.

Conservative scheme. To guarantee that the numerical method do not converge to non-solutions (see an example in [23], p.123 in a non-linear case (Burgers equation)), it turns out that it exists a very simple and natural requirement we can impose on our numerical method. One says that the schemes are **conservative** in the following sense (it is directly inspired from (3.38)) :

Definition 3. *: A finite volume scheme is conservative if and only if it exists a function F of p+q+1 arguments such that the scheme can be written on the form :*

$$\frac{a_i^{n+1}-a_i^n}{\Delta t}+\frac{F(a_{i-p}^n,a_{i-p+1}^n,...,a_{i+q}^n)-F(a_{i-p-1}^n,a_{i-p}^n,...,a_{i+q-1}^n)}{\Delta x_i}=0, \tag{3.39}$$

the function F is called the ***numerical flux function****.* •

In the simplest case p=0 and q=1 , so that F is a function of only two variables, we have a three points explicit finite volume scheme, and (3.39) becomes

$$\frac{a_i^{n+1}-a_i^n}{\Delta t}+\frac{F(a_i^n,a_{i+1}^n)-F(a_{i-1}^n,a_i^n)}{\Delta x_i}=0. \tag{3.40}$$

Then $\tilde{F}_{i+\frac{1}{2}}=F(a_i^n,a_{i+1}^n)$ is the numerical flux between the Ω_i and Ω_{i+1} cells, and it is an approximation of the exact flux (3.35). By the way it is more convenient to rewrite (3.38). For example in the case p=0 and q=1, the finite volume formulation is :

$$a_i^{n+1}=a_i^n-\frac{\Delta t}{\Delta x_i}\left(F(a_i^n,a_{i+1}^n)-F(a_{i-1}^n,a_i^n)\right). \tag{3.41}$$

Consistent scheme. The definition of the consistency given in the previous subsection 3.1.3 for finite difference schemes and regular solutions is not usable. First because, for conservative non-linear laws, the solutions can be discontinuous and then Taylor expansions do not allow to study the consistency; on the other hand Taylor expansions use point values and we are working with average values, it seems it is less convenient. We suppose that the state a takes its values in a domain $K \subset \mathcal{R}$. The "good" consistency notion is the following :

Definition 4. *: The conservative scheme (3.41) is said to be consistent with the equation (3.1) if and only if the numerical flux function verifies :*

$$F(\bar{a},\bar{a})=\bar{a}u, \qquad \forall \bar{a}\in K\bullet \tag{3.42}$$

That means the numerical flux function F is reduced to the true flux au in the case of constant state, i.e. when $a(x,t)=\bar{a}=constant$.

Remark 4. *: More generally the consistency definition for all p and q is written :*

$$F\ \underbrace{(\bar{a},\bar{a},...,\bar{a})}_{(p+q+1\ terms)}=\bar{a}u, \qquad \forall \bar{a}\in K\bullet \tag{3.43}$$

Remark 5. *: The use of conservative and consistent schemes prevents the convergence of the numerical solution towards a "bad" solution of the problem (for instance towards a solution with shock moving with a "bad" speed). The Lax-Wendroff theorem justifies this result in the more general framework*

$$w_t + (f(w))_x = 0.$$

where f is non linear. This qualifies the pertinence of the consistency and conservative definitions. We don't give the details in this framework because it is then necessary to introduce the notion of ***weak solution*** *valid for discontinuous solution. In the case of the linear advection equation the weak solution is unique. Furthermore for the pollution problem the solution doesn't really present shocks discontinuities and the notion of weak solution is not appropiate (the weak solution is the solution of the equation when the solution is discontinuous).*

We will see that the next Godunov schemes are conservative and consistent.

3.2.3 Godunov type methods : General principle and practice

In 1959, Godunov proposed a way to make use of the characteristic information within the framework of a conservative method type finite volume method. Furthermore the Godunov method is the good making for formally generalize the upwind method seen in subsubsection 3.1.2.

The principal aim is to calculate numerical flux functions $\tilde{F}_{i+\frac{1}{2}} = F(a_i^n, a_{i+1}^n)$ which appear in (3.41), we are using the *Godunov method* for that.

General principle. Let us recall first that Godunov-type schemes for scalar conservation laws can be seen (Harten [18], LeVeque [23] p.138) as a sequence of **three steps** :

Step 1. *Reconstruction* : given the cell averages $(a_i^n)_{i\in\mathbb{Z}}$, construct a function $\tilde{a}^n(x)$ which preserves these cell-averages :

$$\forall\, i \in \mathbb{Z}, \quad \frac{1}{\Delta x_i} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \tilde{a}^n\,(x)\; dx = a_i^n. \tag{3.44}$$

Step 2. *Solution* : solve **exactly** the conservation law on $[t^n, t^{n+1}]$ with $\tilde{a}^n(x)$ as the initial datum, thus obtain a solution $\tilde{a}(x, t^{n+1})$ at time t^{n+1}. For the transport equation this step consists solving the following Cauchy problem :

$$\left\{ \begin{array}{rcl} \frac{\partial \tilde{a}}{\partial t} + \frac{\partial}{\partial x}\,(\tilde{a}\,u) & = & 0 \\ \\ \tilde{a}(\cdot, t^n) & = & \tilde{a}^n \end{array} \right. \tag{3.45}$$

Step 3. *Cell averaging* : define $(a_i^{n+1})_{i\in\mathbb{Z}}$ as the cell averages of the function $\tilde{a}(\cdot, t^{n+1})$:

$$a_i^{n+1} = \frac{1}{\Delta x_i} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \tilde{a}(x, t^{n+1})\, dx.$$

In all the sequel (we will see for instance in subsubsections 3.3.1, 3.3.2 and 3.3.3 the reason of this condition ; furthermore we know it is a standard one for an **explicit** scheme associates to an evolution equation where exits always a stability condition) we impose furthermore a **stepsize restriction** :

$$\forall\, i, \quad -\Delta x_{i+1} \leq u_{i+\frac{1}{2}}^n\, \Delta t \leq \Delta x_i \tag{3.46}$$

where we put for the wind :

$$u^n_{i+\frac{1}{2}} = u(x_{i+\frac{1}{2}}, t^n).$$

In other words from the numerical values of the solution at time t^n we built a function $\tilde{a}^n$ which preserves the cell averages (in the original Godunov's method this function $\tilde{a}^n$ is a piecewise constant function with $\tilde{a}^n\,(x) = a^n_i$.) that we actualize during a timestep on $[t^n, t^{n+1}]$ in solving exactly the problem (3.45), then we project the function obtained at time t^{n+1} on the space of constant function by cells.

Practice of this method. Two problems are open in this presentation for using the Godunov type method.

- α) The numerical flux is not still defined in the equation (3.38),
- β) the exact resolution of (3.45) is not obvious a priori.

Response to the problem α).

In applying the step 2, the exact solution of (3.45) is known. Now it is easy to see that the previous steps 1 and 3 transform (3.34) in (3.38), where the numerical flux $\tilde{F}_{i+\frac{1}{2}}$ is given by :

$$\tilde{F}_{i+\frac{1}{2}}\,(A^n) = \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} (\tilde{a}\,u)\,(x_{i+\frac{1}{2}}, t)\,dt. \tag{3.47}$$

Hence the Godunov type method can be written under the form of a **conservative scheme** after the three previous phases (reconstruction, resolution and cell averaging calculations).

Response to the problem β).

To solve exactly the problem (3.45) we suppose first that the wind u is constant (i.e. uniform and permanent). Without lack of generality we can suppose that $\tilde{a}^n$ is locally integrable. For all the following methods that we will consider this hypothesis is true. Therefore we can solve (3.45) by the characteristics, that yields :

$$\forall\, x \in \mathbb{R},\ \forall\, t \in\ [t^n, t^{n+1}],\quad \tilde{a}(x,t) = \tilde{a}^n\,(x - u\,(t - t^n))\,,$$

hence, using (3.47) and the change of variable $x = x_{i+\frac{1}{2}} - u\,(t - t^n)$, we have :

$$\tilde{F}_{i+\frac{1}{2}}\,(A^n) = \frac{1}{\Delta t}\int_{x_{i+\frac{1}{2}} - u\,\Delta t}^{x_{i+\frac{1}{2}}} \tilde{a}^n\,(x)\ dx. \tag{3.48}$$

For a speed wind u not constant we are unable to solve (3.45) numerically with exactitude (theoretically it is possible with a few hypothesis for u). Here we suppose only u sufficiently regular to do the approximation :

$$\forall\, x \in \mathcal{V}(x_{i+\frac{1}{2}}),\ \forall\, t \in\ [t^n, t^{n+1}],\quad u\,(x,t) \approx u(x_{i+\frac{1}{2}}, t^n), \tag{3.49}$$

and we put, thanks to (3.48) :

$$\tilde{F}_{i+\frac{1}{2}}\,(A^n) = \frac{1}{\Delta t}\int_{x_{i+\frac{1}{2}} - u^n_{i+\frac{1}{2}}\,\Delta t}^{x_{i+\frac{1}{2}}} \tilde{a}^n\,(x)\ dx. \tag{3.50}$$

$\tilde{F}_{i+\frac{1}{2}}\,(A^n)$ is a function of A^n, because function $\tilde{a}^n$ is also define from A^n.

We shall see, precisely in subsubsection 3.3.1, that it is not possible to solve exactly (therefore with unicity) (3.45) without the CFL condition (3.46).

Remark 6. *: A synthetic formulation*
In defining Φ_u by :

$$\Phi_u : A^n \longmapsto \left(-\frac{\Delta t}{\Delta x_i} \left(\tilde{F}_{i+\frac{1}{2}}(A^n) - \tilde{F}_{i-\frac{1}{2}}(A^n) \right) \right)_{i \in \mathbb{Z}}$$

the Godunov type scheme is spelt :

$$A^{n+1} = (\, Id + \Phi_u)\,(A^n) \tag{3.51}$$

For our transport problem the application Φ_u is ***linear***.●

3.3 Godunov type methods.

We present in this context a few exemples of Godunov schemes :

- the Godunov scheme, the oldest and also the simplest;
- the minmod scheme, which is an extension of the previous method; this method is a particular case of MUSCL schemes introduced by Van Leer.
- the PPM method, extension still more elaborated.

That is really not a complete review of the Godunov type methods, for instance ENO (*Essentially Non-Oscillatory*) schemes have been introduced in this context by Harten [19], Harten and al. [18], Shu and al. [30] .

All the following methods differ only by the reconstruction algorithm.

3.3.1 The first Godunov method.

We build the function $\tilde{a}^n$ by putting

$$\forall i, \;\; \forall\, x \in \Omega_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}[, \quad \tilde{a}^n(x) = a_i^n. \tag{3.52}$$

Remark 7. *: We verify (see figure 5) that in the original Godunov method, $\tilde{a}^n$ is a* ***staircase function***.●

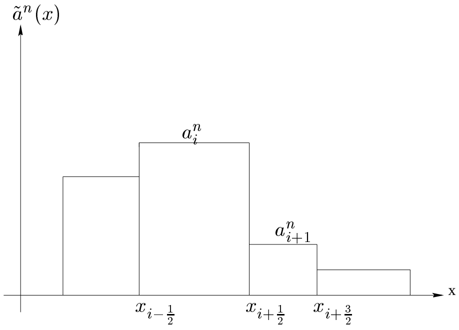


FIG. 5 – Field discretization at time t^n.

.

By direct calculation of the integral (3.50), taking into account the CFL condition (3.46), we can completely explicite the flux function for the original Godunov method with respect to the a_i^n. We have :

$$\tilde{F}_{i+\frac{1}{2}}(A^n) = F(a_i^n, a_{i+1}^n) = \begin{cases} u^n_{i+\frac{1}{2}}\, a_i^n & \text{if } u^n_{i+\frac{1}{2}} \geq 0 \\ \\ u^n_{i+\frac{1}{2}}\, a_{i+1}^n & \text{if } u^n_{i+\frac{1}{2}} < 0 \end{cases} \tag{3.53}$$

N.B. : It is possible to give an explicit expression for the flux, without ambiguity, only if the CFL condition is verified. See also comments about this CFL condition in subsubsection 3.1.1.•

From (3.38), for $u^n_{i+\frac{1}{2}} \geq 0$ (resp. $u^n_{i+\frac{1}{2}} < 0$) we find again the left one-sided method (3.10) (resp. the right one-sided method (3.15)).

We verify immediatly that the Godunov method is conservative. Numerical flux (in (3.41)) F is given by $u^n_{i+\frac{1}{2}}\, a_i^n$ or $u^n_{i+\frac{1}{2}}\, a_i^n$ taking into account the sign of $u^n_{i+\frac{1}{2}}$, and the overall original Godunov scheme depends only on two values, and verifies the consistent relation $F(\bar{a}, \bar{a}) = \bar{a}u$. The original Godunov scheme is conservative and consistent under the CFL condition (3.46).

3.3.2 The minmod scheme (MUSCL).

The idea is to improve **the reconstruction phase**, by using now a function $\tilde{a}^n$ affine by steps such that (3.44) is still true . For that we define a **slope** σ_i^n on each cell $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}[$, and we put :

$$\forall i, \ \forall\, x \in \ [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}[, \quad \tilde{a}^n(x) = a_i^n + \sigma_i^n\,(x - x_i) \tag{3.54}$$

then (3.44) is evidently true by definition of x_i $(x_i = \frac{1}{2}\,(x_{i-\frac{1}{2}} + x_{i+\frac{1}{2}})\ )$.

Furthermore we should limite the slopes σ_i^n by a "slope-limitor". A simple choice of slopes, such that the associated reconstruction let be TVD is given by *minmod* ([23], pp. 183-187) .

This type of method was initially introduced by B. Van Leer (see for instance [35] and [36]) under the appellation of MUSCL (*Monotonic Upstream-centered Scheme for Conservation Laws*) scheme.

The flux function for the minmod scheme is always given by (3.50). Taking into account the CFL condition (3.46), starting from (3.54), like in the subsection 3.3.1 we verify that, for $u^n_{i+\frac{1}{2}} \geq 0$ the integration domain for (3.50) is included in $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}[$ and $\tilde{a}^n(x) = a_i^n + \sigma_i^n\,(x - x_i)$; likewise if $u^n_{i+\frac{1}{2}} < 0$, then the integration domain for (3.50) is included in $[x_{i+\frac{1}{2}}, x_{i+\frac{3}{2}}]$ and $\tilde{a}^n(x) = a_{i+1}^n + \sigma_{i+1}^n\,(x - x_{i+1})$. Then by elementary integral calculations we find that :

$$\tilde{F}_{i+\frac{1}{2}}(A^n) = \begin{cases} \text{if} \quad u^n_{i+\frac{1}{2}} \geq 0 \qquad \text{then} & \\ u^n_{i+\frac{1}{2}}\, a_i^n + \frac{1}{2}\,(1-\nu)\ u^n_{i+\frac{1}{2}}\ \Delta x_i\, \sigma_i^n & \text{where} \qquad \nu = \frac{u^n_{i+\frac{1}{2}}\,\Delta t}{\Delta x_i} \\ \text{if} \quad u^n_{i+\frac{1}{2}} < 0 \qquad \text{then} & \\ u^n_{i+\frac{1}{2}}\, a_{i+1}^n - \frac{1}{2}\,(1+\nu)\ u^n_{i+\frac{1}{2}}\ \Delta x_{i+1}\, \sigma_{i+1}^n & \text{where} \qquad \nu = \frac{u^n_{i+\frac{1}{2}}\,\Delta t}{\Delta x_{i+1}} \end{cases}$$

We verify directly that the MUSCL scheme is conservative (the numerical flux F depends of two values (in the original Godunov scheme it depends on one value) and the overall scheme can depend on four values). It is also easy to see that $F(\bar{a}, \bar{a}) = \bar{a}u$ and the consistency condition is satisfied with respect to the CFL condition (3.46).

3.3.3 The PPM method

Developped by Collela and Woodward [5] since 1984, the PPM (*Piecewise Parabolic Method*) method is also an extension of the original Godunov method, we do the reconstruction **piecewise parabolic** : on each cell Ω_i, $\tilde{a}^n$ is a second degree polynomial. The art of this reconstruction is the pertinent choice of the coefficients of this polynomial, essentially made by appropriate values of this polynomial (see (3.62)) at the boundary of Ω_i.

The reconstruction PPM. It is easy to see that there is an unique polynomial P_i of degree two with given average a_i^n on $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}[$ and given values at $x_{i-\frac{1}{2}}$ and $x_{i+\frac{1}{2}}$. One find

$$P_i\,(x) = P_i(x_{i-\frac{1}{2}}) + \Delta a_i\,\bar{X}_i + a_{6,i}\,\bar{X}_i\,\big(1-\bar{X}_i\big) \tag{3.55}$$

Where $\bar{X}_i$ is defined by $\bar{X}_i = \frac{1}{\Delta x_i}(x - x_{i-\frac{1}{2}})$, Δa_i by $\Delta a_i = P_i(x_{i+\frac{1}{2}}) - P_i(x_{i-\frac{1}{2}})$ and finally $a_{6,i}$ by $a_{6,i} = 6\left(a_i^n - \frac{1}{2}\left(P_i(x_{i-\frac{1}{2}}) + P_i(x_{i+\frac{1}{2}})\right)\right)$.

For our polynomial reconstruction we put naturally

$$\forall\, i,\quad \forall\, x \in [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}[,\quad \tilde{a}^n(x) = P_i(x) \tag{3.56}$$

Thus the condition (3.44) shall be evidently satisfied.

The problem now is to furnish point values at $x_{i-\frac{1}{2}}$ and $x_{i+\frac{1}{2}}$ for P_i to determine it.

A first approximation. We want to obtain, for each i, an approximation $a_{i+\frac{1}{2}}$ of $a(x_{i+\frac{1}{2}}, t^n)$, value of the exact solution at the point $x_{i+\frac{1}{2}}$ and time t^n.

Systematically the method works with averages a_{i-1}^n, a_i^n, a_{i+1}^n et a_{i+2}^n respectively on the cells $[x_{i-\frac{3}{2}}, x_{i-\frac{1}{2}}[$, $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}[$, $[x_{i+\frac{1}{2}}, x_{i+\frac{3}{2}}[$ and $[x_{i+\frac{3}{2}}, x_{i+\frac{5}{2}}[$. There is an unique polynomial $R_{i+\frac{1}{2}} \in \mathbb{R}^3\,[X]$ which satisfies this averages on this intervals.

Then we put $a_{i+\frac{1}{2}} = R_{i+\frac{1}{2}}(x_{i+\frac{1}{2}})$.

After heavy and skillful calculations (it is not possible to give in this framework a simple idea of the demonstration) to get visible the slopes (like $\left(a_{i+1}^n - a_i^n\right)$) and average slopes (like δa_i) we can find :

$$\begin{aligned} a_{i+\frac{1}{2}} &= a_i^n + \frac{\Delta x_i}{\Delta x_i + \Delta x_{i+1}}\left(a_{i+1}^n - a_i^n\right) + \frac{1}{\sum_{k=-1}^{2}\Delta x_{i+k}} \\ &\times \Bigg\{ \frac{2\,\Delta x_i\,\Delta x_{i+1}}{\Delta x_i + \Delta x_{i+1}}\left[\frac{\Delta x_{i-1} + \Delta x_i}{2\,\Delta x_i + \Delta x_{i+1}} - \frac{\Delta x_{i+1} + \Delta x_{i+2}}{\Delta x_i + 2\,\Delta x_{i+1}}\right]\left(a_{i+1}^n - a_i^n\right) \\ &- \Delta x_i\,\frac{\Delta x_{i-1} + \Delta x_i}{2\,\Delta x_i + \Delta x_{i+1}}\,\delta a_{i+1} + \Delta x_{i+1}\,\frac{\Delta x_{i+1} + \Delta x_{i+2}}{\Delta x_i + 2\,\Delta x_{i+1}}\,\delta a_i \Bigg\} \end{aligned} \tag{3.57}$$

where δa_i is defined for all indices i by :

$$\begin{aligned} \delta a_i = \frac{\Delta x_i}{\Delta x_{i-1} + \Delta x_i + \Delta x_{i+1}} &\Bigg[\frac{2\,\Delta x_{i-1} + \Delta x_i}{\Delta x_i + \Delta x_{i+1}}\left(a_{i+1}^n - a_i^n\right) + \\ &\quad \ldots + \frac{\Delta x_i + 2\,\Delta x_{i+1}}{\Delta x_{i-1} + \Delta x_i}\left(a_i^n - a_{i-1}^n\right)\Bigg] \end{aligned} \tag{3.58}$$

Remark 8. *: If all stepsizes in space are constant we find that*

$$\delta a_i = \frac{\left(a_{i+1}^n - a_i^n\right) + \left(a_i^n - a_{i-1}^n\right)}{2} \tag{3.59}$$

It appears that δa_i can be interpreted like the arithmetic mean of slopes. In the general case we can only say that we have an average.•

Remark 9. *: In the case of constant stepsize in space, (3.57) gives*

$$a_{i+\frac{1}{2}} = \frac{7}{6}\left(a_i^n + a_{i+1}^n\right) - \frac{1}{12}\left(a_{i+2}^n + a_{i-1}^n\right) \tag{3.60}$$

This formula alone is out of interest in practice.•

A first correction algorithm. This first approximation doesn't insure against over- or undershooting of the reconstructed polynomial $\tilde{a}^n(x)$. We see from (3.57) that $a_{i+\frac{1}{2}}$ does not belong necessarily to the set of values defined by a_i^n and a_{i+1}^n, this property being at least desirable if we want the reconstruction algorithm and PPM itself to have good properties. A sufficient condition for that is to remplace δa_i by $\delta_m a_i$ defined by :

$$\delta_m a_i = \begin{cases} \text{if } \left(a_{i+1}^n - a_i^n\right)\left(a_i^n - a_{i-1}^n\right) > 0 \quad \text{then} \\ \min\left(\left|\delta a_i\right|, 2\left|a_i^n - a_{i-1}^n\right|, 2\left|a_{i+1}^n - a_i^n\right|\right) \operatorname{sgn}\left(\delta a_i\right) \\ \text{if } \left(a_{i+1}^n - a_i^n\right)\left(a_i^n - a_{i-1}^n\right) \leq 0 \quad \text{then} \\ 0 \end{cases} \tag{3.61}$$

Taking into account (3.55) and (3.56) we have $\tilde{a}^n(x_{i+\frac{1}{2}}) = P_i(x_{i+\frac{1}{2}})$ (because if $x = x_{i+\frac{1}{2}}$ we have $\bar{X}_i = 1$), thus we can first impose

$$P_i(x_{i+\frac{1}{2}}) = P_{i+1}(x_{i+\frac{1}{2}}) = a_{i+\frac{1}{2}}. \tag{3.62}$$

It is proving that $a_{i+\frac{1}{2}}$ belongs necessarily to the set of values defined by a_i^n and a_{i+1}^n with the correction (3.61). In fact the coefficient two (given in the original paper of Collela and Woodward) which appears in (3.61) is not optimal . For instance, in the case of constant space stepsize, the factor three insures a better adaptability to the physical behaviour, in particular with respect to the stiffness in propagation front phenomena.

The second correction algorithm : an algorithm for monotonicity. To insure good properties for the reconstruction (TVD property) we want to insure the **parabolic monotonicity** built on each cell. Parabolic because P_i is a degree two polynomial, monotone because the following behaviours for P_i are impossible after this correction (see figure 6).

For the moment this property is not true : for instance in the standard reconstruction of the cut-off function the monotonicity property is not verified after the first correction. The following algorithm realizes it by ajusting the values at interfaces of the polynomials.

It contains three phases :

First phase :

$$\text{If} \quad \left(P_i(x_{i+\frac{1}{2}}) - a_i^n\right)\left(a_i^n - P_i(x_{i-\frac{1}{2}})\right) \leqslant 0$$

$$\text{then} \quad P_i(x_{i-\frac{1}{2}}) \to a_i^n \quad \text{and} \quad P_i(x_{i+\frac{1}{2}}) \to a_i^n \tag{3.63}$$

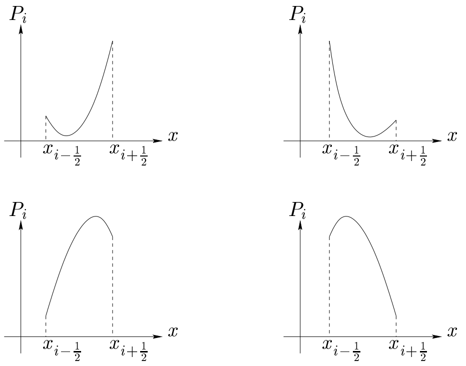


FIG. 6 – Impossible behaviour for P_i after the second correction algorithm.

Remark 10. *This first phase means that if $P_i(x_{i+\frac{1}{2}})$ and $P_i(x_{i-\frac{1}{2}})$ are not on both sides from a_i^n then we put $P_i(x_{i+\frac{1}{2}}) = P_i(x_{i-\frac{1}{2}}) = a_i^n$. Evidently if the test of this first phase is valid it is easy to see, from (3.55), that $P_i(x) \equiv P_i(x_{i+\frac{1}{2}}) = a_i^n$ on Ω_i, and the polynomial P_i is constant therefore monotone on Ω_i. If the test of the first phase is false we are, for instance, in the following configuration (see figure 7) and we do eventually the corrections of the second or third phases (in any order for the tests) so that the previous boundaries values for P_i given by $P_i(x_{i+\frac{1}{2}}) = a_{i+\frac{1}{2}}$ and $P_i(x_{i-\frac{1}{2}}) = a_{i-\frac{1}{2}}$ are eventually modified for secure P_i monotone on Ω_i.*

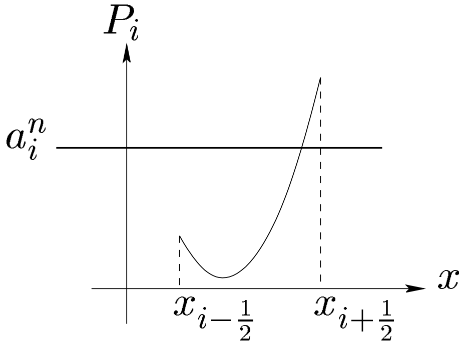


FIG. 7 – Position of a_i^n if the test of the first phase is false.

else :

Second phase :

$$If \quad \Big(P_i(x_{i+\frac{1}{2}}) - P_i(x_{i-\frac{1}{2}})\Big)\left(a_i^n - \frac{1}{2}\Big(P_i(x_{i-\frac{1}{2}}) + P_i(x_{i+\frac{1}{2}})\Big)\right) >$$

$$\frac{1}{6}\Big(P_i(x_{i+\frac{1}{2}}) - P_i(x_{i-\frac{1}{2}})\Big)^2$$

$$\text{then} \quad P_i(x_{i-\frac{1}{2}}) \to 3\,a_i^n - 2\,P_i(x_{i+\frac{1}{2}}) \tag{3.64}$$

else :

Third phase :

$$If \quad \Big(P_i(x_{i+\frac{1}{2}}) - P_i(x_{i-\frac{1}{2}})\Big)\left(a_i^n - \frac{1}{2}\Big(P_i(x_{i-\frac{1}{2}}) + P_i(x_{i+\frac{1}{2}})\Big)\right)$$

$$< -\frac{1}{6}\Big(P_i(x_{i+\frac{1}{2}}) - P_i(x_{i-\frac{1}{2}})\Big)^2$$

$$\text{then} \quad P_i(x_{i+\frac{1}{2}}) \to 3\,a_i^n - 2\,P_i(x_{i-\frac{1}{2}}) \tag{3.65}$$

It is out of order to justify here this algorithm (this justification requires a long and heavy discussion). We give only the outcome of it.

Since $a_{i+\frac{1}{2}}$ is between a_i^n and a_{i+1}^n, after reajusting, by this second algorithm, the values of the polynomial P_i at the boundaries we have necessarily, following $a_{i+1}^n \geq a_i^n$ or $a_i^n \geq a_{i+1}^n$:

$$\begin{cases} a_i^n \leqslant P_i(x_{i+\frac{1}{2}}) \leqslant a_{i+\frac{1}{2}} \leqslant P_{i+1}(x_{i+\frac{1}{2}}) \leqslant a_{i+1}^n, \\ \text{or} \\ a_i^n \geqslant P_i(x_{i+\frac{1}{2}}) \geqslant a_{i+\frac{1}{2}} \geqslant P_{i+1}(x_{i+\frac{1}{2}}) \geqslant a_{i+1}^n. \end{cases} \tag{3.66}$$

$$\text{and :} \quad \begin{cases} P_i(x_{i-\frac{1}{2}}) \leqslant a_i^n \leqslant P_i(x_{i+\frac{1}{2}}), \\ \text{or} \\ P_i(x_{i-\frac{1}{2}}) \geqslant a_i^n \geqslant P_i(x_{i+\frac{1}{2}}). \end{cases} \tag{3.67}$$

Definition 5. *: PPM is define globally by formulas (3.57) and (3.58) taking into account the first and second corrections.*•

The flux function for PPM. After a possible correction of the boundaries values of the polynomial P_i, in injecting (3.56) and (3.55) in (3.50), taking into account (3.46), we find by elementaries calculations :

$$\tilde{F}_{i+\frac{1}{2}}(A^n) = \begin{cases} u_{i+\frac{1}{2}}^n \Big(P_i(x_{i+\frac{1}{2}}) - \frac{\nu}{2}\left(\Delta a_i - \left(1 - \frac{2}{3}\nu\right) a_{6,i}\right)\Big) \\ \text{where } \nu = \frac{u_{i+\frac{1}{2}}^n \Delta t}{\Delta x_i} \quad , \text{ if } u_{i+\frac{1}{2}}^n \geqslant 0 \\ u_{i+\frac{1}{2}}^n \Big(P_{i+1}(x_{i+\frac{1}{2}}) - \frac{\nu}{2}\left(\Delta a_{i+1} + \left(1 + \frac{2}{3}\nu\right) a_{6,i+1}\right)\Big) \\ \text{where } \nu = \frac{u_{i+\frac{1}{2}}^n \Delta t}{\Delta x_{i+1}} \quad , \text{ if } u_{i+\frac{1}{2}}^n < 0 \end{cases} \tag{3.68}$$

From (3.62) and the calculation of $a_{i+\frac{1}{2}}$ by (3.57), we verify that the evaluation $P_i(x_{i+\frac{1}{2}})$ needs three values a_{i-1}^n, a_i^n and a_{i+1}^n ; likewise $P_i(x_{i-\frac{1}{2}})$ needs three values a_{i-2}^n, a_{i-1}^n and a_i^n. Like, by definition, $\Delta a_i = P_i(x_{i+\frac{1}{2}}) - P_i(x_{i-\frac{1}{2}})$ and $a_{6,i} = 6\left(a_i^n - \frac{1}{2}\left(P_i(x_{i-\frac{1}{2}}) + P_i(x_{i+\frac{1}{2}})\right)\right)$, taking into account the expression (3.68), the flux $\tilde{F}_{i+\frac{1}{2}}(A^n)$ depends of five values and the overall PPM method depends of six values. It is evident that PPM is conservative, it is written in the appropriate form (3.39) with $p = 2$ and $q = 2$. Furthermore it is easy to verify from (3.68) that $\tilde{F}\underbrace{(\bar{a},\bar{a},\bar{a},\bar{a},\bar{a})}_{(5\text{ terms})} = \bar{a}u$ (because in this case $P_i(x_{i+\frac{1}{2}}) = P_i(x_{i-\frac{1}{2}}) = a_{i+\frac{1}{2}} = \bar{a}$, hence $\Delta a_i = 0$ and $a_{6,i} = 0$) and the scheme is consistent by definition (3.43). PPM has the required good properties.

3.4 Monotonicity preserving, monotone schemes : TVD property of Godunov, MUSCL and PPM applied to our transport equation (3.1).

In this subsection, we present the stability in the sense of TVD schemes for the scalar reconstructions performed by Godunov-type schemes described in subsubsections 3.3.1, 3.3.2 and 3.3.3.

The key point for a scheme to be TVD is the verification of the total variation diminishing of the scalar reconstruction (see [23] theorem 16.4). Before to present the TVD-property it seems useful to justify the relevance of this notion. We see that a scheme like (3.41) can be writing generally under the form :

$$a_i^{n+1} = \mathcal{H}(a_{i-p-1}^n, a_{i-p}^n, a_{i-p+1}^n, ..., a_{i+q}^n), \tag{3.69}$$

then we have the following definition :

Definition 6. *: A scheme (3.69) is called **monotonicity preserving** if the operator (3.69) maps a monotone sequence into a monotone sequence.* •

That's means if the initial data a_i^0 is monotone (either nonincreasing or decreasing) as a function of i, then the solution a_0^n should have the same property for all n. For example if $a_i^0 \geqslant a_{i+1}^0$ for all i then we should have that $a_i^n \geqslant a_{i+1}^n$ for all i and n. This means in particular that oscillations cannot arise near an isolated propagating discontinuity. For TVD methods we have the following property : any TVD method is monotonicity preserving (see [23], p. 166).

The monotonicity preserving definition ought be not confuse with the **monotonicity** of a scheme :

Definition 7. *: The numerical method (3.69) is said to be **monotone** if $\mathcal{H}$ is a monotone increasing function of each of its arguments.*•

This definition means that $a_i^n \geqslant b_i^n \quad \forall i \Longrightarrow a_i^{n+1} \geqslant b_i^{n+1} \quad \forall i$.

Now it is easy to see that if each step of the subsubsection 3.2.3 is monotone (resp. TVD), so will be the entire scheme. Since the conservation law we consider here is assumed to be scalar, step two is monotone (resp. TVD). And, as the cell-averaging step three is monotone (resp. TVD) too, then if the reconstruction step is monotone (resp. TVD), so will be the entire scheme.

But unfortunately, conservative, consistent and *monotone* methods are at most first order accurate (see [16], p. 125). This is why, in the case of a higher-order Godunov-type scheme like the MUSCL and PPM schemes, we do not expect step one to be monotone. This monotonicity property would be too strong. On the other hand, the TVD property is a stability constraint which is weak enough to allow the method to be more than first order accurate in space.

This shows how important is the TVD-stability of the scalar reconstructions performed by Godunov-type schemes.

It is possible to prove that the reconstruction step is TVD for each method for our problem (3.1). So, the first Godunov method, MUSCL and PPM applied to our transport equation are TVD.

4 Modelling of the advection problem in 3D. Splitting methods.

In this section we emphasize only on the numerical modelling of the advection equation in 3D. We are concern by the advection part of (1.1), let be to solve (recall that $\vec{C}$ is a vector containing all the unknowns concentrations of all model species, $\vec{V}$ is the given three dimensional wind vector and $\vec{C}_0$ is the given vector of initial concentrations) :

$$\left\{ \begin{array}{rclcl} \frac{\partial \vec{C}}{\partial t} + div(\vec{V}\vec{C}) & = & 0 & \forall(x,y,z,t) & \in \mathbb{R}^3 \times [0,T[\\ \vec{C}(x,y,z,0) & = & \vec{C}_0(x,y,z) & \forall(x,y,z) & \in \mathbb{R}^3 \end{array} \right. \tag{4.1}$$

or, for a particular specie $a \in \vec{C}$,

$$\left\{ \begin{array}{rclrcl} \frac{\partial a}{\partial t} + div(a\vec{V}) & = & 0 & \forall (x,y,z,t) & \in & \mathbb{R}^3 \times [0,T[\\ a(x,y,z,0) & = & a_0(x,y,z) & \forall (x,y,z) & \in & \mathbb{R}^3 \end{array} \right. \tag{4.2}$$

This equation is solve with the appropriate finite volume method. We called u,v and w the components of $\vec{V}$:

$$\vec{V} = \begin{pmatrix} u \\ v \\ w \end{pmatrix}$$

4.1 Formulation 3D of the finite volume method.

This subsection is a generalization of the subsection 3.2.1. Let be $\left([t^n, t^{n+1}[\ \right)_{n\in\mathbb{N}}$ a regular partition of $\mathbb{R}^+$ and $\left([x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}[\times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}[\times [z_{k-\frac{1}{2}}, z_{k+\frac{1}{2}}[\right)_{(i,j,k)\in\mathbb{Z}^3}$ a partition of $\mathbb{R}^3$.

We define time and space steps respectively written by $\Delta t = t^{n+1} - t^n$, $\Delta x_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$, $\Delta y_j = y_{j+\frac{1}{2}} - y_{j+\frac{1}{2}}$ and $\Delta z_k = z_{k+\frac{1}{2}} - z_{k+\frac{1}{2}}$, the middle points $x_i = \frac{1}{2}\,(x_{i-\frac{1}{2}} + x_{i+\frac{1}{2}})$, $y_j = \frac{1}{2}\,(y_{j-\frac{1}{2}} + y_{j+\frac{1}{2}})$, $z_k = \frac{1}{2}\,(z_{k-\frac{1}{2}} + z_{k+\frac{1}{2}})$, and the cells by $\Omega_{i,j,k} = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}[\times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}[\times [z_{k-\frac{1}{2}}, z_{k+\frac{1}{2}}[$.

Pratically we should restrict the mesh space to a bounded domain of $\mathbb{R}^3$. Then we ought define boundaries conditions : pratically we impose that outside the domain the solution is given constant. By that the spatial index (i,j,k) describe a finite set of $\mathbb{Z}^3$, says $(i,j,k) \in \{1,\cdots,p\} \times \{1,\cdots,q\} \times \{1,\cdots,r\}$.

Now we introduce the average values of the exact solution, at the time t^n, on each cell by :

$$\bar{a}^n_{ijk} = \frac{1}{\Delta x_i \Delta y_j \Delta z_k} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \int_{z_{k-\frac{1}{2}}}^{z_{k+\frac{1}{2}}} a(x,y,z,t^n)\, dx\, dy\, dz. \tag{4.3}$$

Suppose that we know approximations a^n_{ijk} to $\bar{a}^n_{ijk}$

$$\forall\ (i,j,k), \quad a^n_{ijk} \approx \bar{a}^n_{ijk}, \tag{4.4}$$

we want to calculate a^{n+1}_{ijk} with respect to a^n_{ijk} such that $a^{n+1}_{ijk} \approx \bar{a}^{n+1}_{ijk}$.

By integration of (4.2) on $\Omega_{i,j,k}\times]t^n,\ t^{n+1}[$, by extending in 3D the technic seen in subsubsection (3.2.1), we obtain the **integral form of the conservation law** similar in 3D to (3.34) in 1D

$$\begin{aligned} \bar{a}^{n+1}_{ijk} = \bar{a}^n_{ijk} - \frac{\Delta t}{\Delta x_i} \left(F^n_{i+\frac{1}{2},\,j,\,k} - F^n_{i-\frac{1}{2},\,j,\,k} \right) - \cdots \\ \cdots - \frac{\Delta t}{\Delta y_j} \left(F^n_{i,\,j+\frac{1}{2},\,k} - F^n_{i,\,j-\frac{1}{2},\,k} \right) - \frac{\Delta t}{\Delta z_k} \left(F^n_{i,\,j,\,k+\frac{1}{2}} - F^n_{i,\,j,\,k-\frac{1}{2}} \right). \end{aligned} \tag{4.5}$$

where, for instance, the **exact fluxe** $F^n_{i+\frac{1}{2},\,j,\,k}$ is define by (we leave to the reader the writing of the other fluxes) :

$$F^n_{i+\frac{1}{2},\,j,\,k} = \frac{1}{\Delta t\, \Delta y_j\, \Delta z_k} \int_{t^n}^{t^{n+1}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \int_{z_{k-\frac{1}{2}}}^{z_{k+\frac{1}{2}}} (a\, u)(x_{i+\frac{1}{2}}, y, z, t)\, dy\, dz\, dt, \tag{4.6}$$

Let be $A^n = \left(a^n_{ijk} \right)_{(i,j\,k)\in\{1,\ldots,p\}\times\{1,\ldots,q\}\times\{1,\ldots,r\}}$, suppose we should define **numerical fluxes functions** such that :

$$\tilde{F}_{i+\frac{1}{2},\,j,\,k}\,(A^n) \approx F^n_{i+\frac{1}{2},\,j,\,k}$$

$$\tilde{F}_{i,\,j+\frac{1}{2},\,k}\left(A^n\right) \approx F^n_{i,\,j+\frac{1}{2},\,k}$$
$$\tilde{F}_{i,\,j,\,k+\frac{1}{2}}\left(A^n\right) \approx F^n_{i,\,j,\,k+\frac{1}{2}}$$

finally, from (4.5) and because we have suppose that $\forall\ (i,j,k),\ a^n_{ijk} \approx \bar{a}^n_{ijk}$ (see (4.4)), we calculate the discrete average values (on each cell at time t^{n+1}) by the following scheme :

$$a^{n+1}_{ijk} = a^n_{ijk} - \frac{\Delta t}{\Delta x_i}\left(\tilde{F}_{i+\frac{1}{2},\,j,\,k}\left(A^n\right) - \tilde{F}_{i-\frac{1}{2},\,j,\,k}\left(A^n\right)\right) - \cdots$$
$$\cdots - \frac{\Delta t}{\Delta y_j}\left(\tilde{F}_{i,\,j+\frac{1}{2},\,k}\left(A^n\right) - \tilde{F}_{i,\,j-\frac{1}{2},\,k}\left(A^n\right)\right) - \frac{\Delta t}{\Delta z_k}\left(\tilde{F}_{i,\,j,\,k+\frac{1}{2}}\left(A^n\right) - \tilde{F}_{i,\,j,\,k-\frac{1}{2}}\left(A^n\right)\right) \tag{4.7}$$

Now it suffices to calculate the numerical fluxes functions.

4.2 3D Fluxes calculations.

By simplification we give the details uniquely for $\tilde{F}_{i+\frac{1}{2},\,j,\,k}$.

It is also a generalization of the 1D calculation (see subsubsection 3.2.3). First we suppose that the wind $\vec{V}$ is constant. We can solve exactly (4.2) by the characteristic method, that means that the solution of this equation is given by :

$$\forall\, \vec{x} \in \mathcal{R}^3,\ \forall\, t \geqslant 0, \quad a\left(\vec{x},t\right) = a_0\left(\vec{x} - t\,\vec{V}\right),$$

then

$$\forall\, \vec{x} \in \mathcal{R}^3,\ \forall\, t \geqslant 0, \quad a\left(\vec{x},t\right) = a_0\left(\vec{x} - \left(t - t^n\right)\vec{V} - t^n\vec{V}\right) = a\left(\vec{x} - \left(t - t^n\right)\vec{V}, t^n\right).$$

Therefore (4.6) gives :

$$F^n_{i+\frac{1}{2},\,j,\,k} = \frac{1}{\Delta t\,\Delta y_j\,\Delta z_k} F_x \tag{4.8}$$

with

$$F_x = \int_{t^n}^{t^{n+1}} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \int_{z_{k-\frac{1}{2}}}^{z_{k+\frac{1}{2}}} (a\,u)(x_{i+\frac{1}{2}} - u(t-t^n), y - v(t-t^n), z - w(t-t^n), t^n)\, dy\, dz\, dt. \tag{4.9}$$

For using an unidimensional reconstruction we do the hypotesis :

$$\frac{|v|\ \Delta t}{\Delta y_j} \ll 1 \qquad \text{and} \qquad \frac{|w|\ \Delta t}{\Delta z_k} \ll 1,$$

by that one has :

$$y - v(t-t^n) = y_j + (y - y_j)\,(1 - \underbrace{\frac{v(t-t^n)}{y - y_j}}_{\sim\frac{v\,\Delta t}{\Delta y_j}}) \approx y_j + (y - y_j) = y, \tag{4.10}$$

and, with the same approximation

$$z - w(t - t^n) \approx z. \tag{4.11}$$

Then, taking into account (4.9), (4.6) becomes

$$F^n_{i+\frac{1}{2},\,j,\,k} = \frac{1}{\Delta t\,\Delta y_j\,\Delta z_k}\int_{t^n}^{t^{n+1}}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}}\int_{z_{k-\frac{1}{2}}}^{z_{k+\frac{1}{2}}}(a\,u)(x_{i+\frac{1}{2}}-u(t-t^n),y,z,t^n)\,dy\,dz\,dt$$

and using the change of variable $x = x_{i+\frac{1}{2}} - u(t-t^n)$ we have

$$F^n_{i+\frac{1}{2},\,j,\,k} = \frac{1}{\Delta t\,\Delta y_j\,\Delta z_k}\int_{x_{i+\frac{1}{2}}-u\,\Delta t}^{x_{i+\frac{1}{2}}}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}}\int_{z_{k-\frac{1}{2}}}^{z_{k+\frac{1}{2}}}(a(x,y,z,t^n)\,dy\,dz$$

says,

$$F^n_{i+\frac{1}{2},\,j,\,k} = \frac{1}{\Delta t}\int_{x_{i+\frac{1}{2}}-u\,\Delta t}^{x_{i+\frac{1}{2}}} f_{jk}\,(x,t^n)dx, \tag{4.12}$$

where

$$f_{jk}(x,t^n) = \frac{1}{\Delta y_j\,\Delta z_k}\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}}\int_{z_{k-\frac{1}{2}}}^{z_{k+\frac{1}{2}}} a(x,y,z,t^n)\,dy\,dz.$$

But, from (4.3), we know that the averages, with respect to variable x, of the function $f_{jk}(.,t^n)$ are the $\bar{a}^n_{ijk}$. We know also that, approximatively one has (see (4.4)) $a^n_{ijk} = \bar{a}^n_{ijk}$.

Therefore, by any method seen in subsection 3.3, we can reconstruct a function $\tilde{f}^n_{jk}$ from the a^n_{ijk}, depending only on neighboring cells in direction x, such that :

$$\begin{cases} \forall\, x \in \mathbb{R}, \quad \tilde{f}^n_{jk}(x) \approx f_{jk}(x,t^n) \\[2ex] \dfrac{1}{\Delta x_i}\displaystyle\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \tilde{f}^n_{jk}(x)\,dx = a^n_{ijk} \end{cases}$$

the function $\tilde{f}^n_{jk}(x)$ is similar to the function $\tilde{a}^n\,(x)$ seen in 1D finite volume method.

Finally, when the wind is constant, we define the numerical flux function in **direction x**, from (4.12), by :

$$\tilde{F}_{i+\frac{1}{2},\,j,\,k}\,(A^n) = \frac{1}{\Delta t}\int_{x_{i+\frac{1}{2}}-u\,\Delta t}^{x_{i+\frac{1}{2}}} \tilde{f}^n_{jk}(x) \tag{4.13}$$

In the general case where $\vec{V}$ is not constant we put, like in subsection 3.2.3, see (3.49) and (3.48) :

$$\tilde{F}_{i+\frac{1}{2},\,j,\,k}\,(A^n) = \frac{1}{\Delta t}\int_{x_{i+\frac{1}{2}}-u^n_{i+\frac{1}{2},\,j,\,k}\,\Delta t}^{x_{i+\frac{1}{2}}} \tilde{f}^n_{jk}(x)\,dx, \tag{4.14}$$

where

$$u^n_{i+\frac{1}{2},j,k} = u(x_{i+\frac{1}{2}},y_j,z_k,t^n). \tag{4.15}$$

The formulas for all the numerical fluxes are valid under the hypotesis :

$$\forall\,(i,j,k), \quad \begin{cases} \dfrac{\left|u^n_{i+\frac{1}{2},\,j,\,k}\right|\Delta t}{\Delta x_i} \ll 1, \\[2ex] \dfrac{\left|v^n_{i,\,j+\frac{1}{2},\,k}\right|\Delta t}{\Delta y_j} \ll 1, \\[2ex] \dfrac{\left|w^n_{i,\,j,\,k+\frac{1}{2}}\right|\Delta t}{\Delta z_k} \ll 1. \end{cases} \tag{4.16}$$

Independtly, for the same reason seen in subsection 3.2.3, we impose furthermore a **stepsize restriction**, the following CFL condition :

$$\forall\ (i,j,k),\quad \left\{\begin{array}{l} -\Delta x_{i+1} \leqslant u^n_{i+\frac{1}{2},j,k}\,\Delta t \leqslant \Delta x_i, \\ -\Delta y_{j+1} \leqslant v^n_{i,j+\frac{1}{2},k}\,\Delta t \leqslant \Delta y_j, \\ -\Delta z_{k+1} \leqslant w^n_{i,j,k+\frac{1}{2}}\,\Delta t \leqslant \Delta z_k. \end{array}\right. \tag{4.17}$$

These conditions are **necessary** (see again paragraph “The CFL condition” in 3.1.1) for the construction of a 3D finite volume method. They don't say that they are **sufficient** for stability (see an example in 4.4). The TVD property of a finite volume scheme is a 1D property. Examination of definition for stability (in a sense to precise) and necessary and sufficient condition to make it in a 3D finite volume scheme, is out of our presentation (see [14] for the aficionados).

4.3 A synthetic formulation

To introduce splitting method we define the operators

$$\Phi_u : \left|\begin{array}{ccc} \mathbb{R}^{\{1,\dots,p\}\times\{1,\dots,q\}\times\{1,\dots,r\}} & \longrightarrow & \mathbb{R}^{\{1,\dots,p\}\times\{1,\dots,q\}\times\{1,\dots,r\}} \\ A^n & \longmapsto & \left(-\frac{\Delta t}{\Delta x_i}\left(\tilde{F}_{i+\frac{1}{2},\,j,\,k}\left(A^n\right) - \tilde{F}_{i-\frac{1}{2},\,j,\,k}\left(A^n\right)\right)\right)_{(i,j,k)} \end{array}\right.$$

$$\Phi_v : \left|\begin{array}{ccc} \mathbb{R}^{\{1,\dots,p\}\times\{1,\dots,q\}\times\{1,\dots,r\}} & \longrightarrow & \mathbb{R}^{\{1,\dots,p\}\times\{1,\dots,q\}\times\{1,\dots,r\}} \\ A^n & \longmapsto & \left(-\frac{\Delta t}{\Delta y_j}\left(\tilde{F}_{i,\,j+\frac{1}{2},\,k}\left(A^n\right) - \tilde{F}_{i,\,j-\frac{1}{2},\,k}\left(A^n\right)\right)\right)_{(i,j,k)} \end{array}\right.$$

and :

$$\Phi_w : \left|\begin{array}{ccc} \mathbb{R}^{\{1,\dots,p\}\times\{1,\dots,q\}\times\{1,\dots,r\}} & \longrightarrow & \mathbb{R}^{\{1,\dots,p\}\times\{1,\dots,q\}\times\{1,\dots,r\}} \\ A^n & \longmapsto & \left(-\frac{\Delta t}{\Delta x_i}\left(\tilde{F}_{i,\,j,\,k+\frac{1}{2}}\left(A^n\right) - \tilde{F}_{i,\,j,\,k-\frac{1}{2}}\left(A^n\right)\right)\right)_{(i,j,k)} \end{array}\right.$$

by that our conservative scheme is written :

$$A^{n+1} = \left(\mathrm{Id} + \Phi_u + \Phi_v + \Phi_w\right)\left(A^n\right). \tag{4.18}$$

N. B. : The approximations (4.10) and (4.11) means than the transport of particles in the time interval $[t^n, t^{n+1}]$ are only possible according to the directions (Ox), (Oy) and (Oz), says uniquely between cells with a common face. It don't exists, in this approach, **diagonal fluxes**. Furthermore, due to hypotesis (4.16), we are waiting eventually to small stepsize in time.•

4.4 Splitting methods.

This methods are usually used. Any technics are possible. The described method is one of the simplest.

4.4.1 Principle of the method.

Instead apply to A^n the operator $\mathrm{Id}+\Phi_u+\Phi_v+\Phi_w$ we apply successively the operators $\mathrm{Id}+\Phi_u$, $\mathrm{Id}+\Phi_v$ and $\mathrm{Id}+\Phi_w$, that means that we replace the scheme $A^{n+1} = \left(\mathrm{Id} + \Phi_u + \Phi_v + \Phi_w\right)\left(A^n\right)$ by :

$$A^{n+1} = \left(\mathrm{Id} + \Phi_w\right) \circ \left(\mathrm{Id} + \Phi_v\right) \circ \left(\mathrm{Id} + \Phi_u\right)\left(A^n\right). \tag{4.19}$$

This **time-splitting** is equivalent to do the transport of particles according to the direction (Ox), then according to the direction (Oy) and finally according the direction (Oz). That means, in the time interval $[t^n, t^{n+1}]$, we cancel in three successive steps two of the components of the wind $\vec{V}$: first v and w, then u and w, and finally v and w. In this context the **hypotesis (4.16) is useless**. In any phase of this algorithm, the transfer of particles are **already** parallel to axes. But, in return, the stepsize time limitation (4.17) is always necessary.

4.4.2 Existence of diagonal fluxes.

By this splitting the transfer of particles is possible not only between faces, but also between cells having a common edge or vertex. In this sense the splitting is better well-suited to solve the 3D transport.

For instance exists the 2D classic test of the cone in a 2D rotating wind field monodimensionally divergence free (that is $\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} = 0$), proposed by Crowley [4], for which the splitting method gives better results that the method without splitting (see [3]). Unfortunately, for some other tests, when the monodimensional divergence of the flow field is not zero ($\frac{\partial u}{\partial x} \neq 0$, $\frac{\partial v}{\partial y} \neq 0$) the time-splitting algorithm creates over- and undershooting errors that increases with time (see Easter [12]). For instance consider the 2D strong deformational flow of the Staniforth and al (see [33]) which is divergence free (that is $\mathrm{div}\vec{V} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0$) but not monodimensionally divergence free, which is the general case. Then, when A^n is constant and equal to a on all the cells, the no-splitting method gives the exact solution $A^{n+1} = a$, whereas the time-splitting method produces on artificial gradient in a that increases with the simulation time. It would be interesting to correct the time-splitting method such that it gives also the exact solution. Clappier ([3]) has developed this correction factor. Consider the time-splitting method (4.19) in 2D, let be

$$A^{n+1} = (\mathrm{Id} + \Phi_v) \circ (\mathrm{Id} + \Phi_u)\,(A^n) = [(\mathrm{Id} + \Phi_u) + \Phi_v(\mathrm{Id} + \Phi_u)]\,(A^n)\,, \tag{4.20}$$

then Clappier introduce an appropriate corrector Ψ_u such that

$$A^{n+1} = [(\mathrm{Id} + \Phi_u) + \Phi_v(\mathrm{Id} + \Phi_u - \Psi_u)]\,(A^n)\,. \tag{4.21}$$

By that, the time-splitting method, for the test of Clappier-Staniforth, gives the exact solution because, when A^n is constant and equal to a on all the cells, $\Phi_u = \Psi_u$ and (4.21) gives

$$A^{n+1} = (\mathrm{Id} + \Phi_u + \Phi_v)\,(A^n)$$

that is the scheme without splitting. Furthermore, for the Crowley test for which the field is monodimensionally divergence free, $\Psi_u = 0$ and the corrected time-splitting scheme is the original time-splitting scheme (4.20) which all its performances. For more informations the reader can consult [3].

4.4.3 Alternating splitting.

We can, see [3] for results, inverse at each iteration the order of the factors in (4.19), that gives :

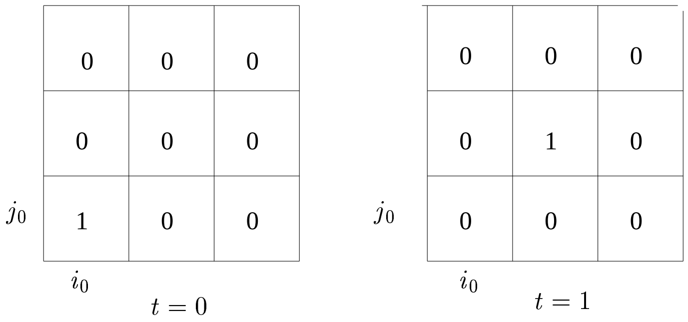
$$A^{n+1} = \begin{cases} (\mathrm{Id} + \Phi_w) \circ (\mathrm{Id} + \Phi_v) \circ (\mathrm{Id} + \Phi_u)\,(A^n) & \text{if } n \text{ is even,} \\ \\ (\mathrm{Id} + \Phi_u) \circ (\mathrm{Id} + \Phi_v) \circ (\mathrm{Id} + \Phi_w)\,(A^n) & \text{if } n \text{ is odd.} \end{cases} \tag{4.22}$$

4.4.4 An illustrative example.

Now, we are going to explore "manually" a very simple numerical case in 2D. We suppose $\Delta x_i = \Delta y_j = 1$ and we are taking a diagonal wind ($u = v = 1$). For the initial condition we put :

$$a^0_{ij} = \begin{cases} 1 & \text{if } (i,j) = (i_0, j_0) \\ \\ 0 & \text{if not} \end{cases}$$

the experience time is fixed to $T = 1$, we consider the square showed in figure 8 (we don't precise, in this example, the domain Ω at which this square belongs ; it is out of interest for our purpose). We suppose in first experiences $\Delta t = \frac{1}{2}$.



	0	0	0
	0	0	0
j_0	1	0	0
	i_0		

$t = 0$

	0	0	0
	0	1	0
j_0	0	0	0
	i_0		

$t = 1$

FIG. 8 – Diagonal wind : initial condition and exact solution.

The figure 8 shows the initial condition and the exact solution obtained by the characteristic method according to the diagonal wind. Following the necessary CFL condition (4.17) the accepted maximum stepsize is 1. For this elementary problem it is easy to see that the three associated schemes 1D (Godunov, MUSCL and PPM) have similar fluxes expressions, near a coefficient, to the Godunov scheme that we recall (see (3.53)) :

$$\tilde{F}_{i+\frac{1}{2}}(A^n) = F(a^n_i, a^n_{i+1}) = \begin{cases} u^n_{i+\frac{1}{2}}\, a^n_i & \text{if } u^n_{i+\frac{1}{2}} \geq 0 \\ \\ u^n_{i+\frac{1}{2}}\, a^n_{i+1} & \text{if } u^n_{i+\frac{1}{2}} < 0 \end{cases}$$

We know that the Godunov scheme is the upwind scheme.

In the **no splitting case**, (4.18) gives in 2D :

$$A^{n+1} = (\,\mathrm{Id} + \Phi_u + \Phi_v)\,(A^n)\,.$$

let be, with the given stepsize in space and time, $\frac{\Delta t}{\Delta x_i} = \frac{1}{2}$, and :

$$a^{n+1}_{i,j} = a^n_{i,j} - \frac{1}{2}(a^n_{i,j} - a^n_{i-1,j}) - \frac{1}{2}(a^n_{i,j} - a^n_{i,j-1}) \tag{4.23}$$

we see that $a^{n+1}_{i,j}$ don't depend of $a^n_{i-1,j-1}$, the diagonal flux is numerically missing in this scheme.

Therefore one has at time $t = 0.5$:

$$a^1_{i_0,j_0} = a^0_{i_0,j_0} - \frac{1}{2}(a^0_{i_0,j_0} - a^0_{i_0-1,j_0}) - \frac{1}{2}(a^0_{i_0,j_0} - a^0_{i_0,j_0-1}) = 1 - \frac{1}{2}(1-0) - \frac{1}{2}(1-0) = 0,$$

similarly $a^1_{i_0+1,j_0} = \frac{1}{2}$, $a^1_{i_0,j_0+1} = \frac{1}{2}$ and $a^1_{i_0+1,j_0+1} = 0$, at time $t = 1$ the cells values are indicated on the figure 9. We see a deformation along the second bisectrix.

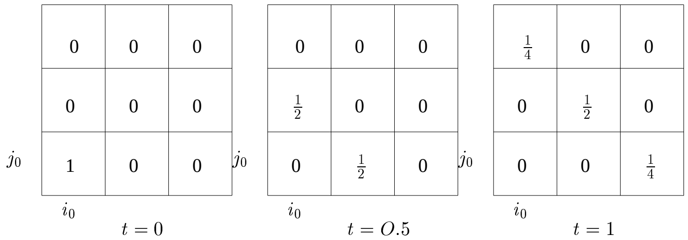


FIG. 9 – Diagonal wind : resolution without splitting for $\Delta t = 0.5$.

In the **splitting case** we have (see (4.19)) in 2D :

$$A^{n+1} = (\,\mathrm{Id} + \Phi_v) \circ (\,\mathrm{Id} + \Phi_u)\,(A^n)\,,$$

with

$$(\,\mathrm{Id} + \Phi_u)\,(A^n) = a^n_{i,j} - \frac{1}{2}(a^n_{i,j} - a^n_{i-1,j}) = \frac{a^n_{i,j} + a^n_{i-1,j}}{2},$$

and

$$(\,\mathrm{Id} + \Phi_v)\,(A^n) = \frac{a^n_{i,j} + a^n_{i,j-1})}{2}.$$

That means

$$a^{n+1}_{i,j} = \frac{a^n_{i,j} + a^n_{i,j-1} + a^n_{i-1,j} + a^n_{i-1,j-1}}{4}, \tag{4.24}$$

we see that $a^{n+1}_{i,j}$ depends on $a^n_{i-1,j-1}$, the diagonal flux is considering.

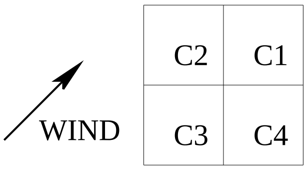


FIG. 10 – Diagonal wind. Splitting 2D

The formula (4.24) shows (see figure 10) that the value in cell $C1$ (at right and top of the figure), for the given wind of our example, is equal at the sum of the values on the four described cells divided by four. Therefore one has at time $t = 0.5$, $a^1_{i_0,j_0} = \frac{1}{4}$, $a^1_{i_0+1,j_0} = \frac{1}{4}$, $a^1_{i_0,j_0+1} = \frac{1}{4}$ and $a^1_{i_0+1,j_0+1} = \frac{1}{4}$, and at time $t = 1$ the cells values indicated in figure 11.

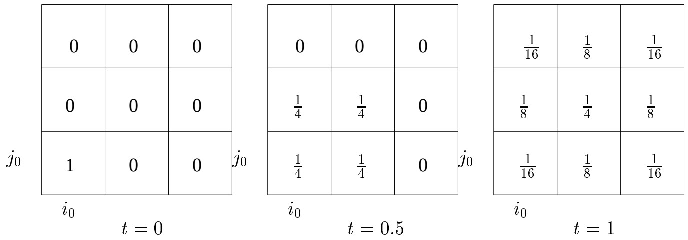


FIG. 11 – Diagonal wind : resolution with splitting for $\Delta t = 0.5$.

In the two cases the solution is flatten out and spreading out by diffusion.

But we know , by the CFL's condition, that $\Delta t = 1$ is possible. With $\Delta t = 1$ then, in the no splitting case we have $a^1_{i_0,j_0} = -1$, $a^1_{i_0+1,j_0} = 1$, $a^1_{i_0,j_0+1} = 1$ and $a^1_{i_0+1,j_0+1} = 0$; the *no splitting method gives, in this example, no positive solution.*

The splitting method is advantageous, it can be seen really in this example. Furthermore many numerical experiments prove it.

On the other hand, for the splitting method we find :

$$(\,\mathrm{Id} + \Phi_u)\,(A^n) = a^n_{i,j} - (a^n_{i,j} - a^n_{i-1,j}) = a^n_{i-1,j},$$

and

$$(\,\mathrm{Id} + \Phi_v)\,(A^n) = a^n_{i,j-1}.$$

that means $a^{n+1}_{i,j} = a^n_{i-1,j-1}$. Therefore *in one iteration* (because $\Delta t = T = 1$) we have $a^1_{i_0,j_0} = 0$, $a^1_{i_0+1,j_0} = 0$, $a^1_{i_0,j_0+1} = 0$ and $a^1_{i_0+1,j_0+1} = 1$, *we obtain the exact solution.*•

N.B. : In all the experiences the total mass is preserved in a time iteration. The scheme is conservative according to the theorical analysis.•

4.4.5 Analysis of the last experience : positivity and sufficient condition for stability.

i) **Positivity :** In the no splitting case it appears no positive solution. In fact it would be necessary to do an analysis of the positivity (and also of the stability, it is possible in this very simple case) of a scheme. A scheme is **positif** if $A^n \geqslant 0 \quad \Longrightarrow \quad A^{n+1} \geqslant 0$. It can be prove that for a splitting (resp. no splitting) scheme a necessary condition for positivity is, for $\Delta x_i = 1$ and $u = v = 1$, $\Delta t \leqslant 1$ (resp. $\Delta t \leqslant \frac{1}{2}$). That explains that in the scheme without splitting, when $\Delta t = 1$, the condition of positivity being violated, it appears negative solution.

ii) **One sufficient condition for stability for this very particular example .** We can given (for this case, we insist) a sufficient stability condition for the no splitting case. In finite dimension (the finite number of cells), all the norms are equivalent. Define the norm l^1 by

$$\|a^n\| = \sum_{i,j} |a^n_{i,j}|.$$

In introducing Δt, (4.23) becomes :

$$a^{n+1}_{i,j} = a^n_{i,j} - \Delta t(a^n_{i,j} - a^n_{i-1,j}) - \Delta t(a^n_{i,j} - a^n_{i,j-1})$$

let be :

$$a^{n+1}_{i,j} = 2(\frac{1}{2} - \Delta t)a^n_{i,j} + \Delta t a^n_{i-1,j} + \Delta t a^n_{i,j-1}$$

Suppose now $0 \leqslant \Delta t \leqslant \frac{1}{2}$, we have

$$\sum_{i,j} |a^{n+1}_{i,j}| \leqslant (2(\frac{1}{2} - \Delta t) + \Delta t + \Delta t) \sum_{i,j} |a^n_{i,j}| = \sum_{i,j} |a^n_{i,j}|.$$

and we obtain the stability relation $\sum_{i,j} |a^{n+1}_{i,j}| \leqslant \sum_{i,j} |a^n_{i,j}|$. We find a sufficient condition (for stability) $0 \leqslant \Delta t \leqslant \frac{1}{2}$ (here it is the same that the condition for positivity), near notations, similar to (18.7) in [23].•

Furthermore we see that the splitting gives better and better results when Δt is increasing until $\Delta t_{max} = 1$. The problem of the positivity, with an associated Δt_{max} different of the Δt given by the CFL's condition, is typical of the transport modeling in dimension strictly greatest than one.

5 Modelling of the chemical differential equations

This presentation summarize the concepts for solving ODE's (Ordinary Differential Equations). For more informations, inside a profuse literature, the reader can refer to [6] (in French), [15] and [20].

5.1 Generalities.

The chemical system of nonlinear differential equations doesn't have generally an explicit solution. It is necessary to dispose of robust and cheap numerical methods. Furthermore the chemical system is stiff, that means the solutions are characterized by very different scale time, and by that simple numerical method, like the Euler-Cauchy method, abort.

5.1.1 A first method : the Euler-Cauchy method.

Let be $I_0 = [t_0,\ t_0 + T]$, $T > 0$, a close interval of $\mathbb{R}$ and a continuous nonlinear function f from $I_0 \times \mathbb{R}^m$ in $\mathbb{R}^m$. One wish to solve the following system of ordinary differential equations :

$$\left\{ \begin{array}{l} y'(t) = f(t,\ y(t)), \quad t \in I_0, \\ y(t_0) = y_0, \quad y_0 \text{ given in } \mathbb{R}^m \end{array} \right. \tag{5.1}$$

We suppose that (5.1) has an unique global solution $y \in \mathcal{C}^1(I_0\ ;\mathbb{R}^m)$ on the interval I_0. Note that the solution of (5.1) verify the following integral equation :

$$y(t) = y_0 + \int_{t_0}^{t} f(s,\ y(s))\ ds, \quad t \in I_0. \tag{5.2}$$

Now let be a subdivision of the interval I_0 :

$$t_0 < t_1 < \ldots < t_n < \ldots < t_N = t_0 + T$$

and we put

$$h_n = t_{n+1} - t_n \qquad ; \qquad h = \max_{0 \leq n \leq N-1} h_n.$$

For built an approximation method for (5.1), we remark from (5.2) that :

$$y(t_{n+1}) = y_n + \int_{t_n}^{t_{n+1}} f(t,\ y(t))\ dt. \tag{5.3}$$

Like

$$\int_{t_n}^{t_{n+1}} f(t,\ y(t))\ dt \simeq h_n f(t_{n,;} y(t_n)),$$

we replace the relation (5.3) by :

$$y_{n+1} = y_n + h_n f(t_n,\ y_n), \quad 0 \leq n \leq N-1. \tag{5.4}$$

It is the **Euler-Cauchy** method or the **explicit Euler method** or the **forward Euler method** : from the initial condition y_0 we do the calculation by (5.4) of the sequence $(y_n)_{0 \leq n \leq N}$.

A natural question is to know if y_n represent an approximation of $y(t_n)$. We put

$$e_n = y(t_n) - y_n \tag{5.5}$$

and we search if

$$\lim_{h \to 0} \max_{0 \leq n \leq N} \parallel e_n \parallel = 0 \tag{5.6}$$

where $\parallel . \parallel$ means a norm in $\mathbb{R}^m$. That is the **convergence** question.

5.1.2 Local troncature error :

It is define by the error done when we replace the numerical solution y_n by $y(t_n)$ in (5.4), so we define the quantity

$$\epsilon_n = y(t_{n+1}) - y(t_n) - h_n f(t_n,\ y(t_n)) \tag{5.7}$$

called the **local error of troncature**. If y is a little more regular, $y \in \mathcal{C}^2(I_0\ ;\mathbb{R}^m)$, a Taylor expansion up to the order two gives :

$$\epsilon_n = \int_{t_n}^{t_{n+1}} (t_{n+1} - s) y''(s)\ ds = \frac{h_n^2}{2} y''(\theta_n), \quad t_n < \theta_n < t_{n+1} \tag{5.8}$$

Another evident relation is useful, from (5.4) and (5.7), by definition of e_n, we deduce

$$e_{n+1} = e_n + h_n(f(t_n,\ y(t_n)) - f(t_n,\ y_n)) + \epsilon_n \tag{5.9}$$

5.1.3 Lipschitz condition for numerical convergence :

For study the convergence of this method we suppose that the function f verify a **global Lipschitz condition** on the interval I_0, i.e. it exists a constant $L > 0$ such that :

$$\| f(t,\ y) - f(t,\ z) \| \leq L\ \| y - z \| \quad \forall\, t\ \in I_0,\ \forall\, y,\ z \in \mathbb{R}^m. \tag{5.10}$$

then one prove that the Euler-Cauchy method is convergent. Furthermore if the solution y of (5.1) belongs to $\mathcal{C}^2(I_0\ ;\mathbb{R}^m)$ we have :

$$\| e_n \| \leq h \int_{t_0}^{t_n} \| y''(s) \| \exp(L(t_n - s)) ds \tag{5.11}$$

the error is in $O(h)$, one says that this method is a first order one, that implies to use small stepsizes and in many cases it is expensive. Happily it is possible to use more sophisticated methods with largest stepsizes.

5.1.4 A second method : The implicit Euler method or backward Euler method

That is define naturally by

$$y_{n+1} = y_n + h_n f(t_{n+1},\ y_{n+1}), \quad 0 \leq n \leq N-1. \tag{5.12}$$

At each step for calculating y_{n+1} one must consequently solve an nonlinear (in general) equation in $\mathbb{R}^m$ and by that this method is called **implicit**, y_{n+1} is not immediatly given. For this reason the coast of this method can be eventually greatest than those of the explicit one.

We define now naturally the local troncature error by

$$\epsilon_n = y(t_{n+1}) - y(t_n) - h_n f(t_{n+1},\ y(t_{n+1})). \tag{5.13}$$

By a taylor expansion up to the order two we have

$$y(t_n) = y(t_{n+1}) - h_n y'(t_{n+1}) + \int_{t_{n+1}}^{t_n} (t_n - s) y''(s)\ ds,$$

therefore one has :

$$\epsilon_n = -\int_{t_{n+1}}^{t_n} (t_n - s) y''(s)\ ds = -\frac{h_n^2}{2} y''(\theta_n), \quad t_n < \theta_n < t_{n+1}, \tag{5.14}$$

and, with the same previous definition of the error e_n, (5.9) becomes

$$e_{n+1} = e_n + h_n(f(t_{n+1},\ y(t_{n+1})) - f(t_{n+1},\ y_{n+1})) + \epsilon_n. \tag{5.15}$$

5.1.5 Stiff problems : explicit or implicit schemes ?

The Euler-Cauchy method allows an approximation of solution for (5.1) under the Lipschitz condition (5.10). When the Lipschitz constant L is "not too great" (resp. is "too great") one says that the differential systeme is **well-conditioned** (resp. **ill-conditioned**) on the interval I_0. Consider for instance the following problem.

We suppose $t_0 = 0$, let be a scalar function slowly varying $g \in \mathcal{C}^1(I_0)$, we consider the differential equation (see [8] p.6)) proposed by Shampine and Gear :

$$\begin{cases} y' = -\lambda(y - g(t)) + g'(t), \quad \lambda \in \mathbb{R} \\ y(0) = y_0 \quad \text{given in } \mathbb{R} \end{cases} \tag{5.16}$$

The solution is

$$y(t) = (y_0 - g(0)) \exp(-\lambda t) + g(t) \tag{5.17}$$

When $\lambda \gg 1$ (it is a very simple verification to see that the Lipschitz constant L (see (5.10)) is then λ), the equation in (5.16) is ill-conditioned. If $y_0 \neq g(0)$ and if g is a regular function with slow evolution, we understand that, in an initial phase, in a time interval of order $\frac{1}{\lambda}$ (the **scale time**), the solution $y(t)$ is strongly varying, it is the **transient regime** or the **boundary layer**. After a time of order $\frac{1}{\lambda}$, the contribution of the exponential term in (5.17) becomes quickly negligible and we have $y(t) \simeq g(t)$, it is the slowly varying solution , this phase is called **permanent regime** or **regular zone**. The word regular is used in the sense of the derivatives being essentially smaller than the derivatives of the transient part of the solution.

For our problem (5.16), the relation (5.9) gives

$$e_{n+1} = (1 - \lambda h_n)e_n + \epsilon_n$$

Like $e_0 = 0$, by recurrence one verify easily that

$$e_n = \sum_{l=0}^{n-1} \left(\prod_{k=l+1}^{n-1} (1 - \lambda h_k) \right) \epsilon_l \tag{5.18}$$

with the convention $\prod_{k=n}^{n-1} = 1$.

Thus the contribution to e_n, up to the time t_n, of the local error troncature ϵ_l evaluate at the time $t_l < t_n$, is equal to

$$\left(\prod_{k=l+1}^{n-1} (1 - \lambda h_k) \right) \epsilon_l .$$

In the boundary layer one has, from (5.17), $y''(t) \simeq \lambda^2$; furthermore from (5.8), the local troncature error ϵ_l behaves like $(\lambda h_l)^2$ and it is necessary to choose h_n very small (because $\lambda \gg 1$) such that the local troncature error let be weak.

In the regular zone one has, from (5.17), $y''(t) \simeq g''(t)$ and the local troncature error behaves like $(g'' h_n)^2$, like g is slowly varying g'' is small and it is a priori possible to use a greatest stepsize h_n such that ϵ_l let be small again. Unfortunately like $\lambda \gg 1$, if we are choosing a greatest timestep h_n, such that ϵ_n let be small but with $\lambda h_n \gg 1$, $\mid 1 - \lambda h \mid$ becomes very great and the errors at time $t_l > t_n$ are expanded (see (5.18)), the method becomes unstable. This condition of numerical stability imposes a severe restriction on the stepsize h (for the simplified case h = constant).

$$\mid 1 - \lambda h \mid < 1, \quad \text{let be} \quad 0 < h < \frac{2}{\lambda}$$

The **explicit Euler method is inadequate** for solving ill-conditionned differential problems.
N. B. : We remark that the h_n choice is essentially controlled, in the case of this explicit method, by numerical stability rather by accuracy. This situation is typical when applying an explicit method to a stiff problem.•

If we apply now the backward Euler method (5.12) to the equation (5.16), by (5.15) we obtain now

$$e_{n+1} = \frac{1}{1+\lambda h_n}(e_n + \epsilon_n),$$

and

$$e_n = \sum_{l=0}^{n-1} \left(\prod_{k=l}^{n-1} \frac{1}{1+\lambda h_k} \right) \epsilon_l. \tag{5.19}$$

Thus the contribution to e_n, up to the time t_n, of the local error troncature ϵ_l evaluate at the time $t_l < t_n$, is equal to

$$\left(\prod_{k=l+1}^{n-1} \frac{1}{1+\lambda h_k} \right) \epsilon_l.$$

In the boundary layer it is still necessary to choose $h_n \ll \frac{1}{\lambda}$ such that ϵ_n let be $\ll 1$. But in the regular zone it is now possible to choose h_n such that we have together

$$\lambda h_n \gg 1 \qquad \text{and} \qquad |\, \epsilon_n \,| \ll 1.$$

In effect like $\frac{1}{1+\lambda h_n}$ is by that very small, the errors at times $t_l < t_n$ are strongly damped. The **implicit Euler method is very stable and well convenient** for the numerical resolution of differential systems ill-conditionned.

This analysis concerning the Euler method is representative for using implicit or explicit method to stiff problem. **The implicit methods are highly recommended for solving stiff problems.**

Remark 11. *: If we suppose that the function f verifies the Lipschitz condition (5.10), then if $Lh_n < 1$, (5.12) has an unique solution $y_{n+1} \in \mathbb{R}^m$ and the backward Euler method is convergent.•*

5.1.6 Consistency and order for onestep methods.

Consider once more time the differential problem

$$\begin{cases} y'(t) = f(t,\ y(t)), & t \in I_0, \\ y(t_0) = \eta, & \eta \text{ given in } \mathbb{R}^m \end{cases} \tag{5.20}$$

A onestep method is defined by

$$\begin{cases} y_{n+1} = y_n + h_n \Phi(t_n\ , y_n\ ; h_n), \\ y_0 = \eta_h \qquad \text{given in } \mathbb{R}^m, \end{cases} \tag{5.21}$$

where Φ is a continuous function from $I_0 \times \mathbb{R}^m \times [0\ , h_0]$ in $\mathbb{R}^m$. Note that for the Euler-Cauchy method

$$\Phi(t\ , y\ ; h) = f(t\ , y),$$

and the function Φ is independant of h.

For the backward Euler method, we suppose that the equation (see (5.12))

$$z - hf(t+h,\ z) = y$$

admits for all $y \in \mathbb{R}^m$ an unique solution $z \in \mathbb{R}^m$ (it is true if the function f verifies the conditions of the remark 11) that we note $z = \Phi(t\ ,y\ ;h)$, then we put

$$\Phi(t\ ,y\ ;h) = f(t+h,\ \Phi(t\ ,y\ ;h)).$$

In a general sense we shall say that the onestep method (5.21) is **explicit** if the exact calculation of $\Phi(t\ ,y\ ;h)$ needs the evaluation of f for a **finite** number (t,y) values. The method is said **implicit** in the contrary case. For any solution y of the differential problem (5.1), we introduce the **local troncature error** defines by

$$\epsilon_n = y(t_{n+1}) - y(t_n) - h_n\Phi(t_n\ ,y(t_n)\ ;h_n), \tag{5.22}$$

so that

$$y(t_{n+1}) = y(t_n) + h_n\Phi(t_n\ ,y(t_n)\ ;h_n) + \epsilon_n, \tag{5.23}$$

the sequence $y(t_n)_{0\leq n\leq N}$ appears like being the numerical perturbated solution.

The order of this perturbation permits to define the order of the scheme (5.21).

Definition 8. *For the onestep method (5.21) we call order of this method the greatest integer p such that*

$$y(t+h) - y(t) - h\Phi(t\ ,y(t)\ ;h) = O(h^{p+1}), \tag{5.24}$$

this for all function f with p continous derivatives and all solution y for (5.1). The method is said ***consistent*** *if the order is $p \geq 1$.*

The onestep scheme (5.21) is of order p under a necessary and sufficient condition (see [6], p.95).

5.1.7 Stability for onestep methods.

It is a continuity result for solution with respect to datas, this uniformly with respect to $h = \max_{0\leq n\leq N-1} h_n$. One consider the perturbated system

$$z_{n+1} = y_n + h_n\Phi(t_n\ ,z_n\ ;h_n) + \eta_n \tag{5.25}$$

Definition 9. *The onestep method (5.21) is said stable if it exists a constant $M > 0$ independent of h such that*

$$\max_{0\leq n\leq N} \mid y_n - z_n \mid \leq M(\mid y_0 - z_0 \mid + \sum_{n=0}^{N-1} \mid \eta_n \mid) \tag{5.26}$$

That means, when the onestep method is stable, a “small” perturbation on the datas implies only a “small” perturbation on the solution even if h tends to zero (with many **rounding errors**).

The onestep scheme (5.21) is stable under a Lipschitz condition carrying on Φ (see [6], lemma 5.3, p.93).

5.1.8 Definition of the convergence for onestep methods.

Definition 10. *We shall say that the onestep method (5.21) is convergent if the conditions* h *tends towards zero and* η_h *tends toward* η *imply*

$$\max_{0\leq n\leq N} \mid y(t_n) - y_n \mid \qquad \textit{tends towards zero,}$$

where y(.) is the solution of (7.42) and y_n *is the solution of (5.21).*

Now taking into account this definitions, one has the following key result, almost evident to prove (see [6], p.92) : **If the onestep method (5.21) is stable and consistent then it is a convergent one**.

5.1.9 A-stability.

What are the stability properties that it is desirable that a numerical method holds ? Consider the scalar differential equation

$$y\prime = -\lambda y, \quad \lambda \in \mathbb{C} \tag{5.27}$$

the general solution $y(t) = y_0 exp(-\lambda\, t)$ remains bounded when $t \to 0$ if and only if $Re\,\lambda \geqslant 0$, and tends towards 0 when $t \to \infty$ if and only if $Re\,\lambda > 0$. It seems natural to ask that a numerical method verifies similar properties.

Consider first the Euler-Cauchy method (5.4). Applied to equation (5.27) with a **constant** stepsize h, in putting $z = \lambda\, h$ one has :

$$y_{n+1} = (1 - z)y_n$$

then

$$y_n = (1 - z)^n y_0$$

Therefore y_n remains bounded when $n \to \infty$ if and only if z belongs to the disk in the complex plan

$$D = \{z \in \mathbb{C};\ |1 - z| \leqslant 1\}.$$

One says that$D$ is the **absolute domain of stability** of the method. Remark that, if z belongs to the interior of D, $y_n \to 0$ when $n \to \infty$.

On the other hand the backward Euler method (5.12) gives :

$$y_{n+1} = \frac{1}{1 + z} y_n$$

hence

$$y_n = (\frac{1}{1 + z})^n y_0.$$

This time, y_n remains bounded if and only if z belongs to :

$$D = \{z \in \mathbb{C};\ \frac{1}{|1 + z|} \leqslant 1\}.$$

In this case, the absolute domain of stability encloses the half-plane $Re(z) \geqslant 0$; one says that the backward Euler method is **A - stable**. Moreover, if $Re(z) > 0$, $y_n \to 0$ when $n \to \infty$; this property is called **strong A-stability**. We understand now the interest to work if possible with A-stable methods.

5.2 Higher order methods : Runge-Kutta methods and multisteps linear methods

The previous Euler methods are only of the first order. In practise one uses higher order methods which present a better "quality-price profit". It exists two principal types of higher order methods which are currently used. The first type consists of the **Runge-Kutta** methods which belongs at the class of the onestep methods. The second consists of the **multisteps linear methods**.

5.2.1 Runge-Kutta methods.

These are onestep methods with order $p > 1$. For simplify this introduction for modelling the differential equations we refer to Crouzeix [6] chapter 5 for these methods.

5.2.2 Multisteps linear methods.

The use of several approached values permits to obtain methods with less higher coast. These methods are called multisteps methods : more precisely when calculation use the q previous values y_n, y_{n-1},$\cdots$, y_{n-q+1}, we are speaking about method with q steps. Such a method requires q starting values y_0, y_1,$\cdots$, y_{q-1}. To simplify we suppose that the step h_n is **constant** and egal to h in all the following. The classical multisteps methods are the **Adams** and the **backward differentiation** methods. We consider always the problem (5.1).

We obtain a large class of multistep methods by the following process : we integrate (5.1) between t_{i-m} and t_{i+n}

$$y(t_{i+n}) = y(t_{i-m}) + \int_{t_{i-m}}^{t_{i+n}} f(t,\ y(t))\,dt, \tag{5.28}$$

and we approach the integral in the second member by $\int_{t_{i-m}}^{t_{i+n}} P_q(t)\ dt$, where $P_q(t)$ is the interpolation polynomial for the function $t \to f(t,\ y(t))$ under the points t_i, t_{i-1}, $\cdots$, t_{i-q}. By noting

$$L_j(t) = \prod_{\substack{k=0 \\ k\neq j}}^{q} \frac{t - t_{i-k}}{t_{i-j} - t_{i-k}} \tag{5.29}$$

the j-th Lagrange polynomial, one has by the elementary Lagrange polynomial formula (see for instance Crouzeix [6], p.6)

$$P_q(t) = \sum_{j=0}^{q} L_j(t) f(t_{i-j}, y(t_{i-j})). \tag{5.30}$$

Remark 12. *: When $q = 0$ we do the convention $L_0(t) \equiv 1$.•*

Therefore we have the following approached formula

$$y(t_{i+n}) - y(t_{i-m}) \simeq f(t_{i-j}, y(t_{i-j})) \int_{t_{i-m}}^{t_{i+n}} L_j(t)\ dt = h \sum_{j=0}^{q} \beta_j f(t_{i-j}, y(t_{i-j})), \tag{5.31}$$

where

$$\beta_j = \frac{1}{h} \int_{t_{i-m}}^{t_{i+n}} L_j(t)\ dt, \tag{5.32}$$

let be, by the change of variable $h\ s = t - t_i$ for the integral, taking into account (5.29) (if $t = t_{i+n}$ then $s = n$ because $t_{i+n} = t_i + hn$, furthermore $t - t_{i-k} = t - t_i + (t_i - t_{i-k} = hs + (ih - (i-k)h) = h(s+k)$ and $t_{i-j} - t_{i-k} = h(k-j)$) one has :

$$\beta_j = \int_{-m}^{n} \prod_{\substack{k=0 \\ k \neq j}}^{q} \frac{s+k}{k-j}\, ds,$$

The previous relation 5.31 suggests the iterative scheme

$$y_{i+n} - y_{i-m} = h \sum_{j=0}^{q} \beta_j f(t_{i-j}, y_{i-j}) \tag{5.33}$$

Adams methods. For $n = 1$ and $m = 0$ we have the **Adams-Bashforth** methods with $q+1$ steps, i.e.

$$y_{i+1} = y_i + h \sum_{j=0}^{q} \beta_j f(t_{i-j}, y_{i-j}) \tag{5.34}$$

where

$$\beta_j = \int_{0}^{1} \prod_{\substack{k=0 \\ k \neq j}}^{q} \frac{s+k}{k-j}\, ds \tag{5.35}$$

N.B. : For $q = 0$, by convention $L_0(t) \equiv 1$ (see remark 12), then $\beta_0 = 1$; from (5.34) we find again the Euler-Cauchy method (5.4). That justifies this construction of **explicit** multisteps methods like Adams-Bashforth methods.•

When $n = 0$ and $m = 1$ we have from (5.33) the **Adams-Moulton** methods with $(q+1)$ steps, i.e.

$$y_i = y_{i-1} + h\beta_0 f(t_i, y_i) + h \sum_{j=1}^{q} \beta_j f(t_{i-j}, y_{i-j}), \tag{5.36}$$

where

$$\beta_j = \frac{1}{h} \int_{-1}^{0} \prod_{\substack{k=0 \\ k \neq j}}^{q} \frac{s+k}{k-j}\, ds. \tag{5.37}$$

We see, from (5.36), that now y_i is implicitly given.

N.B. : For $q = 0$, $L_0(t) \equiv 1$, then $\beta_0 = 1$; from (5.36) we find now the Backward Euler method (5.12). That justifies this construction of **implicit** multisteps methods like Adams-Moulton methods.•

When $n = 1$ and $m = 1$ we obtain the **Nyström** methods. In particular for $q = 0$ we have the **Milne** method

$$y_{i+1} = y_{i-1} + 2hf(t_i, y_i), \tag{5.38}$$

Remark 13. *: When $h_n = h(=$ constant $\forall n)$ (it is the framework of our presentation), it is possible to calculate by recurrence the β_j coefficients for Adams-basforth and Adams-Moulton ([6], p.115-116).•*

Backward Differentiation Formulas : BDF methods. The previous Adams methods can be inefficacious on a really stiff problem. It may be interesting to use BDF methods because they have better stability properties (see [6], chapter 7, where we can find also the following construction of BDF schemes). Suppose known the approached values y_i, y_{i-1},$\cdots$ of $y(t_i)$, $y(t_{i-1})$,$\cdots$. We suppose also, to simplify, that the step $h = t_{i+1} - t_i$ is constant $\forall i$ (in fact all the following is strictly valid if one replace h by $h_i = t_{i+1} - t_i$). To find an approximation y_{i+1} for $y(t_{i+1})$, we consider $P_{i,r}$ the Lagrange interpolation polynomial with degree $\leq r$ such that

$$P_{i,r}(t_{i+1-j}) = y_{i+1-j}\ , \qquad j = 0, 1, \cdots, r. \tag{5.39}$$

We know from (5.30) that :

$$P_{i,r}(t) = \sum_{j=0}^{r} y_{i+1-j} L_j(t), \tag{5.40}$$

with, see (5.29),

$$L_j(t) = \prod_{\substack{k=0 \\ k\neq j}}^{r} \frac{t - t_{i+1-k}}{t_{i+1-j} - t_{i+1-k}}. \tag{5.41}$$

Then we approximate (5.1) by the scheme

$$\frac{d}{dt} P_{i,r}(t_{i+1}) = f(t_{i+1}, y_{i+1}). \tag{5.42}$$

Carry out the change of variable $s = \frac{t_{i+1}-t}{h}$ and put

$$\alpha_j = \frac{h}{t_{i+1} - t_{i+1-j}} \text{ for } j \geq 1, \qquad p_{i,r}(s) = P_{i,r}(t),$$

(5.42) becomes

$$p'_{i,r}(0) = -hf(t_{i+1}, y_{i+1}). \tag{5.43}$$

If we introduce the following polynomial in s

$$\omega_{i,r}(s) = (1 - \alpha_1 s)(1 - \alpha_2 s) \cdots (1 - \alpha_r s), \tag{5.44}$$

one checks easily, by this change of variable, that (5.43) is

$$\omega'_{i,r}(0) y_{i+1} - \sum_{j=1}^{r} \frac{\alpha_j^2 \omega_{i,r}(0)}{\omega'_{i,r}(\frac{1}{\alpha_j})} y_{i+1-j} = -hf(t_{i+1}, y_{i+1}). \tag{5.45}$$

Putting

$$a^i_{j,r} = \frac{\alpha_j^2}{\omega'_{i,r}(\frac{1}{\alpha_j})\omega'_{i,r}(0)}, \qquad b^i_r = -\frac{1}{\omega'_{i,r}(0)}, \tag{5.46}$$

from (5.45), like $\omega_{i,r}(0) = 1$, the **standard form for the implicit BDF scheme** is finally :

$$y_{i+1} - hb^i_r f(t_{i+1}, y_{i+1}) = \sum_{j=1}^{r} a^i_{j,r} y_{i+1-j}. \tag{5.47}$$

Then it is useful to express (5.46). One finds, if $r > 1$

$$b^i_r = -\frac{1}{\omega'_{i,r}(0)} = \frac{1}{\sum_{j=1}^{r} \alpha_j}, \quad a^i_{j,r} = \frac{\alpha_j b^i_r}{\prod_{\substack{k=1 \\ k\neq j}}^{r} (1 - \frac{\alpha_k}{\alpha_j})}, \tag{5.48}$$

Furthermore if $r = 1$, then $\alpha_1 = \frac{h}{t_{i+1}-t_i} = 1$ and (see (5.44)) $\omega_{i,1}(s) = (1-s)$, therefore $\omega'_{i,1}(s) = -1$, $\forall s$. Finally if $r = 1$ from (5.46) one has :

$$b_1^i = 1 \qquad \text{and} \qquad a_{1,1}^i = 1. \tag{5.49}$$

Examples of BDF methods. If $r = 1$, (5.47) and (5.49) give :

$$y_{i+1} = hf(t_{i+1}, y_{i+1}) + y_i, \quad \forall i \geq 0 \tag{5.50}$$

we find again the Backward Euler method.

If $r = 2$, then $\alpha_2 = \frac{1}{2}$ (we have always $\alpha_1 = 1$), therefore by (5.48), one has : $b_2^i = \frac{2}{3}$, $a_{1,2}^i = \frac{4}{3}$ and $a_{2,2}^i = -\frac{1}{3}$. From (5.47), the associate BDF scheme is the well-known scheme with constant stepsize :

$$y_{i+1} = \frac{4}{3}y_i - \frac{1}{3}y_{i-1} + \frac{2}{3}hf(t_{i+1}, y_{i+1}), \quad \forall i \geq 1 \tag{5.51}$$

5.2.3 Any comments and properties about multisteps methods.

For an introduction to a study of Adams methods (with variable stepsize) see [6], chapter 7.

The methods (5.33) and (5.47) belong to the general framework of the multisteps linear methods

$$\alpha_k y_{i+k} + \alpha_{k-1} y_{i+k-1} + \cdots + \alpha_0 y_i = h[\beta_k f_{i+k} + \beta_{k-1} f_{i+k-1} + \cdots + \beta_0 f_i]. \tag{5.52}$$

where $f_i = f(t_i\ , y_i)$.

If $\beta_k \neq 0$ (5.52) is an implicit equation in y_{i+k}. We note that the starting of the method (5.52) requires the knowledge of the initial values $y_i = \eta_i, \quad i = 0, \cdots, k-1$. For avoid this drawback one use methods with number variable steps. We begin by a onestep method, then with two steps for calculate y_2, etc.

The method (5.52) is characterized by the coefficients $\alpha_k, \cdots, \alpha_0, \beta_k, \cdots, \beta_0$, or by the polynomials

$$\begin{cases} \alpha(t) = \alpha_k t^k + \alpha_{k-1} t^{k-1} + \cdots + \alpha_0 \\ \beta(t) = \beta_k t^k + \beta_{k-1} t^{k-1} + \cdots + \beta_0 \end{cases} \tag{5.53}$$

we suppose $\alpha_k \neq 0$ and $\mid \alpha_0 \mid + \mid \beta_0 \mid > 0$, by that the method is really with k steps.

Thanks to (5.53), after convenient definitions for consistency and stability, it exists elegant and simple sufficient and necessary conditions for polynomials $\alpha(t)$ and $\beta(t)$ which secures convergence (always taking into account the theorem : *consistency + stability* $\Longrightarrow$ *convergence*) of the multistep method (5.52) (see [20], theorems 5.5, 5.6 and 5.10 in chapter 5). Summarize the results.

Theorem 1. *The scheme (5.52) is consistent if and only if :*
$\alpha(1) = 0, \quad \alpha'(1) = \beta(1) \neq 0.\bullet$

Theorem 2. *The scheme (5.52) is stable if and only if :*
the roots of $\alpha(z)$ ($z \in \mathbb{C}$) are of modulus inferior or equal to one and those with modulus equal one are simple.•

Theorem 3. *The scheme (5.52) is convergent if and only if the conditions in theorems 1 and 2 are verified.*•

Naturally we should solve, at each time, the implicit equation (5.52). Let be L the Lipschitz condition of the function f (see (5.10)). To insures that the solution y_{i+k} is unique we suppose that $h < h_* = \frac{|\alpha_k|}{L|\beta_k|}$ then, with this hypotesis, (5.52) can be solved numerically by the convergent successive approximation method (point fixed method)

$$y_{i+k}^{(l+1)} = h\frac{\beta_k}{\alpha_k} f(t_{i+k}, y_{i+k}^{(l)}) - \sum_{j=0}^{k-1}(\frac{\alpha_j}{\alpha_k} y_{i+j} - h\frac{\beta_j}{\alpha_k} f_{i+j}),\ l = 0, 1, \cdots$$

Evidently the constant L is pratically irrelevant. That means that for h sufficiently small one can solve the equation (5.52) by any convenient method.

For instance for BDF2 we have $\alpha(z) = z^2 - \frac{4}{3}z + \frac{1}{3}$ and $\beta(z) = \frac{2}{3}z^2$ then, $\alpha(1) = 0$ and $\alpha'(1) = \frac{2}{3} = \beta(1)$ hence, by the theorem 1, BDF2 is consistent. Furthemore the roots of $\alpha(z)$ are one and $\frac{1}{3}$, by the theorem 2 this scheme is stable. Then BDF2 is convergent. Furthermore we suppose the condition $Lh < \frac{3}{2}$, by that (5.51) has an unique solution. In fact the BDF method is really optimal because one can prove that the coefficients α_i, $0 \leqslant i \leqslant q$, are determined such that the associated BDF method has the maximum order possible (i.e. order two for $q = 2$). Furthermore, we have a negative result due to Dahlquist ([7]), it says : "it don't exists method $(\alpha,\ \beta)$ A-stable with order strictly higher than 2", and one can prove that the two step methods with order 2 which are A-stables are characterized by the two conditions $\alpha_2 \geqslant \frac{1}{2}$, $\beta_2 > \frac{\alpha_2}{2}$. Like $\alpha_2 = 1$ and $\beta_2 = \frac{2}{3}$, BDF2 is A-stable. Concerning the domain of stability, BDF2 is also optimal.

5.3 Specific methods for chemistry modelling.

We exhibit here two methods which are usually used by the chimists. The method founded on a **Quasi-Stationnary State Approximation** called by that the **QSSA** method and the **Two-Step** method.

5.3.1 The QSSA method.

Consider for instance the following chemical mechanism :

$$\mathcal{M}_1 + \mathcal{X} \ \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} \ \mathcal{M}_2 + \mathcal{X} \tag{5.54}$$

$$\mathcal{M}_2 \ \xrightarrow{k_2} \ \mathcal{P} \tag{5.55}$$

where the $k's$ are the speed reaction constants and $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{X}$ and $\mathcal{P}$ are chemical species with concentrations C_1, C_2, C_X and C_P. The reaction speed for any reaction are given by the elementary kinetic law, i.e. the product of the reaction constant (for this reaction) by the concentrations of consumed species. If one suppose that the concentration C_X is constant, the dynamic of this system is linear and the system of ordinary differential equations is given by :

$$\left\{\begin{array}{l} \frac{dC_1}{dt} = -(k_1 C_X)C_1 + (k_{-1}C_X)C_2 \\ \\ \frac{dC_2}{dt} = (k_1 C_X)C_1 - (k_{-1}C_X + k_2)C_2 \end{array}\right. \tag{5.56}$$

The specie P is the product of a noninvertible reaction and don't have an effect under the evolution of species $\mathcal{M}_1$ and $\mathcal{M}_2$. The first equation in (5.56) means that the specie $\mathcal{M}_1$ is lossed (resp. producted) by the direct (resp. inverse) way of the first chemical reaction in (5.54). It is usual to note, for a specie i, the production term by P_i and loss term by L_i. For instance for the specie $\mathcal{M}_1$ we have

$$\frac{dC_1}{dt} = P_1 - L_1 C_1, \quad \text{with } P_1 = (k_{-1}C_X)C_2 \quad \text{and } L_1 = k_1 C_X. \tag{5.57}$$

Let be now the equation associated to the specie i

$$\frac{dC_i}{dt} = P_i - L_i C_i. \tag{5.58}$$

frequently the P_i and L_i terms depend on $\vec{C}$, the vector of the concentrations for all the species which constitute the chemistry mechanism. It is more appropriate to write

$$\frac{dC_i}{dt} = P_i(\vec{C}) - L_i(\vec{C})C_i. \tag{5.59}$$

All the system can be written in the nonlinear form

$$\frac{d\vec{C}(t)}{dt} = f(t,\ \vec{C}(t)) = P(t,\ \vec{C}(t)) - L(t,\ \vec{C}(t))\vec{C}(t). \tag{5.60}$$

where $\vec{C}(t) = (C_1(t),\ C_2(t), \cdots, C_m(t))^T$, $P(t,\ \vec{C}(t))$ is a vector and $L(t,\ \vec{C}(t))$ is a diagonal matrix.

Now we suppose that all the concentrations are known at time t. Let be a time interval $[t,\ t+\Delta t]$. The principle of the QSSA method (see [21]) consists to suppose that during the stepsize Δt the quantities $P_i(\vec{C})$ and $L_i(\vec{C})$ are constants (it is the quasi-stationnary state approximation valid for Δt sufficiently small) and equals respectively to P_i and L_i. Taking into account this assumption, for all specie i, (5.59) becomes :

$$\frac{dC_i}{dt} = P_i - L_i C_i, \tag{5.61}$$

which is easy to solve explicitly between times t and $t + \Delta t$. One finds

$$C_i(\bar{t}) = \frac{P_i}{L_i} + \lambda \exp(-L_i \bar{t}), \quad \bar{t} \in [t,\ t+\Delta t], \tag{5.62}$$

where λ is a constant. Let be $C_{ie} = \frac{P_i}{L_i}$ the **equilibrium concentration** associated to $\frac{dC_i}{dt} = 0 = P_i - L_i C_{ie}$, then from (5.62) with $\bar{t} = t + \Delta t$ one has

$$C_i(t+\Delta t) = C_{ie} + \lambda \exp(-L_i t) \exp(-L_i \Delta t),$$

finally, always from (5.62) with $\bar{t} = t$, the solution of (5.61) is :

$$C_i(t+\Delta t) = C_{ie} + (C_i(t) - C_{ie}) \exp(-L_i \Delta t), \tag{5.63}$$

We can now, from (5.63), define an characteristic reaction time for specie i, let be $\tau_i = \frac{1}{L_i}$. Knowing this characteristic times it is possible to class the species in three categories :
- the **slow species** for which $\tau_i \geq 100\Delta t$,
- the **middle species** for which $\frac{\Delta t}{10} \leq \tau_i \leq 100\Delta t$,
- the **quick species** for which $\tau_i \leq \frac{\Delta t}{10}$.

How it runs ? For the **quick species**, supposed to be almost instantaneously at equilibrium, (5.61) becomes

$$0 = P_i(t + \Delta t) - L_i(t + \Delta t)C_i(t + \Delta t)$$

and we use the approximation

$$C_i(t + \Delta t) = C_{ie}(t + \Delta t) = \frac{P_i(t + \Delta t)}{L_i(t + \Delta t)} \tag{5.64}$$

But during the transient periods, i.e. at the beginning of the simulation or at sunset and sunrise any number of species quickly develop between an equilibrium with photochimical effect and an equilibrium without photochimical effect. Then it is necessary to a small stepsize time (in order to 1s). Furthermore the quick species highly coupled with other species ask a particular treatment because the P_i and L_i terms change inside the steptime, the quasi-stationnary state assumption is not valid. For avoid this problem we do some iterations for the resolution of (5.61) for a quick specie, inside the stepsize $\Delta t \approx 1s$, with actualization of quick species concentrations in P_i and L_i terms. Practically in five iterations the quick species are in equilibrium. During this transient periods we don't use (5.63).

For the **middle species**, we use explicitly the solution given by (5.63).

Finally for the **slow species**, the associated concentration C_i is sufficiently slow such that a linear approximation for (5.64) is permissible (remember that the evaluation of an exponential term is expensive) during a stepsize time. One decides to do the approximation $\exp(-L_i\Delta t) = I - L_i\Delta t$. Therefore, if $\tau_i \geq 100\Delta t$, one evaluate $C_i(t + \Delta t)$ by :

$$C_i(t + \Delta t) = C_i(t) + (C_{ie} - C_i(t))L_i\Delta t. \tag{5.65}$$

Computational errors connected with QSSA methods are difficult to assess.

5.3.2 A particular case of multistep method : the two-step method :

Founded on a implicit BDF formula of order two (similar to (5.51)) solved by a Gauss-Seidel iteration.

For $r = 2$, with **variable stepsize**, the BDF formula is written (see (5.47))

$$y_{i+1} - h_i b_2^i f(t_{i+1}, y_{i+1}) = \sum_{j=1}^{2} a_{j,2}^i y_{i+1-j} \tag{5.66}$$

where $a_{j,2}^i$ (for $j = 1, 2$) and b_2^i are given by (5.48).

Putting $c = \frac{t_i - t_{i-1}}{t_{i+1} - t_i}$, by elementary calculations, one checks that $b_2^i = \frac{c+1}{c+2}$, $a_{1,2}^i = \frac{(c+1)^2}{c^2+2c}$ and $\frac{(c+1)^2}{c^2+2c} = -\frac{1}{c^2+2c}$.

If we put away the explicit part, by the previous expressions of $a_{1,2}^i$ and $a_{2,2}^i$, from (5.66) one has :

$$Y_i = \sum_{j=1}^{2} a_{j,2}^i y_{i+1-j} = \frac{(c+1)^2 y_i - y_{i-1}}{(c^2 + 2c)}. \tag{5.67}$$

Hence, if we put $\gamma = b_2^i$,

$$y_{i+1} - h_i \gamma f(t_{i+1}, y_{i+1}) = Y_i. \tag{5.68}$$

we find the formula given in the original work of J.G. Verwer ([38], formulas (2.1) and (2.2)), with the same expressions with respect to c for γ and Y_i.

We emphasize that second order is sufficient because for reactive flow problems a low level of accuracy of, says, 1% is good enough. A higher level is thought to be superfluous, due to errors made in other (operator splitting) processes and uncertainties in the reaction constants of the chemistry models.

The Gauss-Seidel method exploits the chemical kinetics form (5.60), by which (5.68) can be written as

$$y_{i+1} = Y_i + h_i\gamma[P(t_{i+1}, y_{i+1}) - L(t_{i+1}, y_{i+1})y_{i+1}],$$

or

$$y_{i+1} = F(y_{i+1}) = (I + h_i\gamma L(t_{i+1}, y_{i+1}))^{-1}(Y_i + h_i\gamma P(t_{i+1}, y_{i+1}), \tag{5.69}$$

where index means that the unknown values of the concentrations are taken at time t_{i+1}. That is the appropriate formulation for a straightforward application of Gauss-Seidel iteration to the nonlinear system of equations $y = F(y)$ because, like the matrix L is diagonal, the matrix $(I + h_i\gamma L(t_{i+1}, y_{i+1}))$ is diagonal and immediatly nonsingular, that makes this process essentially an explicit one.

For simplify we denote by $y = (y_{i+1}) \in \mathbb{R}^m$ the vector of unknown concentrations at time t_{i+1}, and let be $y = (y_1, y_2, \cdots, y_k, \cdots, y_m)$ the components of y. The **Gauss-Seidel** method consists to calculate, when $y_m^{(l)}$, $y_{m-1}^{(l)}$, $\cdots$, $y_{k+1}^{(l)}$, $y_k^{(l)}$, $y_{k-1}^{(l+1)}$, $\cdots$, $,y_2^{(l+1)}$, $y_1^{(l+1)}$ are known, $y_k^{(l+1)}$ by :

$$y_k^{(l+1)} = F(y_1^{(l+1)}, y_2^{(l+1)}, \cdots, y_{k-1}^{(l+1)}, y_k^{(l)}, y_{k+1}^{(l)}, \cdots, y_m^{(l)}), \tag{5.70}$$

then by a similar formula we calculate $y_{k+1}^{(l+1)}$ until $y_m^{(l+1)}$. After, if necessary, we pass to the $(l+2)$-iteration for Gauss-Seidel.

N.B. : The two-step method means principally that we calculate y_{i+1} from y_i and y_{i-1}, these are the two-steps ! After we use the Gauss-Seidel method for solving the implicit equation (5.68). Due to the special form of the chemical modeling the Gauss-Seidel method is very convenient for solving this implicit equation. In this particular situation it is desirable to rename the BDF method of order two by two-step.•

Note that this technic can be implemented in a completely similar way into a diagonally implicit Runge-Kutta code, says, into one or order with two stages.

6 One example of the modelling of the air pollution problem : the CHIMERE software.

It is not possible to give a description of all the models in this field. Mainly I roughly describe CHIMERE-continental, a newly developed European scale Eulerian chemistry transport model developed at *Laboratoire de Météorologie Dynamique, Ecole Polytechnique, Palaiseau, France* which, I thing, gives a good idea of the problems encountered and associated solutions and mechanisms for solve it. A comparison between models should be desirable but, following the objectives and associated constraints the modelling ought be adapted, and this comparison is not easy to do, any one has its field of interest. By that the description of one model can done more precise information than a survey.

Recall that the mass continuity equation for the concentrations of chemical species in the air pollution modeling is (1.1). Independtly of the numerical resolution and meteorological input, the major steps of the modelization concerns the chemical mechanism, emissions, deposition, together with the advection, diffusion and boundary layer modelling.

6.1 Chemical mechanism.

Naturally it is desirable that a model can include different gas phase chemical mechanisms. For instance CHIMERE offers this possibility but usually use the original scheme MELCHIOR (Lattuati [22]), an updated and extended version of the EMEP mechanism (Simpson [31]) suited to polluted as well as unpolluted environments. MELCHIOR describes more than 200 reactions of 80 gaseous species. In MELCHIOR the hydrocarbon degradation is fairly similar to the EMEP gas phase mechanism and represents the major pathways of ozone build-up from industrial, traffic or biogenic (isoprene) emissions, and it has been used for many pollutions studies (Simpson [31], Simpson [32]). Six hydrocarbon species (ethane, butane, ethene, propene, isoprene, and O-xylene) represent alkane, alkene and aromatic compounds. Their oxidation by OH, O_3, yields 11 oxygenated compounds, including formaldehyde and higher aldehydes, ketones, bicarbonyls, unsaturated carbonyls and alcohols. Several adaptations have been made with respect to this mechanism, in particular for low $N0_x$ conditions at nighttime for instance. All rate constants have been updated from Atkinson and al. [1] and De Moore and al. [9]. Heteregeneous reactions are not included, but the effect of the clouds on the photolysis frequency is taken into account. A list of the model species and of the complete set of chemical reactions is presented on the internet (http ://euler.polytechnique.fr/chimere).

6.2 Emissions in CHIMERE.

The estimation of emissions and depositions is a physical, technical and difficult evaluation problem : a lot af data are missing together in space, in time, for species ; there are also uncertainties and it is evidently an important cause of errors in the models for pollution. A effort will be made in the future for improve these datas and their use.

CHIMERE requires input emission data for 16 model species. Annual data for anthropic emissions for the four classes NOx, SO_2, CO and nonmethane volatile organic compounds (NMVOC) are taken from the EMEP data base for 1997 (see the web site http ://www.emep.int). NMVOC emissions have been to be split into 12 classes represented within the models chemical mechanism. To this purpose they are first distributed for each country into different broad activity sectors (traffic, solvents, industrial and residential combustion, others), according to data prepared by IER (Institute for Energy Economics and Rational Use of Energy) in the framework of the EUROTRAC/GENEMIS project (GENEMIS [10]). Second, for each sector, NMVOC emissions are split into 32 classes with similar structure and reactivity, following a classification of Middleton and al. [25], and using VOC profiles again from IER. Third, VOCs from these 32 classes are aggregated into the 12 classes represented within CHIMERE, by applying mass and reactivity weighting as proposed by Middleton and al. [25]. It is clear that the uncertainty in the spatial distribution of individual VOC emissions is large, as VOC profiles are the same all over Europe, with the exception of traffic, where national differences (for example, distribution gasoline/diesel) are taken into account. Monthly, daily and hourly variations of the emissions are modeled by imposing variations available from the GENEMIS [10] data base. For biogenic emissions of isoprene and terpernes in CHIMERE we refer to Schmidt and al. [28].

6.3 Deposition in CHIMERE.

Dry deposition is considered for 23 of the model species i and is parametrized as a downward flux $F_{d,i} = -u_{d,i}c_i$ out of the lowest model layer with c_i being the concentration of species in this layer, and $u_{d,i}$ is the deposition velocity. This velocity is, as commonly, described through a resistance analogy (Wesley and Hicks [39]) by

$$u_{d,i} = \frac{1}{R_a + R_{b,i} + R_{c,i}}$$

The aerodynamic $R_a = (c_D v)^{-1}$ is calculated from the drag coefficient for momentum c_D (see [39]) and the wind speed v in the lowest layer.

According to Hicks and al [11] the quasi-laminar resistance $R_{b,i}$ is calculated as

$$R_{b,i} = \frac{2}{ku_*}\left(\frac{Sc_i}{Pr}\right),$$

where Sc_i and Pr are respectively the Schmidt and Prandtl number, k is the von Karman's constant and the friction velocity u_* is inferred from the meteorological data by means of the Louis [24] parametrization.

Following Erisman and al. [13] the surface resistance $R_{c,i}$ can be calculated from a number of different other resistances accounting mainly for stomatal and surface processes which are again dependent on the land use type, taken from the RIVM data base ([34]). For the exact calculation of $R_{c,i}$ the reader is referred to [13]. For more informations see [28].

6.4 Advection, diffusion and boundary layer modelling

Recall that the wind is given. The ECMWF (European Centre for Medium-Range Weather Forecasts) data are used for CHIMERE. Only the horizontal advection is calculated with piecewise parabolic method (PPM) or with a simple first order approach (the original Godunov method) depending of species. Contrary to the horizontal wind components, vertical wind is not taken directly from the meteorological input, but calculated to enforce mass balance. This is necessary as the vertical resolution of the models for meteorology and pollution are often different (that is the case for the ECMWF and CHIMERE grids). Vertical mass fluxes are calculated with a first order approach (the original Godunov method). Furthermore the lateral diffusion is often neglected as it is commonly done in mesoscale models with horizontal resolution of several tens of kilometers (see Chang and al. [2]). Only the vertical diffusion is parametrized with an eddy diffusion approach, as indicated in equation (1.1). Meanly the pollution essentially exists within the boundary layer or mixed layer, between the top of the surface layer (between the ground at 0 m and about 50 m) and a maximum height of 2500 m. The vertical eddy diffusion K (depending of z) is implemented, following a profile approach after O'Brien [27] by :

$$K(z) = K_{z_i} + \frac{(z_i - z)^2}{(z_i - z_s)^2}[K_{z_s} - K_{z_i} + (z - z_s)(\frac{\partial K}{\partial z}(z_s) + 2\frac{K_{z_s} - K_{z_i}}{z_i - z_s})]$$

Due to the bulk Richardson number, the height z_i of the boundary layer is defined. The height of the surface layer is assumed to be equal to the lowest model layer (i.e. about 50 m). K_{z_i}, the diffusion coefficient at the top of the boundary layer, is fixed at 0.05 $m^2 s^{-1}$ and considered valid in the free atmosphere (above the mixed layer). The diffusion coefficient at the top of the surface layer K_{z_s} and the respective gradient $\frac{\partial K}{\partial z}(z_s)$ are parametrized as linear functions of the height of the boundary layer only. The fluxes resulting from the vertical diffusion are calculated with a first order spatial differentiation.

The bulk Richardson number $R(z)$ is defined as

$$R(z) = g\frac{(\theta(z) - \theta_0)(z - z_0)}{\theta(z)(\tilde{V}(z) + \tilde{V}_{min})}$$

where z_0 is the altitude of the first ECMWF model level (about 35 m), $\tilde{V}_{min} = 0.1ms^{-1}$ is a minimal wind speed in order for the Richardson number not to go to infinity, $\tilde{V}(z) = (u(z), v(z))$ is the horizontal wind profile in z and $\theta(z)$ denote the vertical profile of virtual potential temperature. Finally $\theta_0 = \theta(z_0) + \theta\prime$ is assumed to represent the virtual potential temperature at the base of the thermals ; here $\theta\prime$ is a ramdom temperature departure with exponential distribution and mean 1 K. We calculate a boundary layer height z_i as the altitude at which $R(z) = 0.15$. The modelization of this height together with the vertical eddy diffusion K is also an uncertainty problem. For more details we refer to [28] and Vautard and al. [37].

6.5 The model in 3D.

Here we emphasize only on the link between the modelling of the advection equation and the modelling of the chemical differential equations. In CHIMERE the second-order two-step algoritm is used. Recall it (see (5.69)) :

$$y_{i+1} = F(y_{i+1}) = (I + h_i \gamma L(t_{i+1}, y_{i+1}))^{-1}(Y_i + h_i \gamma P(t_{i+1}, y_{i+1})),$$

in this formulation P indicates the chemical production (such that we define it, for instance, in (5.57)) and **emissions**, and L indicates the chemical loss (see (5.57)) and **deposition**.

In the last equation, one has the following definitions (see the two-step method in subsubsection 5.3.2) : $\gamma = b_2^i = \frac{1}{1+\alpha_2}$ with $\alpha_2 = \frac{t_{i+1}-t_i}{t_{i+1}-t_{i-1}}$ and (see (5.67)) $Y_i = \frac{(c+1)^2 y_i - y_{i-1}}{(c^2+2c)}$ with $c = \frac{t_i - t_{i-1}}{t_{i+1}-t_i}$. In the case of constant stepsize time we easily verify that $\gamma = \frac{2}{3}$ and $Y_i = \frac{4y_i - y_{i-1}}{3}$ because $c = 1$. Therefore we have

$$y_{i+1} = (I + h_i \frac{2}{3} L(t_{i+1}, y_{i+1}))^{-1}(\frac{4}{3}y_i - \frac{1}{3}y_{i-1} + h_i \frac{2}{3} P(t_{i+1}, y_{i+1})), \tag{6.1}$$

we find again, near notations, the formula (8) in [28].

In (6.1), the previous production and loss terms P and L are now replaced by the modified terms $\tilde{P} = P + P_h + P_v$ and $\tilde{L} = P + L_h + L_v$. By that we take into account the 3D transport term $div(\vec{V}\vec{C})$ and the 3D eddy diffusion term $div(\mathbf{K}\vec{C})$. P_h and P_v denote the temporal evolution of the concentrations due to horizontal (only advection) and vertical (advection and diffusion) inflow into the concerned cell, L_h and L_v the temporal evolution due to the respective outflow divided by the concentration itself.

The question is now : how we determine P_h and L_h ? Like we are working in 2D we use the following conservative scheme (see (4.18))

$$A^{n+1} = \left(\mathrm{Id} + \Phi_u + \Phi_v \right) (A^n) ,$$

with

$$\Phi_u(A^n) = -\frac{\Delta t}{\Delta x_i} \left(\tilde{F}_{i+\frac{1}{2},j}(A^n) - \tilde{F}_{i-\frac{1}{2},j}(A^n) \right)_{(i,j)} ,$$

and

$$\Phi_v(A^n) = -\frac{\Delta t}{\Delta y_j} \left(\tilde{F}_{i,j+\frac{1}{2}}(A^n) - \tilde{F}_{i,j-\frac{1}{2}}(A^n) \right)_{(i,j)} .$$

The term P_h (resp.L_h) sum the positive (resp. negative) terms (the inflow (resp. outflow) term) in $-\tilde{F}_{i+\frac{1}{2},j}(A^n)$, $\tilde{F}_{i-\frac{1}{2},j}(A^n)$, $-\tilde{F}_{i,j+\frac{1}{2}}(A^n)$ and $\tilde{F}_{i,j-\frac{1}{2}}(A^n)$. Finally we estimate these four terms by one (evidently the same for these four terms) of the methods seen in section 3.3.

We can find a more complete description and evaluation of this model designed for seasonal simulations and real time forecasts without the use of super-computers in [28], where details about the implementation and evaluations of the modeling are given. The reader can also complete its information, and appreciated validation of the CHIMERE model, in [37] for forecasts of the ozone over Paris area and in Menut and al. [26] for sensitivity studies of photochemical pollution using the adjoint of this simplified chemistry-transport model.

7 Appendix

7.0.1 Direct proof of the convergence of the one-sided scheme (3.16) for a regular solution.

We are to do the demonstration of the convergence of the upwind scheme for illustrate the technical proceed. In this very simple case it is possible to do it without any particular difficulty. We shall see how the Courant number occurs.

Theorem 4. *One suppose that* $a_0 \in \mathcal{C}^2(\mathcal{R})$ *is periodic in time with period L, we note a the exact solution of the problem (3.1). Let* (a_i^n) *be the numerical solution obtained by the one-sided scheme (3.16) with* $(a_i^0) = a_0(x_i)$ *for all index i. We suppose also that* $|\nu| = \left|u\frac{\Delta t}{\Delta x}\right| \leq 1$*. Then, for all* $T > 0$*, there exists one constant* $C_T > 0$ *(only dependant of* a_0 *and T) such that*

$$\sup_{1\leq i\leq M, 0\leq t^n\leq T} |a_i^n - a(x_i,t^n)| \leq C_T(\Delta t + \Delta x).\bullet \tag{7.1}$$

Proof We do it in the case $u > 0$ and for the left one-sided scheme. Since the exact solution given by $a(x,t) = a_0(x-ut)$ belongs to the class $\mathcal{C}^2$ on $\mathcal{R}\times\mathcal{R}_+$, we have from (3.30) :

$$\begin{aligned}\frac{a(x_i,t^{n+1}) - a(x_i,t^n)}{\Delta t} &+ u\frac{a(x_i,t^n) - a(x_{i-1},t^n)}{\Delta x} \\ &= \frac{\Delta t}{2}a_{tt}(x_i,t^n+\lambda_i^n) - u\frac{\Delta x}{2}a_{xx}(x_i-\mu_i^n,t^n),\end{aligned} \tag{7.2}$$

with $0\leq\lambda_i^n\leq\Delta t, 0\leq\mu_i^n\leq\Delta x$. Call ϵ_i^n the right member of (7.2) and define $K = \sup_{0\leq x\leq L}|(a_0)''(x)|$. Like $a_{xx}(x,t) = (a_0)''(x-ut)$, $a_{tt}(x,t) = u^2(a_0)''(x-ut)$, we obtain the majoration :

$$\epsilon_i^n \leq \frac{|u|\,K}{2}\Delta x + \frac{u^2K}{2}\Delta t. \tag{7.3}$$

Now introduce the numerical error $e_i^n = a_i^n - a(x_i,t^n)$. By difference between (3.30) and (7.2) we obtain :

$$\frac{e_i^{n+1}-e_i^n}{\Delta t} + u\frac{e_i^n - e_{i-1}^n}{\Delta x} = -\epsilon_i^n, \tag{7.4}$$

therefore

$$e_i^{n+1} = (1-\nu)e_i^n + \nu e_{i-1}^n - \Delta t\epsilon_i^n, \tag{7.5}$$

where ν is still the Courant number. For $n\geq 0$, denote $E^n = \sup_{1\leq i\leq M}|e_i^n|$. As (here appears the hypotesis about ν) $1-\nu\geq 0$ and $\nu\geq 0$, from (7.3) and (7.5) we deduce :

$$\left|e^{n+1}\right| \leq E^n + \Delta t(\frac{|u|\,K}{2}\Delta x + \frac{u^2K}{2}\Delta t), \tag{7.6}$$

and

$$E^{n+1} \leq E^n + \Delta t\frac{|u|\,K}{2}(1+|u|)[\Delta t + \Delta x). \tag{7.7}$$

As $E^0 = 0$ (because $a_i^0 = a_0(x_i)$ for all index i), we deduce easily from (7.7) that :

$$E^n \leq n\Delta t\frac{|u|\,K}{2}(1+|u|)(\Delta t+\Delta x), \tag{7.8}$$

which is exactly the desired inequality (7.1), with $C_T = \frac{|u|KT}{2}(1+|u|).\bullet$

7.0.2 Analyse de Fourier. Stability l^2 and convergence in this norm.

The previous proof, can be less easy for more complicate scheme. Generally to prove the convergence of a scheme it is useful to use the fundamental Lax theorem (see for instance [23] p.107). This fundamental convergence theorem for linear difference methods says that for a *consistent* linear method, *stability* is necessary and sufficient for *convergence.* Noting

$$e^n = \left(e_j^n\right)_{1\leq j\leq M} \qquad \text{and} \qquad e_j^n = a_j^n - a(x_j, t^n).$$

We can first define the l^∞-norm by

$$\| e^n \|_\infty = \sup_j \left|e_j^n\right|.$$

Ideally one might hope for convergence in this norm but it is unrealistic in approximating discontinuous solutions. The pointwise error in the neighbooring of a discontinuity typically does not go to zero uniformly as the grid is refined even though the numerical results may be perfectly satisfactory. So a more popular norm is the l^2-norm define by

$$\| e^n \|_2 = (\Delta x \sum_{j=1}^{M} \left|e_j^n\right|^2)^{\frac{1}{2}}.$$

What follows show the l^2-stability for the two-level schemes (in time).

All begin which the following result. Consider for instance the matrix A defines by :

$$\mathcal{A} = \begin{pmatrix} \beta & \gamma & 0 & \dots & 0 & \alpha \\ \alpha & \beta & \gamma & \ddots & \ddots & 0 \\ 0 & & \ddots & \ddots & \ddots & \vdots \\ \vdots & \dots & & \beta & \gamma & 0 \\ 0 & & & \alpha & \beta & \gamma \\ \gamma & 0 & \dots & 0 & \alpha & \beta \end{pmatrix} \tag{7.9}$$

This matrix is said *circulating* for all real numbers α, β and γ. Observe that all the previous two-level schemes can be write under the formulation $a^{n+1} = \mathcal{A}a^n$. Then we have :

Proposition 2. *Let be a circulating matrix $\mathcal{A}$ of the order M. Then $\mathcal{A}$ is diagonalizable and has for eigenvectors the vectors X_k (for $1 \leq k \leq M$) with coordinates :*

$$(X_k)_j = \frac{\exp\left(\frac{2ijk\pi}{M}\right)}{\sqrt{M}} \quad for \quad 1 \leq j \leq M.\bullet \tag{7.10}$$

Proof Introduce the following circulating matrix of the order M :

$$E = \begin{pmatrix} 0 & 1 & 0 & \dots & & 0 \\ 0 & 0 & 1 & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & & & \vdots \\ 0 & \dots & & & & 1 \\ 1 & 0 & \dots & & 0 & 0 \end{pmatrix} \tag{7.11}$$

It is a permutation matrix, and one verifies easily that :

$$E^{-1} = \begin{pmatrix} 0 & 0 & 0 & \dots & & 1 \\ 1 & 0 & 0 & \ddots & \ddots & 0 \\ 0 & 1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & & & \vdots \\ 0 & \dots & & & & \vdots \\ 0 & 0 & \dots & & 1 & 0 \end{pmatrix} \tag{7.12}$$

Calculating all the powers of E, we see (it is an immediate verification) :

$$E^2 = \begin{pmatrix} 0 & 0 & 1 & \dots & & 0 \\ 0 & 0 & 0 & 1 & 0 & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & & & 1 \\ 1 & 0 & & & & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \end{pmatrix} \tag{7.13}$$

it is also a circulating matrix : the diagonals of the matrix E are shift to the right. By recurrence on the powers of E we verify that E^{M-1} is precisely the same that E^{-1} and therefore $E^M = I$ (I being the identity matrix of the order M). Hence any circulating matrix is written simply :

$$\mathcal{A} = \mathcal{A}_{11} I + \mathcal{A}_{12} E + \cdots + \mathcal{A}_{1M} E^{M-1}, \tag{7.14}$$

where $\mathcal{A}_{1j}$ for $1 \leq j \leq M$ are the elements of the first line of $\mathcal{A}$.

Now it suffices to verify that the M eigenvectors define by (7.10) are the eigenvectors of E (we know evidently that an eigenvector of E is also an eigenvector of E^j for $2 \leq j \leq M-1$), indeed one has :

$$EX_k = \lambda_k X_k \qquad \lambda_k = exp\left(\frac{2ik\pi}{M}\right). \tag{7.15}$$

The M eigenvalues $(\lambda_k)_{1 \leq k \leq M}$ of the matrix E being distinct, this matrix (hence all circulating matrix $\mathcal{A}$) is diagonalizable.•

For instance we see that (7.9) can be written :

$$\mathcal{A} = \beta I + \gamma E + \alpha E^{M-1} = \beta I + \gamma E + \alpha E^{-1}. \tag{7.16}$$

From (7.14) (or in the particular case (7.16)), we can determine the eigenvalues of $\mathcal{A}$ taking into account the eigenvalues of E given by (7.15), the eigenvalues of E^{-1} being evidently inverse of the previous.

The previous proposition engage us to work in the eigenvectors basis $(X_k)_{1 \leq k \leq M}$ (that we call the *Fourier modes*). For all $a = (a_j)_{1 \leq j \leq M} \in \mathcal{C}^M$, we can write $a = \sum_{k=1}^{M} b_k X_k$ and the coordinates $(b_k)_{1 \leq k \leq M}$ in the eigenvector basis are connected to the coordinates $(a_j)_{1 \leq j \leq M}$ by a "*discrete Fourier transform*", from (7.10) one has directly :

$$a_j = \frac{1}{\sqrt{M}} \sum_{k=1}^{M} b_k \exp\left(\frac{2ijk\pi}{M}\right) \quad for \quad 1 \leq j \leq M. \tag{7.17}$$

Now it is an algebraic exercise to verify, like $(X_k)_{1\leq k\leq M}$ is an orthonormalized basis : if $a=(a_j)_{1\leq j\leq M}$ and $b=(b_k)_{1\leq k\leq M}$ are two vectors in $\mathcal{C}^M$ verifying (7.17), then b is given by the *discrete inverse Fourier transform* :

$$b_k=\frac{1}{\sqrt{M}}\sum_{j=1}^{M}a_j\exp\left(-\frac{2ijk\pi}{M}\right)\quad for\quad 1\leq k\leq M, \tag{7.18}$$

and

$$\sum_{k=1}^{M}|b_k|^2=\sum_{j=1}^{M}|a_j|^2\,. \tag{7.19}$$

Now it is easy to study the convergence of the implicit or explicit schemes in the l^2 norm. Begin first by the left one-sided scheme (3.10).

Proposition 3. *When $0\leq\nu\leq 1$, the left one-sided scheme (3.10) converge in the l^2 norm : if the solution of (3.1) is sufficiently regular, then for all $T>0$:*

$$\lim_{\Delta x,\Delta t\to 0}\left(\sup_{t^n\leq T}\parallel e^n\parallel_2\right)=0, \tag{7.20}$$

by noting $e^n=\left(e_j^n\right)_{1\leq j\leq M}$ and $e_j^n=a_j^n-a(x_j,t^n)$.•

Proof From (3.10) and (7.16), it is here where the periodicity condition ($a_0^n=a_M^n$) of the problem is important, the left one-sided scheme can be written $a^{n+1}=\mathcal{A}a^n$ with :

$$\mathcal{A}=\begin{pmatrix}1-\nu & 0 & 0 & \dots & 0 & \nu\\ \nu & 1-\nu & 0 & \ddots & \ddots & 0\\ 0 & & \ddots & \ddots & \ddots & \vdots\\ \vdots & \dots & & 1-\nu & 0 & 0\\ 0 & & & \nu & 1-\nu & 0\\ 0 & 0 & \dots & 0 & \nu & 1-\nu\end{pmatrix} \tag{7.21}$$

hence $\mathcal{A}=(1-\nu)I+\nu E^{-1}$ and, from (7.15), the eigenvalues of $\mathcal{A}$ are :

$$\lambda_k=1-\nu+\nu exp\left(-\frac{2ik\pi}{M}\right)\quad for\quad 1\leq k\leq M. \tag{7.22}$$

An easy calculation carries away :

$$|\lambda_k|^2=1+2\nu(\nu-1)\left(1-\cos\left(\frac{2k\pi}{M}\right)\right), \tag{7.23}$$

then, like by hypotesis $0\leq\nu\leq 1$, one has $|\lambda_k|^2\leq 1$ for all index k.

By decomposition of a^n in the basis $(X_k)_{1\leq k\leq M}$ such that $a^n=\sum_{k=1}^{M}b_k^nX_k$, we see that :

$$a^{n+1}=\mathcal{A}a^n=\sum_{k=1}^{M}b_k^n\mathcal{A}X_k=\sum_{k=1}^{M}b_k^n\lambda_kX_k \tag{7.24}$$

like we have also $a^{n+1} = \sum_{k=1}^{M} b_k^{n+1} X_k$, by identification one finds $b_k^{n+1} = \lambda_k b_k^n$. But $|\lambda_k|^2 \leq 1$ for all index k, consequently we deduce the following inequalities :

$$\left|b_k^{n+1}\right|^2 \leq |b_k^n|^2 \leq \cdots \leq \left|b_k^0\right|^2 \quad \forall k. \tag{7.25}$$

Finally by (7.19) (these being independant of time t^n) we can conclude in writing :

$$\Delta x \sum_{j=1}^{M} \left|a_j^n\right|^2 = \Delta x \sum_{k=1}^{M} |b_k^n|^2 \leq \Delta x \sum_{k=1}^{M} \left|b_k^0\right|^2 = \Delta x \sum_{k=1}^{M} \left|a_k^0\right|^2 \tag{7.26}$$

it is exactly the l^2 stability and by the Lax theorem the left one-sided scheme converge, i.e. one has the inequality (7.20).•

Remark 14. *The key point of the l^2 stability is the proof of the inequalities $|\lambda_k|^2 \leq 1$ for all index k.•*

l^2 unstability of the centered scheme. For the centered scheme these inequalities are not satisfied whatever the ν value. This scheme can be written (see (**??**)) :

$$a_i^{n+1} = a_i^n - \frac{\nu}{2} a_{i+1}^n + \frac{\nu}{2} a_{i-1}^n$$

Hence $\mathcal{A} = I - \frac{\nu}{2}E + \frac{\nu}{2}E^{-1}$ and, from (7.15), the eigenvalues of $\mathcal{A}$ are :

$$\lambda_k = 1 - \frac{\nu}{2} exp\left(\frac{2ik\pi}{M}\right) + \frac{\nu}{2} exp\left(-\frac{2ik\pi}{M}\right), \quad for \quad 1 \leq k \leq M, \tag{7.27}$$

then

$$\lambda_k = 1 - i\nu \sin\left(\frac{2k\pi}{M}\right), \quad for \quad 1 \leq k \leq M, \tag{7.28}$$

and

$$|\lambda_k|^2 = 1 + \nu^2 \sin^2\left(\frac{2k\pi}{M}\right). \tag{7.29}$$

Like $|\lambda_k|^2 \geq 1$ $\forall k$ and $\forall \nu$ the centered scheme is always l^2 unstable.•

Remark 15. *We verify that if $|\nu| < 1$ the centered scheme is unstable. This necessary CFL condition is well not sufficient for stability. This justify the comment about the CFL condition at the end of the subsection 3.1.1.•*

Remark 16. *The previous l^2 stability analysis is valid for the two-level schemes in time. For the three-level schemes the analysis is a little more technic.•*

Schemes	l^2- stability
Left One sided (3.10)	Yes, only if $u \geq 0$ and $0 \leq \nu \leq 1$
Centered (3.18)	No, whatever the ν value , except $\nu = 0$
Lax-Friedrichs (3.19)	Yes, under the condition $\|\nu\| \leq 1$
Lax-Wendroff (3.23)	Yes, under the condition $\|\nu\| \leq 1$
Centered Implicit (3.26)	Yes, whatever the ν value

7.0.3 Numerical experiments.

Consider the following problem where $u = 1$:

$$\left\{ \begin{array}{l} a_t + a_x = 0 \quad \text{for} \quad (x,\ t) \in \mathcal{R} \times \mathcal{R}_+ \\ a(x,\ 0) = \left\{ \begin{array}{lll} 0 & \text{if} & 0 \leq x \leq 0.3 \\ 1 & \text{if} & 0.3 \leq x < 0.7 \\ 0 & \text{if} & 0.7 \leq x < 1 \end{array} \right. \\ a(x,\ 0) = a(x+1,\ 0) \quad \forall x \in \mathcal{R} \end{array} \right. \tag{7.30}$$

which a_0 is periodic (evidently in space) with period 1 but contains two discontinuities by period. This type of function is called a cut-off function or a "square wave".

At the time t=1, the exact solution and the numerical solution obtained by the left one-sided, Lax-Friedrichs and Lax-Wendroff schemes are displayed, for the Courant number $\nu = \frac{1}{2}$, on the figures 12, 13 and 14.

N. B. : The exact solution of (7.30) is given by $a(x,\ t) = a_0(x - t)$, at time $t = 1$ that means $a(x,\ 1) = a_0(x - 1)$ and by periodicity the exact solution on the space interval $[0,\ 1]$ is the same that the initial condition.•

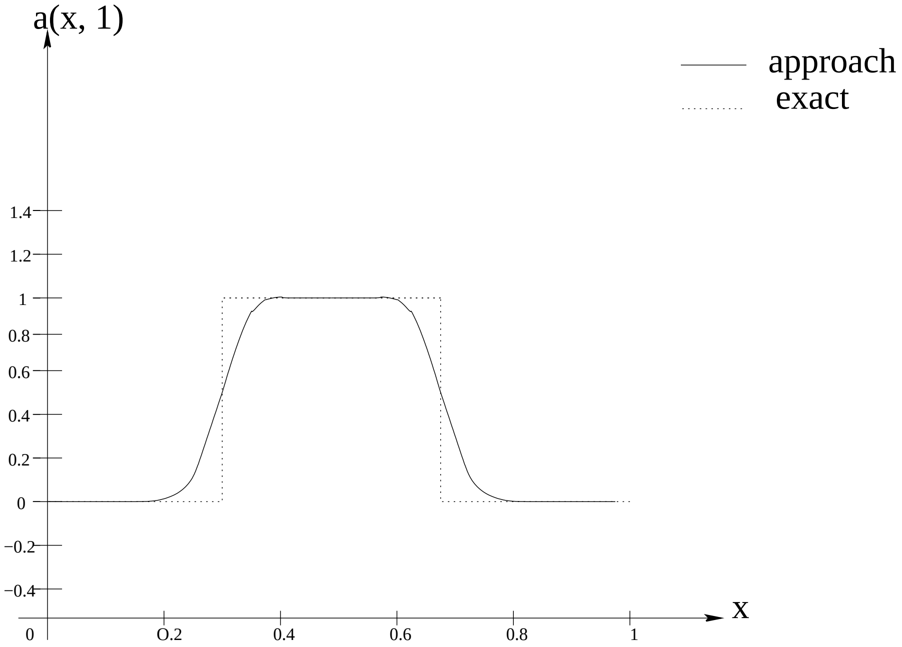


FIG. 12 – Advection of a cut-off function with the left one-sided scheme.

We see that the discontinuities for the solution are highly spreading out by the first order schemes. The second order Lax-Wendroff scheme restore better the shape of the cut-off function but it appears spurious oscillations. The following subsection explains these phenomena.

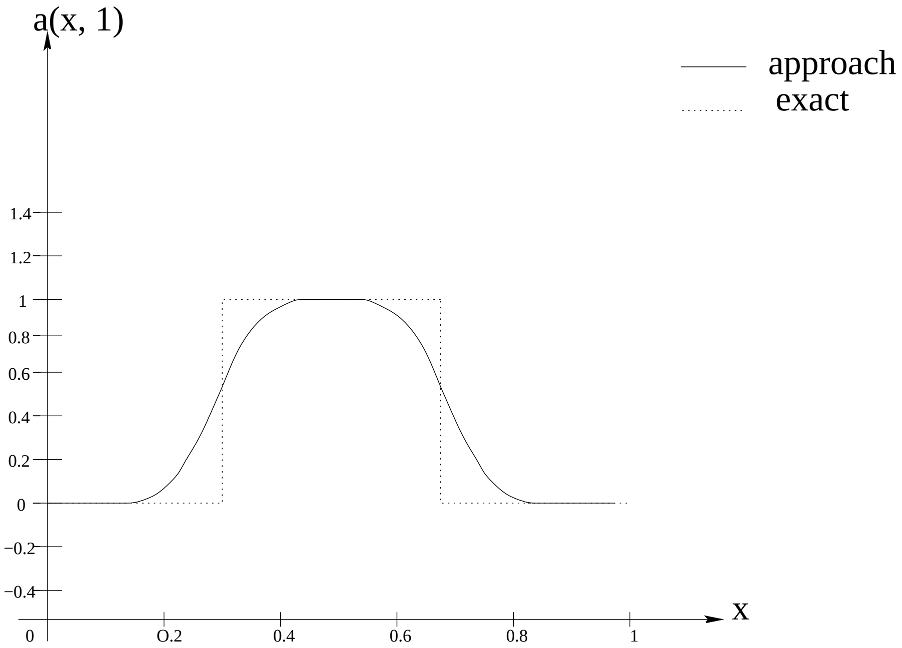


FIG. 13 – Advection of a cut-off function with the Lax-Friedrichs scheme.

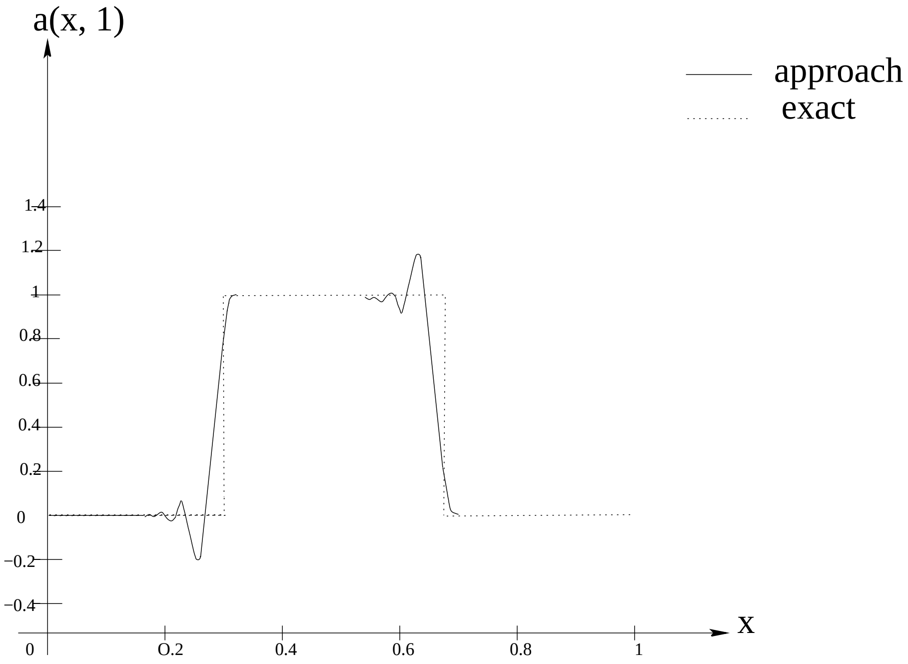


FIG. 14 – Advection of a cut-off function with the Lax-Wendroff scheme.

7.0.4 Equivalent equations. Dissipative and dispersive effects.

Introduce now the *equivalent equation* for a *numerical scheme* which is a powerful tool to analyse the behaviour of a scheme. Let us write the left one-sided scheme (3.10) under the form :

$$\frac{a(x,t+\Delta t)}{\Delta t} + u\frac{a(x-\Delta x,t)}{\Delta x} = 0. \tag{7.31}$$

If a is a regular function satisfying the difference-equation (7.31) for any point $(x,t) \in \mathcal{R} \times \mathcal{R}_+$, then one can write the following Taylor expansions

$$a_t + \frac{\Delta t}{2}a_{tt} + O(\Delta t^2) = -u(a_x - \frac{\Delta x}{2}a_{xx} + O(\Delta x^2)) \tag{7.32}$$

Then we say that the partial derivative equation :

$$a_t + \frac{\Delta t}{2}a_{tt} = -ua_x + u\frac{\Delta x}{2}a_{xx} \tag{7.33}$$

is “approached up the order 2” by the left one-sided scheme.
It is more convenient to rewrite the equation (7.33) with no term a_{tt}. Derivating (7.32) with respect of t, then by respect of x, we obtain :

$$a_{tt} = -ua_{xt} + O(\Delta x + \Delta t), a_{xt} = -ua_{xx} + O(\Delta x + \Delta t), \tag{7.34}$$

hence $a_{tt} = u^2 a_{xx} + O(\Delta x + \Delta t)$. Reporting in (7.33), we have :

$$a_t + \frac{\Delta t}{2}u^2 a_{xx} = -ua_x + u\frac{\Delta x}{2}a_{xx} + O(\Delta x^2 + \Delta t^2) \tag{7.35}$$

By neglecting term in $O(\Delta x^2 + \Delta t^2)$, we obtain now, instead of (7.33), the equation :

$$a_t + ua_x = (u\frac{\Delta x}{2} - u^2\frac{\Delta t}{2})a_{xx} \tag{7.36}$$

which is the *equivalent equation of order two* for the left one-sided scheme; like (7.33), (7.36) is an advection-diffusion equation (the partial derivative a_{xx} is a diffusion term) approached to the order two by the left one-sided scheme.
By Taylor expansions up to the order three, like in the equation,(7.32), and by appropiate eliminations, we can obtain the *equivalent equation of order three* for the left one-sided scheme :

$$a_t + ua_x = (u\frac{\Delta x}{2} - u^2\frac{\Delta t}{2})a_{xx} - (\frac{u\Delta x^2}{6} - \frac{u^2\Delta x\Delta t}{2} + \frac{u^2\Delta t^2}{3})a_{xxx} \tag{7.37}$$

or, taking into account the Courant number ν,

$$a_t + ua_x = \frac{u\Delta x}{2}(1-\nu)a_{xx} - \frac{u\Delta x^2}{6}(2\nu^2 - 3\nu + 1)a_{xxx} \tag{7.38}$$

Remark 17. *: It is evidently possible to do the same work about the Lax-Friedrichs and the Lax-Wendroff schemes, up the order three we have respectively*

$$a_t + ua_x = \frac{\Delta x^2}{2\Delta t}(1-\nu^2)a_{xx} + \frac{u\Delta x^2}{3}(1-\nu^2)a_{xxx} \tag{7.39}$$

$$a_t + ua_x = \frac{u\Delta x^2}{6}(\nu^2 - 1)a_{xxx}.\bullet \tag{7.40}$$

Now it is possible to do, with the equivalent equations, a precise analysis of the numerical errors. Suppose we advect a cut-off function (for the definition see subsubsection 7.0.3. Numerically (see figure 13) the Lax-Friedrichs scheme tends to damp the solution, the discontinuities of the solution are smeared out. Now it is easy to give a first comment. Consider, for instance, the equivalent equations of order two for the left one-sided and Lax-Friedrichs schemes. The modified equation is written :

$$a_t + ua_x = Da_{xx} \tag{7.41}$$

with $D = (u\frac{\Delta x}{2} - u^2\frac{\Delta t}{2})$ for the left one-sided scheme, and $D = \frac{\Delta x^2}{2\Delta t}(1-\nu^2)$ for the Lax-Friedrichs scheme. In a numerical experiment with $u = 1$ and $\frac{\Delta t}{\Delta x} = \frac{1}{2}$, like in subsubsection 7.0.3, we see that $D = \frac{\Delta x}{4}$ for the upwind and $D = \frac{3\Delta x}{4}$ for Lax-Friedrichs. So we expect that upwind is less diffusive that Lax-Friedrichs, as confirmed in the numerical results (see figures 12 and 13).

Furthermore (see figure 14) with the Lax-Wendroff scheme the cut-off function is better represented, but spurious oscillations appears at the discontinuities (one checks also that the discrete maximum principle (see subsection 3.1.1) is not verified by this scheme). To understand this behaviour observe that the second member of (7.38), (7.39) and (7.40) are of the following form $\lambda a_{xx} + \mu a_{xxx}$. So we consider the following problem.

Let be λ, μ, ω,$\varphi \in \mathcal{R}$ with $\lambda > 0$ and the equation :

$$\begin{cases} a_t + ua_x = \lambda a_{xx} + \mu a_{xxx} & \forall (x,t) \in \mathcal{R} \times \mathcal{R}_+ \\ a(x,0) = \sin(\omega x + \varphi) & \forall x \in \mathcal{R} \end{cases} \tag{7.42}$$

the choice of this initial condition is carried out by a Fourier analysis approach of any initial condition. We verify that the solution is given by :

$$a(x,t) = \exp^{-\lambda\omega^2 t} \sin[\omega(x - (u + \mu\omega^2)t) + \varphi] \tag{7.43}$$

This solution is to be compared with the solution of the following problem (i.e. the initial problem with the previous initial condition), which is :

$$\begin{cases} a_t + ua_x = 0 & \forall (x,t) \in \mathcal{R} \times \mathcal{R}_+ \\ a(x,0) = \sin(\omega x + \varphi) & \forall x \in \mathcal{R} \end{cases} \tag{7.44}$$

given immediatly (by the characteristic method) by $u(x,t) = \sin(\omega(x - ut) + \varphi)$. Then we see , comparing the two last exact solutions, that the effect of the term λa_{xx} in the second member of (7.42) bears upon the solution amplitude. If $\lambda > 0$ the solution is damped in such a way that the pulsation ω is raised, it is the *dissipation* or *diffusion* effect. For a difference scheme, **dissipation** yields the attenuation (damping) of the amplitude of the Fourier components of a discrete equation. With an initial data componed of several Fourier modes, that is for the problem

$$\begin{cases} a_t + ua_x = \lambda a_{xx} + \mu a_{xxx} & \forall (x,t) \in \mathcal{R} \times \mathcal{R}_+ \\ a(x,0) = \sum_{k=1}^{K} A_k \sin(\omega_k x + \varphi_k) & \forall x \in \mathcal{R} \end{cases} \tag{7.45}$$

the solution

$$a(x,t) = \sum_{k=1}^{K} A_k \exp^{-\lambda\omega_k^2 t} \sin[\omega_k(x - (u + \mu\omega_k^2)t) + \varphi_k] \tag{7.46}$$

is “smoothed” more and more with the time because high frequence modes are damped more rapidly. We observe this behaviour for the first order schemes (one left-sided and Lax-Friedrichs).

For these schemes the principal term of right members for the equivalent equations (7.38) and (7.39) is well on the form λa_{xx} with $\lambda \geq 1$, this is true if $u \in [0,1]$ for the one left-sided scheme (because $\lambda = \frac{u\Delta x}{2}(1-\nu)$) and $u \in [-1,+1]$ for Lax-Friedrichs (because $\lambda = \frac{\Delta x^2}{2\Delta t}(1-\nu^2)$). This schemes are called principally **dissipative**.

We see also about (7.46) that the term μa_{xxx} modifies the propagation speed u of the solution and that all the modes progress to a speed higher than u for $\mu > 0$ and lower than u if $\mu < 0$; in this two cases the perturbation affects more modes with high frequency than modes with low frequency. That effect which modulates the propagation speed according to the frequency is called *dispersive*. For the first order schemes this dispersive effect is not very perceptible because it is hided by the dissipative effect : in fact observe the second members of (7.39) and (7.40) are of the following form $\lambda a_{xx} + \mu a_{xxx}$ with $\mu = O(\Delta x^2) << \lambda = O(\Delta x)$. On the other hand, this dispersive effect is very perceptible for the Lax-Wendroff second order scheme (with no dissipative effect in $O(\Delta x)$) : this dispersive effect is due to the oscillations observed numerically because it preferentially delayes the high frequency modes (as time evolves these highly components disperse, leading to an oscillatory solution) ; this oscillations are delay on the exact solution because the coefficient μ about the third derivative a_{xxx} is negative in the equivalent equation (7.40). In summary a scheme is *dispersive* if different Fourier components travel at different speeds.

Remark 18. *: In general, the lower-order odd (resp even) term gives the leading dispersive (resp. diffusive) error term.* •

Références

[1] R. Atkinson, D.L. Baulsch, R.A. Cox, R.F. Hampton, J.A. Kerr, M.J. Rossi, J. Troe, 1997. Evaluated kinetics, photochemical and heterogeneous data. *Journal of Physical and Chemical Reference Data*, **26(3)**, pp. 521-1012.

[2] J. S. Chang, R. R. Brost, I. S. A. Isaksen, S. Madronich, P. Middleton, W. R. Stockwell, C. J. Walcek, 1987. A three-dimensional eulerian acid deposition model physical concepts and formulation. *Journal of Geophysical Research*, **92**, pp. 14681-14700.

[3] A. Clappier, 1998 : A correction method for use in multidimensional time-splitting advection algorithms : application to two- and three-dimensional transport. *Monthly Weather Review*, **126**, 232-242.

[4] W. P. Crowley, 1968 : Numerical advection experiments, *Monthly Weather Review*, **96**, 1-11.

[5] P. Collela and P. R. Woodward, 1984 : The piecewise parabolic method (PPM) for gas-dynamical simulations. *Journal of Computational Physics*, **54**, 174-201.

[6] M. Crouzeix et A.L. Mignot : Analyse numérique des équations différentielles. Collection mathématiques appliquées pour la maîtrise, Masson, 1984.

[7] G. Dahlquist. *A special stability problem for linear multistep methods.* BIT, vol. **3**, pp. 27-43, 1963.

[8] K. Dekker and J.G. Verwer : *Stability of Runge-Kutta methods for Stiff nonlinear differential equations.* CWI Monographs, North Holland, 1984.

[9] W.B. De Moore, S.P. Sander, D.M. Golden, R.F. Hampton, M.J. Kurylo, C.J. Howard, A.R. Ravishankara, C.E. Kolb, M.J. Molina, 1994. Chemical kinetics and photochemical data for use in stratospheric modelling. Evaluation 11, JPL Publication 94, 26, JPL, Pasadena, US.

[10] GENEMIS (Generation of European Emission Data for Episodes) project, 1994. EUROTRAC Annual Report 1993, Part 5. EUROTRAC International Scientific Secretariat. Garmisch-Partenkirchen, Germany.

[11] B.B. Hicks, D.D. Baldocchi, T.P. Meyers, R.P. Hosker, D.R. Matt, 1987. A preliminary multiple resistance routine for deriving dry deposition velocities from measured quantities. *Water, Air and Soil Pollution* 36, PP. 311-330.

[12] R. C. Easter : Two modified versions of Bott's positive-definite numerical advection scheme. *Mon. Wea. Rev*, **121**, pp. 297-304.

[13] J.W. Erisman, W. Van Pul, G. Wyers, 1994. Parametrization of surface rsistance for the quantification of atmospheric deposition of acidifying pollutants and ozone. *Atmospheric Environment*, 28, pp. 2595-2607.

[14] R. Eymard, T. Galloüet, and R. Herbin *Finite Volume methods*, Handbook of Numerical Analysis, P. G. Ciarlet and J.-L. Lions (Eds), North Holland, Amsterdam.

[15] C.W. Gear : *Numerical Initial Value Problems in Ordinary Differential Equations.* Prentice Hall, Englewood Cliffs N. J., 1971.

[16] E. Godlewski and P. A. Raviart : *Hyperbolic systems of conservation laws.* Mathematiques & Applications, Ellipses, Paris, 1991

[17] T.E. Graedel and P.J. Crutzen, *Atmosphere, Climate and Change*, Scientific American Library. Freeman and Company, New York-Oxford, 1995.

[18] A. Harten, B. Engquist, S. Osher and S. R. Chakravarthy, 1987 : Uniformly high order accurate essentially non-oscillatory schemes, III. *Journal of Computational Physics*, **71**, 231-303.

[19] A. Harten,1987 : ENO schemes with subcell resolution. *Journal of Computational Physics*, **83**, 357-393.

[20] P. Henrici : *Discrete Variable Methods in Ordinary Differential Equations.* John Wiley and sons, New-York, 1962.

[21] E. Hesstvedt, O. Hov, and I. S. A. Isaksen Quasi-Steady-State Approximations in Air Pollution Modelling : Comparison of Two Numerical Schemes for Oxydant Prediction. *International Journal of Chemical Kinetics*, **Vol. X**, 971-994, 1978.

[22] M. Lattuati, 1997 : Impact des émissions européennes sur le bilan des émissions troposphériques à l'interface de l'Europe et de l'Atlantique nord : apport de la modélisation lagrangienne et des mesures en altitude. Thèse de l'Université, P.& M. Curie, Paris.

[23] R. J. Le Veque, 1990 : *Numerical Methods for Conservation Laws.* Birkäuser Verlag.

[24] J.F. Louis, 1979. A parametric model of vertical eddy fluxes in the atmosphere. *Boundary-Layer Meteorology*, 17, pp. 187-202.

[25] P. Middleton, W.R. Stockwell, W.P.L. Carter, 1990. Aggregation and Analysis of volatile organic compound emissions for regional modeling. *Atmospheric Environment* **24**, pp. 1107-1133.

[26] L. Menut and R. Vautard, M.Beekmann, C. Honoré Sensitivity of photochemical pollution using the adjoint of a simplified chemistry-transport model *Journal of Geophysical Research*, vol. 105, N°D12, pp. 15379-15402, June 27, 2000.

[27] J.J. O'Brien, 1970. A note on the vertical structure of the eddy exchange coefficient in the planetary boundary layer ; *Journal Of Atmospheric Sciences* 27, pp. 1213-1215.

[28] H. Schmidt, C. Derognat, R. Vautard, M. Beekmann A comparison of simulated and observed ozone mixing ratios for the summer of 1998 in Western Europe. *Atmospheric Environment*, **35**,2001, pp ; 6277-6297.

[29] J.H. Seinfeld, S.N. Pandis *Atmospheric Chemistry and Physics*, John Wiley and Sons, Inc., New-York, 1998.

[30] C. Shu and S. Osher (1988) : Efficient Implementation of essentially non-oscillatory schemes. *J. Comput. Phys.*,77, 439-471.

[31] D. Simpson, 1992 : Long period modelling of photochemical oxidants in Europe calculations for July 1985. *Atmospheric Environment* **26**, 1609-1634.

[32] D. Simpson, Y. Andersson-Skold, M.E. Jenkin, 1993 Updating the chemical scheme for the EMEP MSC-W model current status. EMEP MSC-W Note 2/93. The Norwegian Meteorological Institute, Oslo.

[33] A. Staniforth, J. Côté and J. Pudykiewicz, 1987 : Comments on "Smolarkiewikcz's deformational flow." *Monthly Weather Review*, **115**, 894-900.

[34] R. van de Velde, W. Faber, V. Katwijk, H.J. Shoten, T.J.M. Thevessen, M. Verspuy, M. Zevenbergen, 1994. The preparation of an European land-use data base. RIVM Report 712401001, Bilthoven, The Netherlands.

[35] B. Van Leer, 1977 : Towards the ultimate conservative difference scheme. IV. A new approach to numerical convection. *Journal of Computational Physics*, **23**, 276-299.

[36] B. Van Leer, 1979 : Towards the ultimate conservative difference scheme. V. A second-order sequel to Godunov's method. *Journal of Computational Physics*, **32**, 101-136.

[37] Robert Vautard, Matthias Beekmann, Jean Roux, Dominique Gombert. Validation of a hybrid forecasting system for the ozone concentrations over the Paris area. *Atmospheric Environment*, **35**, pp. 2449-2461, 2001.

[38] J. G. Verwer, September 1994. Gauss-Seidel Iteration for Stiff ODES from Chemical Kinetics. *SIAM J. Sci Comput.*, vol. 15, No. 5, pp. 1243-1250.

[39] M.L. Wesley, B.B. Hicks, 1977. Some factors that affect the deposition rates of sulfur dioxide and similar gases on vegetation. *Journal of Air Pollution Controm Association*, 27, PP. 1110-1116.