J. Fluid Mech. (1971), *vol.* 49, *part* 1, *pp.* 75–112
Printed in Great Britain

Numerical simulation of incompressible flows within simple boundaries: accuracy

By STEVEN A. ORSZAG[†]
National Center for Atmospheric Research, Boulder, Colorado

(Received 8 June 1970 and in revised form 25 March 1971)

Galerkin (spectral) methods for numerical simulation of incompressible flows within simple boundaries are shown to possess many advantages over existing finite-difference methods. In this paper, the accuracy of Galerkin approximations obtained from truncated Fourier expansions is explored. Accuracy of simulation is tested empirically using a simple scalar-convection test problem and the Taylor–Green vortex-decay problem. It is demonstrated empirically that the Galerkin (Fourier) equations involving N^p degrees of freedom, where p is the number of space dimensions, give simulations at least as accurate as finite-difference simulations involving $(2N)^p$ degrees of freedom. The theoretical basis for the improved accuracy of the Galerkin (Fourier) method is explained. In particular, the nature of aliasing errors is examined in detail. It is shown that 'aliasing' errors need not be errors at all, but that aliasing should be avoided in flow simulations. An eigenvalue analysis of schemes for simulation of passive scalar convection supplies the mathematical basis for the improved accuracy of the Galerkin (Fourier) method. A comparison is made of the computational efficiency of Galerkin and finite-difference simulations, and a survey is given of those problems where Galerkin methods are likely to be applied most usefully. We conclude that numerical simulation of many of the flows of current interest is done most efficiently and accurately using the spectral methods advocated here.

1. Introduction

This paper compares Galerkin (spectral) methods with finite-difference methods for the numerical simulation of incompressible flows within simple boundaries. An introduction to Galerkin flow approximations and to methods to implement them efficiently is given in Orszag (1971*a*). All the Galerkin approximations used in the present paper are obtained from Fourier expansions. We show that, if a given number of degrees of freedom is used to represent an approximate solution of the Navier–Stokes equations, the Galerkin (Fourier) procedure gives results that are substantially more accurate than are obtained by existing finite-difference methods. Also, in those important special applications where fast transform methods apply (Orszag 1969, 1970, 1971*a*, *b*; Patterson & Orszag 1971), the Galerkin approximations are implementable roughly as efficiently

[†] Permanent address: Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

as finite-difference simulations involving the *same* numbers of degrees of freedom. Hence, to achieve a reasonable standard of accuracy, the Galerkin approximations require considerably less computer memory storage and somewhat less computer time than do finite-difference approximations.

The conclusions reached in the present paper concerning the relative efficiency and accuracy of Galerkin approximations to flows within periodic rectangular boundaries appear to apply with even greater force to Galerkin approximations to flows within slabs, spheres, and cylinders with rigid no-slip or free-slip boundary conditions. Preliminary results obtained using Chebyshev series approximations to flows within no-slip boundaries (Orszag 1971*b*) indicate improvement in accuracy with little cost in efficiency relative to finite-difference methods, in close analogy with the periodic boundary condition results presented here. Furthermore, there may be significant loss of accuracy by the various kinds of special difference approximations used to impose the boundary conditions: the Galerkin (Chebyshev) approximations account for the boundary conditions with infinite-order accuracy. A full account of the comparison between Galerkin (Chebyshev) approximations and finite-difference approximations will be given later.

In §2, we present empirical computational evidence for the accuracy of Galerkin approximations applied to a simple two-dimensional scalar convection problem that has become a rather standard test problem for numerical methods (Crowley 1968; Molenkamp 1968; Burstein & Mirin 1970). In §3, we present further empirical comparisons of accuracy for a three-dimensional vortex-decay problem (Taylor & Green 1937; Goldstein 1940; Orszag 1971*a*). In §4 we explain some theoretical reasons for the relative accuracy of Galerkin approximations over finite-difference approximations. The principal source of inaccurate results with finite-difference schemes is phase error; the Galerkin approximations discussed here have essentially no phase errors. In §5 we give a more satisfactory theoretical treatment of accuracy for simulations of passive scalar convection. In particular, it is explained why the accuracy of Galerkin simulations deteriorates less rapidly with time than the accuracy of finite-difference simulations. Finally, in §6 we compare the computational efficiency of Galerkin and finite-difference simulations. In §6 we also indicate how Galerkin methods may usefully be applied in conjunction with finite-difference methods for accurate simulation of a wide variety of incompressible flows.

Before proceeding, it is appropriate to comment on the sense in which we test accuracy of simulation in this paper. Suppose that an approximate solution is sought to an initial-value problem for the system of m partial differential equations,

$$\partial \mathbf{v}(\mathbf{x},t)/\partial t = \mathbf{F}(\mathbf{v},\mathbf{x},t), \tag{1.1}$$

where $\mathbf{v} = (v_1, v_2, \ldots, v_m)$, $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{F}(\mathbf{v},\mathbf{x},t)$ is a partial differential operator involving derivatives of $\mathbf{v}(\mathbf{x},t)$ in n space dimensions. Here we mean accuracy of simulation to signify accuracy of representation of $\mathbf{F}(\mathbf{v},\mathbf{x},t)$, not accuracy of evaluation of $\partial\mathbf{v}/\partial t$. For all the methods described herein, $\partial\mathbf{v}/\partial t$ in (1.1) is evaluated by finite-difference approximation, and (1.1) is solved as a marching problem in time. In the simulations reported below, the time step is chosen sufficiently small that there is no appreciable error due to time discretization.

The significance of space differencing errors owes to the fact that, while halving a time step requires double the number of computations to evolve (1.1) some fixed finite time interval, halving a space-discretization interval requires at least a factor 2^{n+1} as many computations. (The factor 2^{n+1} is accounted for by the number of points of the space mesh being increased by a factor 2^n, and the time step required for numerical stability of the finite-difference approximation being halved due to the refinement of the mesh.) With a three-dimensional simulation, halving the space-discretization interval increases the required computation time by at least a factor 16, and the computer memory required for a marching calculation by a factor 8, while halving a time step just doubles the computer time. Therefore, it is generally recognized in the literature (cf. e.g. Roberts & Weiss 1966) that accurate space differencing is the primary requisite of accurate simulation of solutions of multi-dimensional partial-differential equations of the form (1.1).

2. Empirical investigation of accuracy: passive scalar convection

Two-dimensional convection of a passive scalar by a uniform rotation velocity is a simple model problem that gives an effective test of the accuracy of numerical simulations (Crowley 1968; Molenkamp 1968; Burstein & Mirin 1970). The scalar field (called a field of 'colour' by Crowley) is denoted by $A(\mathbf{x},t)$ and the convecting velocity is $\mathbf{v}(\mathbf{x},t)$. In a two-dimensional rectangular co-ordinate system, $A(\mathbf{x},t)$ satisfies the equation,

$$\partial A(\mathbf{x},t)/\partial t = -\mathbf{v}(\mathbf{x},t).\nabla A(\mathbf{x},t), \tag{2.1}$$

which is a mathematical statement of the fact that $A(\mathbf{x},t)$ remains constant along particle orbits. The velocity field $\mathbf{v}(\mathbf{x},t)$ is assumed incompressible, so that a stream function $\psi(\mathbf{x},t)$ may be introduced with

$$v_1(\mathbf{x},t) = \partial\psi(\mathbf{x},t)/\partial x_2, \quad v_2(\mathbf{x},t) = -\partial\psi(\mathbf{x},t)/\partial x_1. \tag{2.2}$$

In terms of $\psi(\mathbf{x},t)$, (2.1) becomes

$$\partial A(\mathbf{x},t)/\partial t = J(\psi, A), \tag{2.3}$$

where $J(\cdot,\cdot)$ is the two-dimensional Jacobian. Uniform rotation about the origin with angular velocity Ω (positive for counter-clockwise rotation) corresponds to the stream function,

$$\psi(\mathbf{x},t) = -\tfrac{1}{2}\Omega x^2 \quad (x^2 = x_1^2+x_2^2). \tag{2.4}$$

The initial conditions $A(\mathbf{x},0)$ are chosen as

$$A(\mathbf{x},0) = \left\{\begin{matrix} 1-\bar{x}/r & (\bar{x} \leqslant r) \\ 0 & (\bar{x} > r), \end{matrix}\right\} \tag{2.5}$$

where r is a positive parameter and

$$\bar{x}^2 = (x_1-x_0)^2+x_2^2. \tag{2.6}$$

Here $(x_0, 0)$ is the centre of the conical distribution (2.5). In most of the calculations reported below $x_0 = -\frac{1}{2}$. The initial scalar field (2.5) is plotted in three-dimensional (x_1, x_2, A) perspective in figure 1 (*a*): the initial A field has the shape of an inverted cone of base radius r centred at the point $(-\frac{1}{2}, 0)$. A top view of the same cone is shown in figure 1 (*b*) in which the contours $A = 0{\cdot}2, 0{\cdot}4, 0{\cdot}6, 0{\cdot}8$ are plotted. All the contour plots and perspective drawings included in this paper are unretouched computer plots; the deviations from circular contours in figures 1 (*a*), (*b*) arise from the discreteness of the grid on which the values of A are contoured.

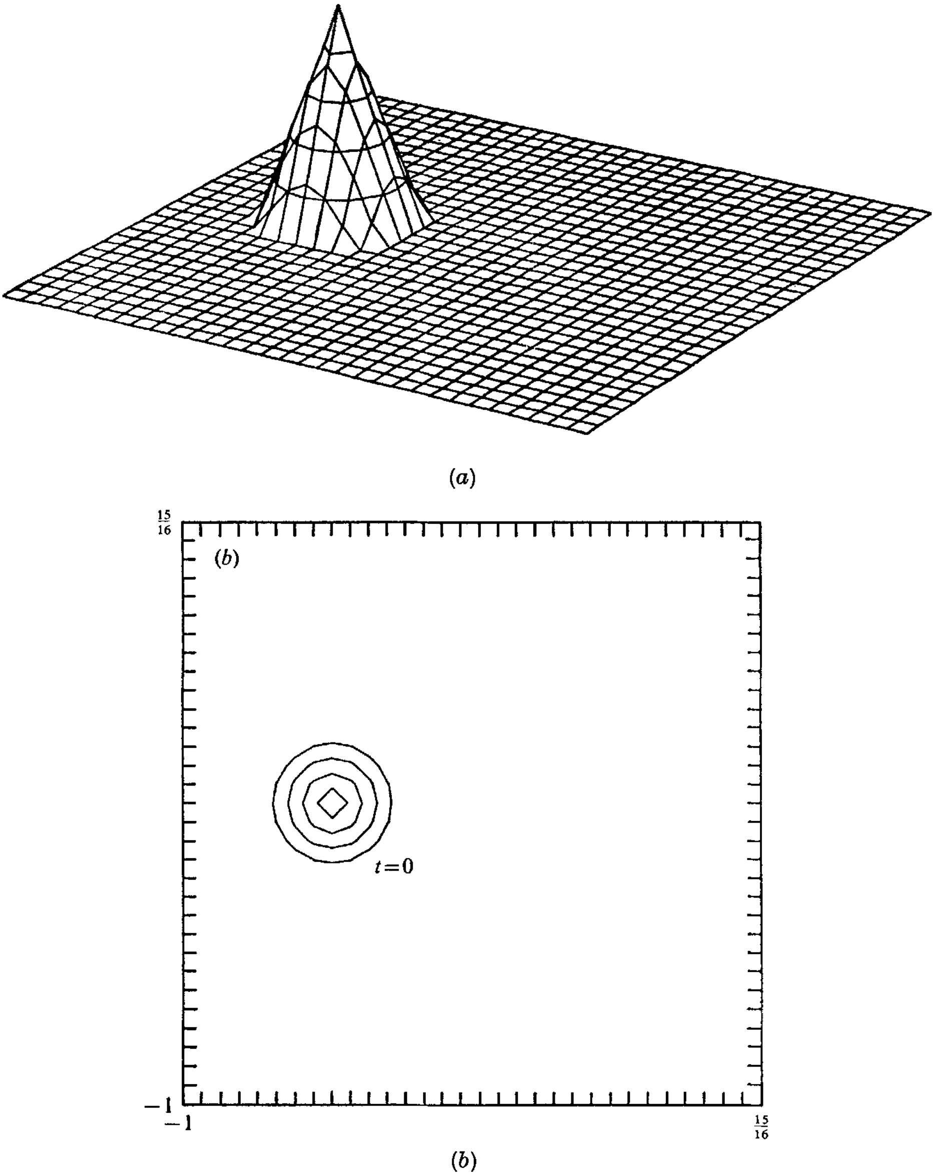


FIGURE 1. (*a*) Three-dimensional (x_1, x_2, A) perspective plot of initial conditions (2.5) with $x_0 = -\frac{1}{2}$, $r = \frac{1}{4}$ on 32 × 32 space grid. (*b*) Top view of cone shown in perspective in (*a*). Contours $A = 0{\cdot}2, 0{\cdot}4, 0{\cdot}6, 0{\cdot}8$ are plotted.

In the exact solution of (2.3) with (2.4), (2.5), the centre of the inverted cone in figure 1 (*a*), (*b*) rotates about the origin with angular velocity Ω, while the cone itself undergoes no change of shape. This solution is used below to test various numerical schemes. First, it is necessary to modify the solution slightly by imposing the periodic boundary conditions,

$$\mathbf{v}(\mathbf{x}+2\mathbf{n},t) = \mathbf{v}(\mathbf{x},t), \quad A(\mathbf{x}+2\mathbf{n},t) = A(\mathbf{x},t), \tag{2.7}$$

where $\mathbf{n}$ has integral components. If $r < \frac{1}{2}$ (with $x_0 = -\frac{1}{2}$), the periodic boundary conditions do not disturb the solution in the sense that, within the square $-1 \leqslant x_\alpha < 1$ ($\alpha = 1, 2$), the conical distribution (2.5) rotates uniformly about $\mathbf{x} = 0$ without change of shape. Applying periodic boundary conditions to $\mathbf{v}(\mathbf{x},t)$ causes no difficulty for an inviscid convecting fluid: the rotation field (2.4) reproduced periodically does not allow fluid to be created or destroyed at the boundaries of the periodicity box (i.e. $\boldsymbol{\nabla}.\mathbf{v} = 0$), although there are infinitely thin shear layers at the box boundaries.

The three numerical schemes that are compared for accuracy here are the second- and fourth-order Arakawa schemes (Arakawa 1966, 1970) and a Galerkin approximation using Fourier expansions. Properties of these schemes are stated in §2(i)–(iii) below.

(i) *Second-order Arakawa scheme*

Arakawa's (1966, 1970) scheme was chosen as a representative and popular second-order difference method. Let $A^n_{jk} = A(x_1,x_2,t)$, $\psi^n_{jk} = \psi(x_1,x_2,t)$ when $x_1 = j\Delta x - 1$, $x_2 = k\Delta x - 1$, $t = n\Delta t$, where Δx, Δt are the space and time differences, respectively. The space grid is termed $N \times N$ if $\Delta x = 2/N$, so that there are N grid points along the x_1 and x_2 axes lying within the periodicity square. The second-order Arakawa space-differencing scheme approximates the Jacobian by

$$\begin{aligned} J^n_{jk} = (1/12\Delta x^2)\,[&\psi^n_{j+1,k}(A^n_{j,k-1} + A^n_{j+1,k-1} - A^n_{j,k+1} - A^n_{j+1,k+1}) \\ &+ \psi^n_{j-1,k}(A^n_{j,k+1} + A^n_{j-1,k+1} - A^n_{j,k-1} - A^n_{j-1,k-1}) \\ &+ \psi^n_{j,k+1}(A^n_{j+1,k} + A^n_{j+1,k+1} - A^n_{j-1,k} - A^n_{j-1,k+1}) \\ &+ \psi^n_{j,k-1}(A^n_{j-1,k} + A^n_{j-1,k-1} - A^n_{j+1,k} - A^n_{j+1\ k-1}) \\ &+ \psi^n_{j+1,k+1}(A^n_{j+1,k} - A^n_{j,k+1}) + \psi^n_{j+1,k-1}(A^n_{j,k-1} - A^n_{j+1,k}) \\ &+ \psi^n_{j-1,k+1}(A^n_{j,k+1} - A^n_{j-1,k}) + \psi^n_{j-1,k-1}(A^n_{j-1,k} - A^n_{j,k-1})]. \end{aligned} \tag{2.8}$$

Equation (2.3) is solved numerically using (2.8) to approximate the Jacobian at grid points and 'leapfrog' (or mid-point rule) time differencing (Richardson 1910)

$$A^{n+1}_{jk} = A^{n-1}_{jk} + 2\Delta t J^n_{jk}, \tag{2.9}$$

to march forward in time. The truncation error of the scheme (2.9) with (2.8) is $O(\Delta t^2) + O(\Delta x^2)$. The scheme is termed second-order, because the error involved in space differencing is $O(\Delta x^2)$. All the results reported in this paper were obtained using second-order time differencing, but, as noted in §1, time steps were always so small that time-differencing errors are negligible.

In some of the calculations reported in §§2, 3, Adams–Bashforth time differencing was used instead of leapfrog. The Adams–Bashforth scheme is

$$A^{n+1}_{jk} = A^{n}_{jk} + \tfrac{3}{2}\Delta t J^{n}_{jk} - \tfrac{1}{2}\Delta t J^{n-1}_{jk}, \tag{2.10}$$

which also gives errors of order $O(\Delta t^2)$. Lilly (1965) showed that, while the leapfrog method is susceptible to 'weak instability' where the solutions at odd and even time steps become uncoupled, the Adams–Bashforth method is not susceptible to this instability. However, we have discovered by numerical experiment that, to achieve given accuracy, the Adams–Bashforth method requires a time step roughly half that of the leapfrog method. The truncation error in $\partial A/\partial t$ for the Adams–Bashforth method is asymptotically $-\frac{5}{12}(\Delta t)^2\, \partial^3 A/\partial t^3$ as $\Delta t \to 0$, while the truncation error of the leapfrog scheme is asymptotically $-\frac{1}{6}(\Delta t)^2\, \partial^3 A/\partial t^3$. It seems that leapfrog differencing is more efficient and the weak instability can be removed by averaging over neighbouring even and odd time steps every 100 time steps or so.

The Arakawa scheme (2.8), (2.9) has the property that, in the limit $\Delta t \to 0$, Δx fixed,

$$\Sigma_{j,k}(A^n_{jk})^2 \tag{2.11}$$

is conserved in time, when the boundary conditions are periodic. The Arakawa scheme also conserves ΣA_{jk}, independently of the size of Δt, Δx. We say that the quadratic quantity (2.11) is 'semi-conserved', because it is conserved in the absence of time-differencing errors. These conservation properties are analogous to the exact conservation of $\int A^2\, dx$, $\int A\, dx$ by (2.3).

Combination of the Arakawa difference approximation (2.8) with the implicit Crank–Nicolson time-differencing scheme (Crank & Nicolson 1947),

$$A^{n+1}_{jk} = A^{n}_{jk} + \Delta t J^{n+\frac{1}{2}}_{jk}, \tag{2.12}$$

where $J^{n+\frac{1}{2}}_{jk}$ is obtained by using $A^{n+\frac{1}{2}}_{jk} = \frac{1}{2}(A^{n+1}_{jk} + A^{n}_{jk})$, $\psi^{n+\frac{1}{2}}_{jk}$ instead of A^n_{jk}, ψ^n_{jk} in (2.8), gives a scheme that conserves (2.11) exactly for any Δt. Conservation of ΣA^2 is important, because it ensures that the numerically determined values of A are bounded. However, the Crank–Nicolson scheme (2.12) is difficult to implement in the present case, while the easily implemented leapfrog scheme only semi-conserves (2.11), so that there is no absolute guarantee that the numerical results are bounded for all time.

In practice, leapfrog time-differencing with sufficiently small Δt (usually $\Delta t < \Delta x/v_{\max}$, where $v_{\max} = \max|\mathbf{v}(\mathbf{x}, t)|$), and the Arakawa approximation (2.8) gives a stable numerical scheme for any fixed finite time interval. The stability of the Arakawa scheme is usually much greater than that of centred-difference approximations with no quadratic semi-conservation properties (cf. Grammeltvedt 1969). Schemes with no quadratic semi-conservation properties may be unstable due to aliasing errors.

Another second-order spatial difference scheme has also been examined, viz. second-order space-differencing of the primitive equation (2.1) on a staggered grid (Lilly 1965). The approximation to $-\mathbf{v}.\nabla A = -\nabla.(\mathbf{v}A)$ at the grid point j, k is

$$\begin{aligned} J^n_{jk} = \frac{1}{2\Delta x}[&u^n_{j-\frac{1}{2},k}(A^n_{j-1,k} + A^n_{jk}) - u^n_{j+\frac{1}{2},k}(A^n_{jk} + A^n_{j+1,k}) \\ &+ v^n_{j,k-\frac{1}{2}}(A^n_{j,k-1} + A^n_{jk}) - v^n_{j,k+\frac{1}{2}}(A^n_{jk} + A^n_{j,k+1})], \end{aligned} \tag{2.13}$$

where the x_1-component of velocity, denoted by u in (2.13), is stored at the grid points $j+\frac{1}{2}$, k and the x_2-component v is stored at $j, k+\frac{1}{2}$. Alternatively, if u, v are stored at the same grid points as A, i.e. j, k, then we may define

$$u_{j+\frac{1}{2},k} = \tfrac{1}{2}(u_{j+1,k}+u_{jk}), \quad v_{j,k+\frac{1}{2}} = \tfrac{1}{2}(v_{j,k+1}+v_{jk})$$

as the values of u, v on cell boundaries. If the velocity field is incompressible in the sense that

$$\frac{1}{\Delta x}[(u_{j+\frac{1}{2},k}-u_{j-\frac{1}{2},k})+(v_{j,k+\frac{1}{2}}-v_{j,k-\frac{1}{2}})] = 0,$$

then (2.13) semi-conserves ΣA^2.

In general, we have found that (2.8) and (2.13) give results that are usually within 5 % of each other, with the staggered-mesh scheme (2.13) usually more accurate. The improved accuracy of the staggered-mesh scheme is explained by the fact that it is not necessary to use finite-difference approximations to derivatives of the stream function in (2.13). However, the differences in accuracy among second-order schemes are much less than the differences between second order, fourth order, and Galerkin schemes reported below.

It is possible to improve the results obtained by second-order schemes by means of Richardson's extrapolation technique (Richardson 1927; Gaunt 1927). Here grid-point solutions to (2.8), (2.9) with grid spacings Δx, Δt and $2\Delta x$, $2\Delta t$, are combined with weight factors $\frac{4}{3}$, $-\frac{1}{3}$, respectively, to give a fourth-order solution. The resulting extrapolation costs little computer time and storage, and usually gives very much improved results. However, the results obtained are still inferior to those obtained by fourth-order methods on the Δx, Δt grid. It should be noted that Richardson extrapolation gives improved values only on points of the $2\Delta x$, $2\Delta t$ grid; values at other points of the Δx, Δt grid must be obtained by interpolation. Also, it should be observed that Richardson extrapolation cannot be applied in conjunction with a sub-grid-scale-eddy viscosity coefficient of the kind suggested by Smagorinsky (1963) (see also Lilly 1967; Deardorff 1971), since the extrapolation treats the eddy viscous term (which contains an explicit factor $(\Delta x)^2$) as an error term and formally cancels it. In the results reported below, Richardson extrapolation was not used. However, it is strongly recommended that any future numerical hydrodynamics calculations employing second-order schemes also employ Richardson extrapolation to improve their accuracy.

(ii) *Fourth-order Arakawa scheme*

Arakawa (1966) also devised a finite-difference approximation to J^n_{jk} with error $O(\Delta x^4)$ that semi-conserves (2.11). This approximation, which will not be written out here,† is combined with leapfrog time-differencing (2.9) to give a fourth-order scheme.

† There is an error in sign in Arakawa (1966, equation (58)). All the signs within round parentheses should be minus signs.

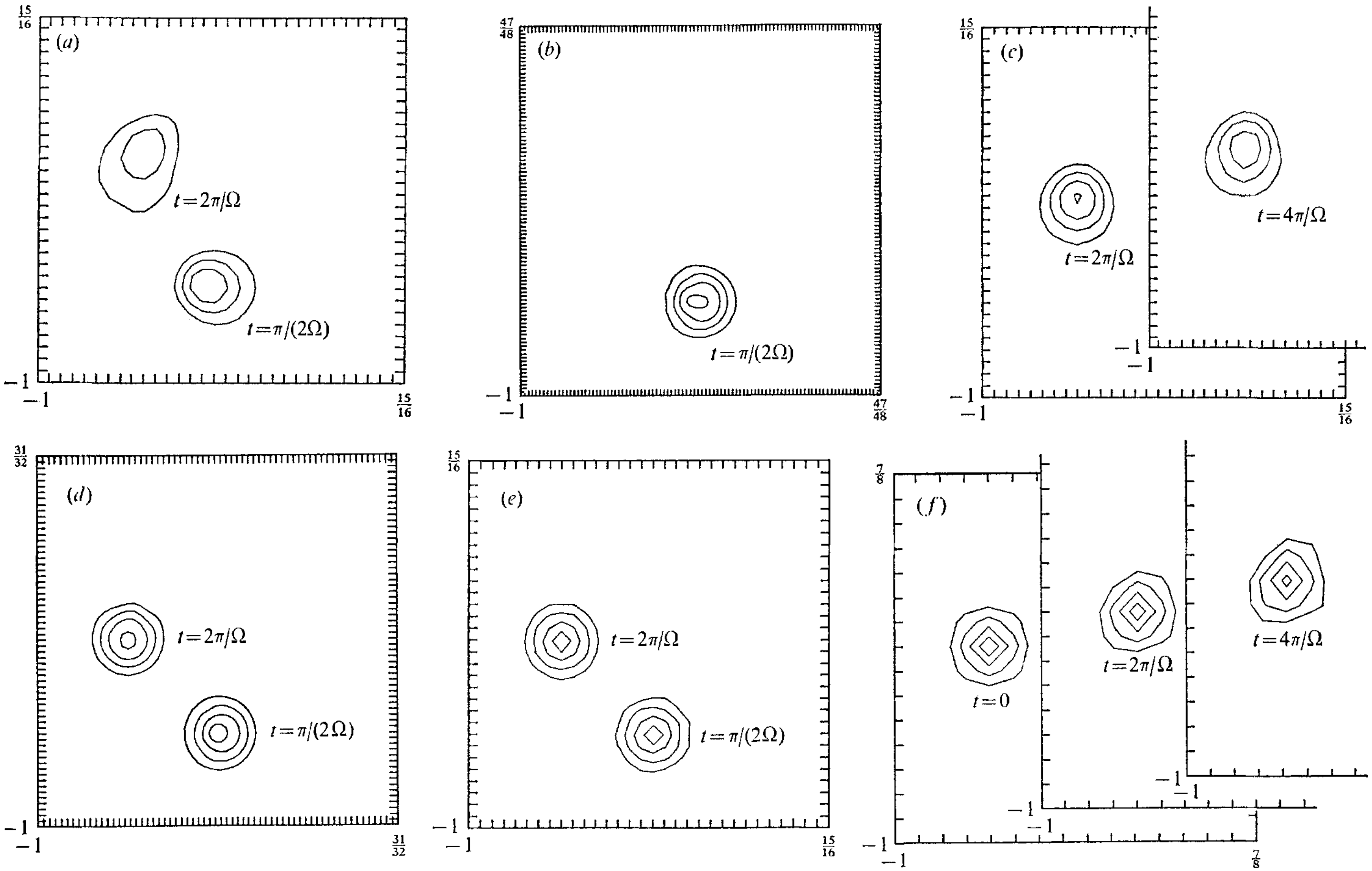


FIGURE 2. Contours of $A(\mathbf{x}, t)$ obtained after (*a*) $\frac{1}{4}$ and 1 revolution, using second-order Arakawa scheme on 32×32 space grid; (*b*) $\frac{1}{4}$ revolution, using second-order Arakawa scheme on 96×96 grid; (*c*) 1 and 2 revolutions using fourth-order Arakawa scheme on 32×32 grid; (*d*) $\frac{1}{4}$ and 1 revolution using fourth-order Arakawa scheme on 64×64 grid; (*e*) $\frac{1}{4}$ and 1 revolution, using cut-off ($K = 16$) Fourier-expansion scheme on 32×32 grid; (*f*) 0, 1 and 2 revolutions, using cut-off ($K = 8$) Fourier-expansion scheme on 16×16 grid. Initial radius of the cone is $r = \frac{1}{4}$.

(iii) *Galerkin* (*Fourier*) *approximation*

Following the technique outlined in Orszag (1971*a*, §2), we seek an approximate solution of (2.3) of the form,

$$A(\mathbf{x},t) = \sum_{\|\mathbf{k}\| \leqslant K} A(\mathbf{k},t)\, e_{\mathbf{k}}(\mathbf{x}), \tag{2.14}$$

where $\|\mathbf{k}\| \leqslant K$ is defined, as in Orszag (1971*a*), to mean $-K \leqslant k_\alpha < K (\alpha = 1, 2)$, $\mathbf{k}$ has integer components, K is an integer cut-off (usually a power of 2), and

$$e_{\mathbf{k}}(\mathbf{x}) = \left\{\begin{array}{lll} \exp(i\pi\mathbf{k}.\mathbf{x}) & \text{if} \quad -K < k_\alpha < K & (\alpha = 1, 2), \\ \cos(\pi K x_1)\exp(i\pi k_2 x_2) & \text{if} \quad k_1 = -K, & -K < k_2 < K, \\ \cos(\pi K x_2)\exp(i\pi k_1 x_1) & \text{if} \quad k_2 = -K, & -K < k_1 < K, \\ \cos(\pi K x_1)\cos(\pi K x_2) & \text{if} \quad k_1 = k_2 = -K. & \end{array}\right\} \tag{2.15}$$

The periodicity interval 2 in both space directions in (2.7) requires that all wave-vectors in the Fourier series representation of $A(\mathbf{x},t)$ have components that are integral multiples of π. Reality of the physical-space $A(\mathbf{x})$ field requires that

$$[A(\mathbf{k},t)]^* = A(\mathbf{k}^*,t), \tag{2.16}$$

where $k_\alpha^* = -k_\alpha$ if $-K < k_\alpha < K$, $k_\alpha^* = -K$ if $k_\alpha = -K$ $(\alpha = 1, 2)$. Notice that $e_{\mathbf{k}}(\mathbf{x})$ satisfies (2.16) as a function of $\mathbf{k}$.

It is necessary to explain the reason for expanding $A(\mathbf{x},t)$ in terms of the functions $e_{\mathbf{k}}(\mathbf{x})$, rather than in terms of pure complex exponential functions, as in Orszag (1971*a*). With the choice of $\{e_{\mathbf{k}}(\mathbf{x})\}$ as expansion functions, the description of the A field in terms of the Fourier coefficients $A(\mathbf{k},t)$ for $\|\mathbf{k}\| \leqslant K$ contains exactly as much information as the values of $A(\mathbf{x},t)$ on the $2K \times 2K$ space grid $\mathbf{x}_{mn} = ([m-K]/K, [n-K]/K)$, $m, n = 0, 1, \ldots, 2K-1$. In fact, if

$$A(\mathbf{x}_{mn}) = \sum_{\|\mathbf{k}\| \leqslant K} A(\mathbf{k})\, e_{\mathbf{k}}(\mathbf{x}_{mn}), \tag{2.17}$$

then

$$A(\mathbf{k}) = \frac{1}{4K^2} \sum_{m=0} \sum_{n=0} A(\mathbf{x}_{mn}) \exp(-i\pi\mathbf{k}.\mathbf{x}_{mn}) \quad (\|\mathbf{k}\| \leqslant K), \tag{2.18}$$

and vice versa, as follows from (14)–(16) of Orszag (1971*a*) using the fact that

$$e_{\mathbf{k}}(\mathbf{x}_{mn}) = \exp(i\pi\mathbf{k}.\mathbf{x}_{mn}) \quad (\|\mathbf{k}\| \leqslant K, m, n = 0, \ldots, 2K-1). \tag{2.19}$$

Equation (2.19) states that wave-vector components equal to $+K$ and $-K$ are indistiguishable on the discrete grid $\mathbf{x}_{mn}$ (a fact related to aliasing on the grid $\mathbf{x}_{mn}$ (cf. §4)). The basis functions (2.15) allow the derivation of a consistent set of Galerkin equations for $A(\mathbf{k},t)$ $(\|\mathbf{k}\| \leqslant K)$ that maintains the reality condition (2.16) for all t; also, the $4K^2$ independent (real or imaginary) parts of $A(\mathbf{k},t)$ satisfying (2.16) are equivalent to $4K^2$ independent real values of $A(\mathbf{x},t)$ on the grid $\mathbf{x}_{mn}$. On the other hand, the Orszag (1971*a*) expansion (8) contains (in two space-dimensions for a single scalar quantity such as A) only $(2K-1)^2$ independent pieces of real information, not enough for inversion of $A(\mathbf{x})$ on the full grid $\mathbf{x}_{mn}$, while directly extending the sum in Orszag (1971*a*, (8)) to $\|\mathbf{k}\| \leqslant K$ would give a set of Galerkin equations that would not maintain the reality conditions.

Using an expansion similar to (2.14), it is possible to approximate $\psi(\mathbf{x},t)$ as a truncated Fourier series with coefficients $\psi(\mathbf{k},t)$ ($\|\mathbf{k}\| \leqslant K$) satisfying (2.16). Also, the functions $e_{\mathbf{k}}(\mathbf{x})$ have the orthogonality property that

$$\frac{1}{4}\int_{-1}^{1} dx_1 \int_{-1}^{1} dx_2 [e_{\mathbf{k}}(\mathbf{x})]^* \, e_{\mathbf{p}}(\mathbf{x}) = \begin{cases} 0 & \text{if} \quad \mathbf{k} \neq \mathbf{p}, \\ n(k_1)\, n(k_2) & \text{if} \quad \mathbf{k} = \mathbf{p}, \end{cases} \tag{2.20}$$

where $n(k) = 1$ if $-K < k < K$, $n(-K) = \frac{1}{2}$. The Galerkin procedure of Orszag (1971*a*, §2) applied to (2.3) using the expansion (2.14) and the orthogonality property (2.20) gives the equations,

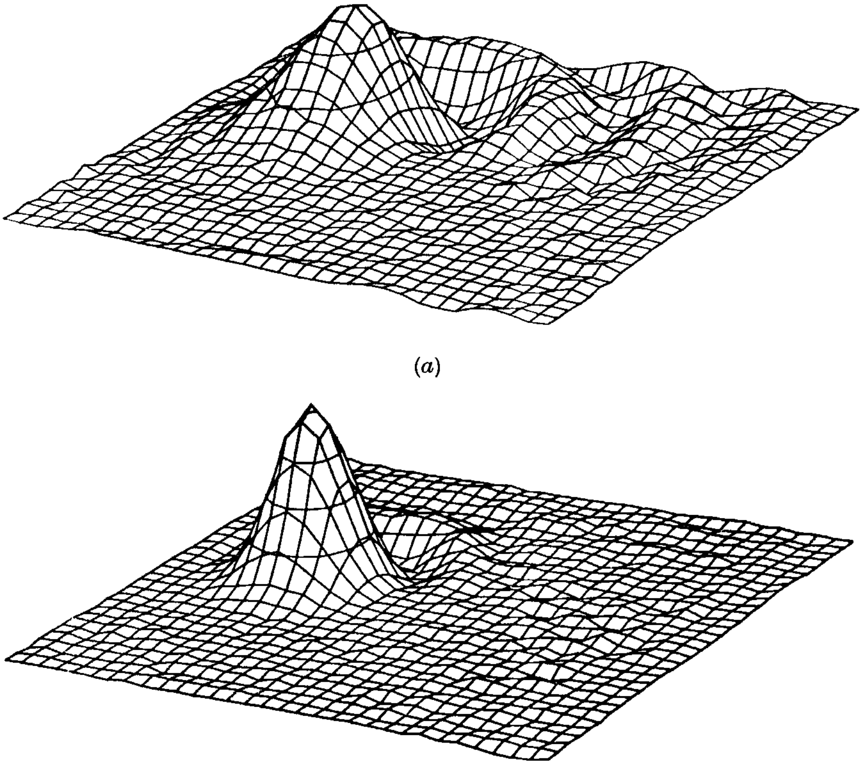
$$\frac{\partial A(\mathbf{k},t)}{\partial t} = \sum_{\|\mathbf{p}\| \leqslant K} \sum_{\|\mathbf{q}\| \leqslant K} I(\mathbf{k}|\mathbf{p},\mathbf{q})\, \psi(\mathbf{p},t)\, A(\mathbf{q},t) \quad (\|\mathbf{k}\| \leqslant K) \tag{2.21}$$

where $$I(\mathbf{k}|\mathbf{p},\mathbf{q}) = [4n(k_1)\, n(k_2)]^{-1} \int_{-1}^{1} dx_1 \int_{-1}^{1} dx_2 [e_{\mathbf{k}}(\mathbf{x})]^* \, J(e_{\mathbf{p}}(\mathbf{x}), e_{\mathbf{q}}(\mathbf{x})). \tag{2.22}$$

Property (2.16) is preserved in time by (2.21), since $[I(\mathbf{k}|\mathbf{p},\mathbf{q})]^* = I(\mathbf{k}^*|\mathbf{p}^*,\mathbf{q}^*)$.

It may easily be verified that

$$I(\mathbf{k}|\mathbf{p},\mathbf{q}) = \begin{cases} 0 & \text{if} \quad \bar{\mathbf{k}} \neq \bar{\mathbf{p}} + \bar{\mathbf{q}}, \\ \pi^2 n(p_1)\, n(p_2)\, n(q_1)\, n(q_2)\, [\bar{q}_1 \bar{p}_2 - \bar{q}_2 \bar{p}_1] & \text{if} \quad \bar{\mathbf{k}} = \bar{\mathbf{p}} + \bar{\mathbf{q}}, \end{cases} \tag{2.23}$$



(*a*)

(*b*)

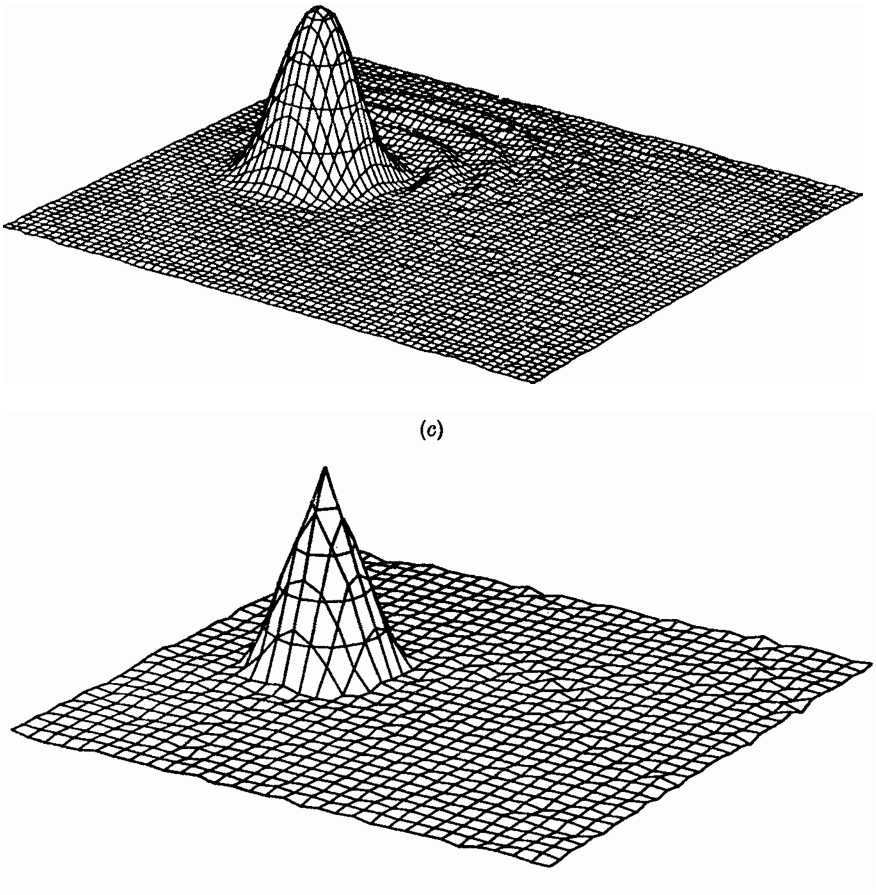
FIGURE 3 (*a*)–(*b*). For legend see facing page.

where barred wave vectors are obtained from their unbarred counterparts by interpreting a component k_α equal to $-K$ as $\bar{k}_\alpha$ equal to either $+K$ or $-K$. For example, if $\mathbf{p} = (-K, -K), k_1 = K+q_1, k_2 = K+q_2$, and $-K < k_1, k_2, q_1, q_2 < K$, then $I = -\frac{1}{4}\pi^2 K(q_1+q_2)$; if $\mathbf{k} = (-K, k_2)$, $\mathbf{p} = (p_1, -K)$, $K = p_1+q_1, k_2 = -K+q_2$, and $-K < k_2, p_1, q_1, q_2 < K$, then $I = -\frac{1}{2}\pi^2(Kq_1+p_1q_2)$.

Aside from interactions with modes having at least one component $-K$, the right-hand side of (2.21) equals

$$\pi^2 \sum_{\substack{\mathbf{p}+\mathbf{q}=\mathbf{k}\\ \|\mathbf{p}\|,\,\|\mathbf{q}\|<K}} (q_1 p_2 - q_2 p_1)\,\psi(\mathbf{p},t)\,A(\mathbf{q},t), \tag{2.24}$$

where $\|\mathbf{k}\| < K$ means $-K < k_\alpha < K$ $(\alpha = 1, 2)$. Therefore, aside from the slightly more complicated interactions with wave-vectors having a component $-K$, the right-hand side of (2.21) involves two convolution sums, *viz.* the



(*c*)

(*d*)

FIGURE 3. Three-dimensional (x_1, x_2, A) perspective plot of the $A(\mathbf{x}, t)$ field obtained after 1 revolution using (*a*) second-order Arakawa scheme on 32 × 32 space grid (see figure 2(*a*) for top view); (*b*) fourth-order Arakawa scheme on 32 × 32 space grid (see 2(*c*)); (*c*) fourth-order Arakawa scheme on 64 × 64 space grid (see 2(*d*)); (*d*) cut-off Fourier-expansion scheme on 32 × 32 space grid (see 2(*e*)).

convolution of $ip_2\psi(\mathbf{p})$ with $iq_1 A(\mathbf{q})$ and $ip_1\psi(\mathbf{p})$ with $iq_2 A(\mathbf{q})$. These convolution sums are most efficiently evaluated by the transform methods of Orszag (1971*a*). In fact, the right-hand side of (2.21) for $\|\mathbf{k}\| \leqslant K$ can be evaluated by the transform method with essentially no additional work required to include exactly the interactions with wave vectors having components equal to $-K$. This latter result is explained in the appendix. The principal result of the appendix is that the right-hand side of (2.21) can be evaluated in 27 real (or half-complex) discrete Fourier transforms on $K \times K$ points, if the Fourier transforms of $ip_1\psi(\mathbf{p})$, $ip_2\psi(\mathbf{p})$ are computed and stored before the start of the calculation of $A(\mathbf{k}, t)$.

The numerical scheme for solution of (2.21) is completed by using leapfrog time differencing (2.9) to approximate $\partial A/\partial t$. The initial conditions $A(\mathbf{k}, 0)$ are determined from (2.5) by (2.18), while the physical-space $A(\mathbf{x}, t)$ field is reconstructed on the $2K \times 2K$ space grid $\mathbf{x}_{mn}$ from (2.17) using the solution to (2.21).

It can easily be shown that (2.21) with (2.22) semi-conserves

$$A(0,0,t), \quad \sum_{\|\mathbf{k}\| \leqslant K} n(k_1)\, n(k_2)\, A(\mathbf{k}, t)\, A(\mathbf{k}^*, t), \tag{2.25}$$

which are the Fourier space analogs of $\int A\, d\mathbf{x}$, $\int A^2 d\mathbf{x}$, respectively. The quadratic semi-conservation property (2.25) assists the stability of numerical solution of (2.21), so that, if Δt is small enough, there is no numerical instability of (2.21), using leapfrog time differencing for any fixed finite time of evolution.

If the Galerkin procedure is applied to (2.1) directly using cut-off Fourier expansions of $\mathbf{v}(\mathbf{x}, t)$, the results are improved slightly over (2.21). However, to facilitate direct comparison with the Arakawa schemes, these Galerkin equations (which may be implemented as efficiently as (2.21)) are used only in § 5.

Another modified Galerkin approximation gives greatly improved results. If the problem (2.3) with (2.4) is reformulated in polar co-ordinates (r, θ) with the origin as pole, then

$$\partial A(r, \theta, t)/\partial t = -\Omega\, \partial A(r, \theta, t)/\partial \theta. \tag{2.26}$$

If the Galerkin procedure is applied using an expansion of $A(r, \theta, t)$ in the functions $\exp(in\theta)$, i.e. a complex Fourier series in θ, then the only error in simulation is time differencing. However, to use results obtained in this way for comparison with results obtained by finite-difference methods is very unfair. The functions $\exp(in\theta)$ are obviously well suited to the solution of (2.3) with the *special* stream function (2.4): the functions $\exp(in\theta)$ are eigenfunctions of $\Omega\, \partial A/\partial \theta$. Equation (2.26) is not used in the sequel. On the other hand, the expansion (2.14) is not exact, the functions $e_{\mathbf{k}}(\mathbf{x})$ bear no special relation to the eigenfunctions of (2.3) with (2.4), and it is not trivially evident that (2.21) gives a good approximation to $A(\mathbf{x}, t)$. The comparison of the results given by (2.21) with results given by finite-difference approximations is a fair test of the accuracy of the methods.

(iv) *Results*

Contour plots of $A(\mathbf{x}, t)$ using the schemes described in § 2 (i)–(iii) above with $r = \frac{1}{4}$ in (2.5) are shown in figures 2(*a*)–(*f*), while the results of calculations with $r = \frac{1}{16}$ are contoured in figures 4(*a*), (*b*). The results for fixed cone radius r

obtained using different numerical schemes and grids are suitable for critical comparison of the schemes and grids.

The contours of the cone with $r = \frac{1}{4}$ convected according to the second-order Arakawa scheme (2.8) on 32×32 space grid are shown in figure 2(*a*) at times of $\frac{1}{4}$ and 1 revolution of the convecting velocity (2.4), i.e. at $t = \pi/(2\Omega)$ and $t = 2\pi/\Omega$.

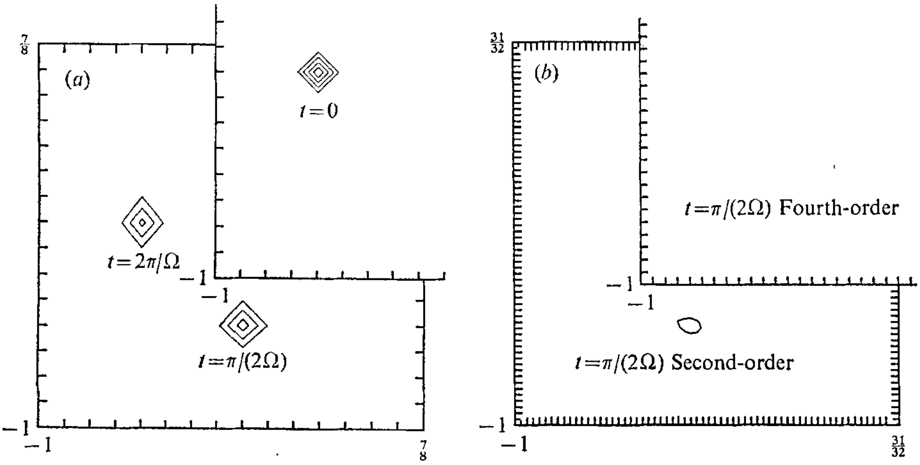


FIGURE 4. Contours of $A(\mathrm{x}, t)$ obtained after (*a*) 0, $\frac{1}{4}$, and 1 revolution, using cut-off ($K = 8$) Fourier-expansion scheme on 16×16 grid; (*b*) $\frac{1}{4}$ revolution, using second-order Arakawa scheme on 64×64 grid and fourth-order Arakawa scheme on 32×32 grid. Initially, $r = \frac{1}{16}$.

In this figure and succeeding figures, missing contours are those for the highest A values, e.g. $A = 0{\cdot}8$ and $A = 0{\cdot}6$ after 1 revolution in figure 2(*a*). The contours obtained after $\frac{1}{4}$ revolution using the second-order Arakawa scheme (with $r = \frac{1}{4}$) on a 96×96 space grid are shown in figure 2(*b*). The results presented graphically in figures 2(*a*), (*b*), together with the quantitative comparisons provided in table 1, show that refinement of the mesh by a factor 3 in each space direction (9 times as many grid points) gives surprisingly little improvement in the results, especially in the maximum of $A(\mathrm{x})$ which should always be 1.

Results obtained using the fourth-order Arakawa scheme on a 32×32 space grid are plotted in figure 2(*c*) at times of 1 and 2 revolutions. The contours obtained using a 64×64 grid are shown in figure 2(*d*) at times of $\frac{1}{4}$ and 1 revolution.

Comparison of the contours in figures 2(*a*), (*b*) with those in figures 2(*c*), (*d*) shows the important improvement in the quality of the results obtained using fourth-order schemes rather than second-order schemes. In fact, comparisons closely analogous to those just made have led many numerical analysts (e.g Roberts & Weiss 1966; Crowley 1968; Molenkamp 1968; Fromm 1969; Burstein & Mirin 1970; Price & Varga 1970) to suggest the abandonment of second-order schemes in favour of those of fourth-order for the numerical simulation of solutions of multi-dimensional partial-differential equations. (Presumably, Richardson extrapolation of second-order solutions is more efficient for one-dimensional problems.) Our results support the use of fourth-order methods. However, an important qualification will be made at the end of §4.

The results for $A(\mathbf{x}, t)$ obtained using the Galerkin (cut-off Fourier-expansion) approximation (2.21), with (2.17) and (2.18) to translate the results onto a discrete space-grid, are contoured in figures 2(*e*), (*f*). With cut-off $K = 16$, the Fourier-space results translate into values on a 32×32 physical-space grid. Contour plots of $A(\mathbf{x}, t)$ determined by the Galerkin method with cut-off $K = 16$ are shown in figure 2(*e*) at times of $\frac{1}{4}$ and 1 revolution. With cut-off $K = 8$, contouring is done on a 16×16 space grid so that even the contours of A at $t = 0$ are not closely circular. In figure 2(*f*), we plot the contours of A when $K = 8$ at times of 0, 1, and 2 revolutions of the convecting velocity field.

Numerical method	Order of scheme	Mesh	Number of quarter-revolutions	Maximum at a grid point	Minimum	Lag of maximum (in radians)
Arakawa	Second	96×96	1	0·85	−0·04	0·08
Arakawa	Second	32×32	1	0·79	−0·13	0·19
Arakawa	Second	32×32	4	0·51	−0·23	0·44
Arakawa	Fourth	64×64	1	0·92	−0·03	0·03
Arakawa	Fourth	64×64	4	0·84	−0·04	0
Arakawa	Fourth	32×32	1	0·89	−0·05	0
Arakawa	Fourth	32×32	4	0·83	−0·10	0·08
Arakawa	Fourth	32×32	8	0·74	−0·15	0·08
Fourier	Infinite	32×32 ($K = 16$)	1	0·98	−0·02	0
Fourier	Infinite	32×32 ($K = 16$)	4	0·98	−0·02	0
Fourier	Infinite	16×16 ($K = 8$)	1	0·97	−0·03	0
Fourier	Infinite	16×16 ($K = 8$)	4	0·96	−0·04	0
Fourier	Infinite	16×16 ($K = 8$)	8	0·89	−0·07	0
Fourier	Infinite	16×16 ($K = 8$)	12	0·79	−0·11	0

TABLE 1. Accuracy of numerical simulation of scalar convection

Three-dimensional perspective plots of $(x_1, x_2, A(\mathbf{x}, t))$ are shown in figures 3(*a*)–(*d*). These figures are drawn after 1 revolution from the results obtained by the second-order Arakawa scheme on a 32×32 grid (figure 3(*a*)), fourth-order Arakawa scheme on a 32×32 grid (figure 3(*b*)) and on a 64×64 grid (figure 3(*c*)), and the cut-off Fourier-expansion scheme with cut-off $K = 16$ (figure 3(*d*)). It is strikingly evident from figures 2(*a*)–3(*d*) that results obtained using the cut-off Fourier-expansion scheme on a 32×32 space grid (cut-off $K = 16$) are at least as good as those obtained using the fourth-order scheme on a 64×64 grid, and significantly better than those obtained by the fourth-order scheme on a 32×32 grid and the second-order scheme on 32×32 and 96×96 grids. This appraisal of the relative accuracy of the various schemes is supported by the quantitative results listed in table 1. If the simulation were exact, the maximum at a grid point after any integral number of quarter-revolutions would be 1·0, the minimum would be 0 (since $A(\mathbf{x}, t) = 0$ outside the cone shown in figure 1), and the lag of the maximum would be 0. Here the lag of the maximum is the angular lag (in radians) of the (interpolated) maximum of $A(\mathbf{x}, t)$ from its exact position. In all the calculations reported in table 1, radial displacements of the maxima are small.

The 'wakes of bad numbers' that lag the counterclockwise rotation of the cone in figures 3 (*a*)–(*c*) should be susceptible to theoretical explanation. Such an explanation is likely related to the fact that finite-difference schemes (i) and (ii) have mostly lagging phase errors (cf. § 4), so that some of the information needed to reconstruct the cones at later times lags the correct position of the cones.

When $r = \frac{1}{16}$, the initial $A(\mathbf{x}, 0)$ field on a 16×16 space grid is zero everywhere except at the grid point located at $(-\frac{1}{2}, 0)$, where A is initially 1. Such a point-excitation solution to the linear equation (2.3) is a Green's function (fundamental solution) that may be used as a building block out of which the general solution to (2.3) may be constructed. Numerical methods are not very well adapted to direct calculation of Green's functions because of the large gradients involved. The fidelity of a numerical simulation of (2.3) with (2.4), (2.5) deteriorates with decreasing r. Contours of the results of numerical calculations for $r = \frac{1}{16}$ using the three numerical schemes outlined above are shown in figures 4 (*a*), (*b*). The contours of $A(\mathbf{x}, t)$ after 0, $\frac{1}{4}$, 1 revolution obtained using the cut-off Fourier-expansion scheme on a 16×16 grid (cut-off $K = 8$) are shown in figure 4 (*a*); the contours at $t = 0$ are included to illustrate how the contour plotter treats this singular case. The contours obtained after $\frac{1}{4}$ revolution by the second-order scheme on a 64×64 grid and the fourth-order scheme on a 32×32 grid are shown in figure 4 (*b*). After $\frac{1}{4}$ revolution, the maximum of $A(\mathbf{x})$ determined by the second-order scheme is 0·27, the fourth-order scheme 0·21, and the Fourier-expansion scheme 0·73. After 1 revolution, the second- and fourth-order schemes give maxima 0·12 and 0·16, respectively, so that no contours remain to be plotted. With the cut-off Fourier-expansion scheme, the maximum of $A(\mathbf{x})$ is 0·67 after 1 revolution.

It seems a safe conclusion from the numerical results reported above that the cut-off Fourier-expansion scheme offers significant improvement in accuracy in comparison with finite-difference schemes. For the scalar convection problem studied here, it seems that to achieve a reasonable standard of accuracy second-order schemes require at least twice as many grid intervals in *each* space direction as fourth-order schemes, which themselves require at least twice as many intervals in each direction as the cut-off Fourier-expansion scheme.

Comparisons of accuracy between spectral and finite-difference methods made previously (cf. e.g. Ellsaesser 1966) were, in the author's opinion, inconclusive. These early studies concerned principally energy conservation properties of surface harmonic representations of flows on the surface of a sphere and comparisons with real meteorological data under uncontrolled computational conditions; these studies did not lead to estimates of the discrete-grid and spectral resolutions necessary to achieve reasonable standards of accuracy.

Perhaps the most remarkable feature of the spectral method that is brought out by the cone problem studied here is the result that the cone is better localized in space using the cut-off Fourier-expansion scheme than by finite-difference schemes formulated directly in physical space. Graphical evidence of this property is provided by figures 3 (*a*)–(*d*), while the grid point minima listed in table 1 provide quantitative evidence. It is remarkable that the cone is better localized by expanding the initial grid-point excitation in a discrete Fourier series, evolving the Fourier coefficients to a later time, and then reconstituting the grid-point field out of the evolved Fourier coefficients, than by direct evolution of the grid-point field using a finite-difference method. In other words, physical-space intermittent flow features (local structures with large excitations) seem to be more easily followed in terms of their Fourier coefficients than directly in physical space! Some reasons for this attractive, but perhaps unexpected, feature of the cut-off Fourier-expansion method are given in §4. Essentially, the point is that Fourier expansion allows much better interpolation between grid points than is given by the local grid-point values alone.

The property, that the Fourier coefficients (collective co-ordinates measuring the excitation in a certain scale over the whole flow) provide good descriptions of local flow structure, is subject to further test by modifying the convecting stream function (2.4) to

$$\psi(\mathbf{x},t) = \begin{Bmatrix} -\tfrac{1}{2}\Omega x^2 & (x \leqslant a), \\ -\tfrac{1}{2}\Omega a^2 & (x > a), \end{Bmatrix} \tag{2.27}$$

for $\mathbf{x}$ within the periodicity square $-1 \leqslant x_\alpha < 1$ $(\alpha = 1, 2)$ and maintaining (2.7) for other $\mathbf{x}$. The velocity field determined by (2.27) is such that, within each periodicity square, there is a core of radius a of uniform rotation with angular velocity Ω, while there is no motion outside the core. If $r + a \leqslant x_0$, then the cone of base radius r centred initially at $(x_0, 0)$ is not disturbed by the velocity field given by (2.27); the cone remains stationary. The results obtained by the cut-off Fourier-expansion scheme on a 32×32 grid (cut-off $K = 16$) using (2.27) with $x_0 = -\frac{5}{8}$, $r = \frac{1}{4}$, $a = \frac{3}{8}$ are contoured in figure 5. The inner rotating core of radius a just touches the outside lip of the cone, so that, if the simulation were exact, the cone would remain centred at $(-\frac{5}{8}, 0)$. The contours in figure 5 are plotted at $t = \pi/\Omega$, when the inner core has rotated through 180°. It is apparent that the cut-off Fourier-expansion scheme successfully keeps local physical-space structures localized during time evolution, even in the presence of the sharp shear layer provided by (2.27).

In summary, the results obtained in §2 suggest that: (*a*) on a given grid, the cut-off Fourier-expansion scheme gives results significantly more accurate than those of finite-difference methods; and (*b*) the cut-off Fourier expansion method accurately describes local flow structures. These comparisons seem to hold true rather generally for Galerkin approximations of infinite-order accuracy (i.e. for schemes in which the truncation error decreases faster than algebraically with the number of degrees of freedom except for time-differencing errors). In §6 we compare the computational efficiency of the various numerical schemes discussed above.

3. Empirical investigation of accuracy: Taylor–Green vortex

The Taylor–Green vortex-decay problem (Taylor & Green 1937; Goldstein 1940; Orszag 1971*a*) involves solution of the Navier–Stokes equations,

$$\partial \mathbf{v}(\mathbf{x},t)/\partial t = -\mathbf{v}(\mathbf{x},t) . \nabla \mathbf{v}(\mathbf{x},t) - \nabla p(\mathbf{x},t) + \nu \nabla^2 \mathbf{v}(\mathbf{x},t), \tag{3.1}$$

$$\nabla . \mathbf{v}(\mathbf{x},t) = 0, \tag{3.2}$$

with initial conditions

$$\mathbf{v}(\mathbf{x}, 0) = (\cos(x_1)\sin(x_2)\cos(x_3),\ -\sin(x_1)\cos(x_2)\cos(x_3),\ 0), \tag{3.3}$$

where $\mathbf{v} = (v_1, v_2, v_3)$ is the velocity field, p is the pressure (normalized by the density), and ν is kinematic viscosity. The origin of x_3 in the initial conditions (3.3) is shifted by $\frac{1}{2}\pi$ from the co-ordinates used by Taylor & Green, for reasons explained in Orszag (1971*a*).

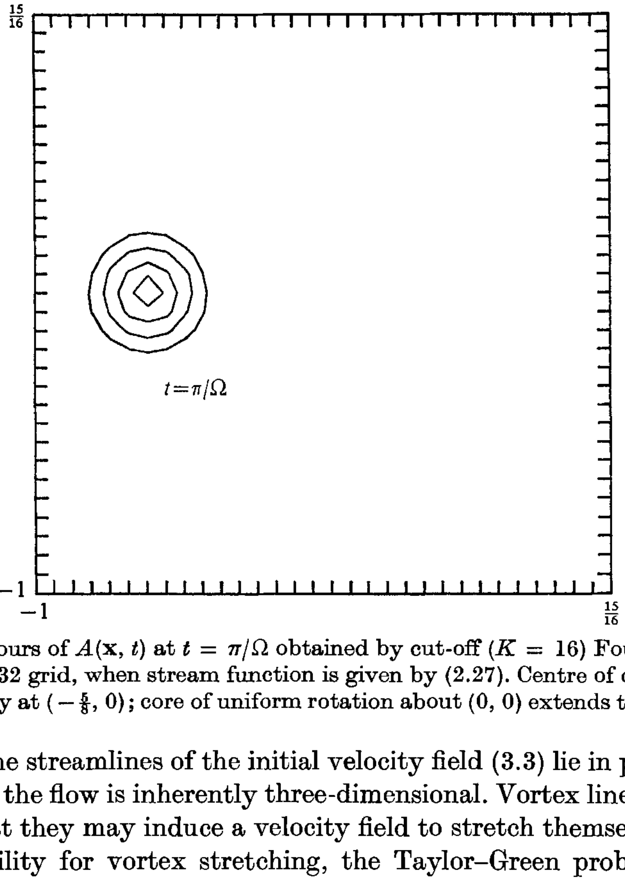


FIGURE 5. Contours of $A(\mathbf{x}, t)$ at $t = \pi/\Omega$ obtained by cut-off ($K = 16$) Fourier-expansion scheme on 32×32 grid, when stream function is given by (2.27). Centre of cone of radius $\frac{1}{4}$ is placed initially at $(-\frac{1}{2}, 0)$; core of uniform rotation about $(0, 0)$ extends to radius $a = \frac{3}{8}$.

Although the streamlines of the initial velocity field (3.3) lie in planes normal to the x_3 axis, the flow is inherently three-dimensional. Vortex lines are initially twisted so that they may induce a velocity field to stretch themselves. Because of the possibility for vortex stretching, the Taylor–Green problem provides insight into the mechanism of enhanced energy dissipation in a turbulent flow. In fact, this was the motivation for new work on the problem (Orszag, manuscript in preparation). However, the importance of the Taylor–Green problem for the present paper lies in the fact that the motion involves an energy cascade, so that appreciable 'aliasing' (cf. §4) errors may result.

The numerical approximations to (3.1)–(3.3) that are compared for accuracy here are the following:

(i) *Second-order centred $2\Delta x$-difference scheme*

Here the velocity and the pressure are recorded at the grid points $i\Delta x_1$, $j\Delta x_2$, $k\Delta x_3$, where Δx_1, Δx_2, Δx_3 specify the grid intervals in the x_1, x_2, x_3 directions, respectively. The centred $2\Delta x$-difference approximation to $\partial p/\partial x_1$, e.g. is

$(\partial p/\partial x_1)_{ijk} \approx (p_{i+1,j,k}-p_{i-1,j,k})/(2\Delta x_1)$, with the difference approximation to the other terms in the Navier–Stokes equations constructed analogously. The pressure field is determined using the fast Fourier transform to solve exactly the finite-difference analog of the Poisson equation $\nabla^2 p = -\boldsymbol{\nabla}.(\mathbf{v}.\boldsymbol{\nabla}\mathbf{v})$ [which follows from (3.1) using (3.2)]. This scheme for numerical solution of the Navier–Stokes equations is second-order. It is well known that scheme (i), described here, may exhibit aliasing instabilities (Phillips 1959). However, if the Reynolds number is sufficiently small, then no instability is observed using leapfrog time differencing for the convective and pressure terms, and an explicit forward time step (with error $O(\Delta t)$) for the viscous term. Since the velocity and length scales of the initial velocity field (3.3) are each of order 1, it follows that a convenient Reynolds number for the Taylor–Green vortex is simply

$$R = 1/\nu.$$

(ii) *Second-order staggered-mesh scheme*

The staggered-mesh scheme (Fromm 1963; Fromm & Harlow 1963; Lilly 1964, 1965; Harlow & Welch 1965; Orszag 1969; Williams 1969; Deardorff 1970), velocities are defined at cell boundaries and pressures at cell centres. Thus, v_1 is discretized so that its values are recorded at the points $(i+\frac{1}{2})\,\Delta x_1, j\Delta x_2, k\Delta x_3$, v_2 at the points $i\Delta x_1$, $(j+\frac{1}{2})\,\Delta x_2$, $k\Delta x_3$, v_3 at $i\Delta x_1$, $j\Delta x_2$, $(k+\frac{1}{2})\,\Delta x_3$, and p at $i\Delta x_1, j\Delta x_2, k\Delta x_3$. The staggered-mesh approximation to the Navier–Stokes equations is written out in the references cited above. The resulting scheme is second-order, and has a number of attractive features in comparison with the centred $2\Delta x$-difference approximation. The staggered-mesh approximation semi-conserves momentum and energy, viz.

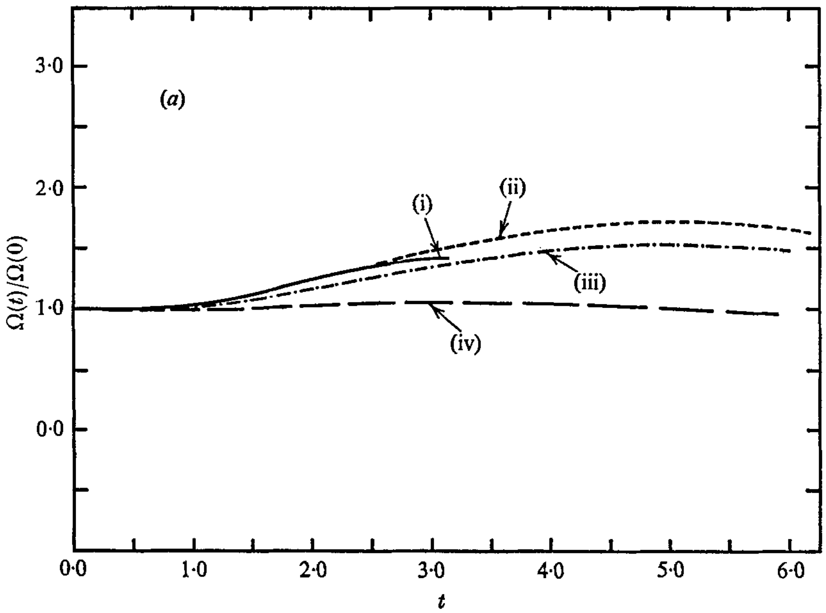
$$\sum_{i,j,k} (v_1, v_2, v_3), \quad \tfrac{1}{2} \sum_{i,j,k} (v_1^2+v_2^2+v_3^2),$$

are conserved in the absence of viscous dissipation and time-differencing errors. The semi-conservation of energy assists in the numerical stability of the scheme. Also, since derivatives in the staggered mesh are approximated by differences over the mesh lengths Δx_1, Δx_2, Δx_3, while in the $2\Delta x$-difference scheme derivatives are approximated by differences over $2\Delta x_1$, $2\Delta x_2$, $2\Delta x_3$, errors are quantitatively reduced by the staggered mesh. The improved accuracy due to the compactness of the staggered mesh is confirmed by the results stated below.

(iii) *Cut-off Fourier-expansion method*

This method involves solution of the Fourier-transformed Navier–Stokes equations using the initial conditions (3.3). The formulation and methods for solution of these equations is described in detail elsewhere (Orszag 1971*a*).

A sensitive test of the accuracy of the solution is given by the quantity $\Omega(t)$, defined as the mean-square vorticity at time t, where the mean is taken as a spatial average over the cube of periodicity. The dominant contribution to $\Omega(t)$ comes from small-scale structures. The ratio $\Omega(t)/\Omega(0)$, which measures the enhancement of energy dissipation due to the cascade to small scales, is plotted



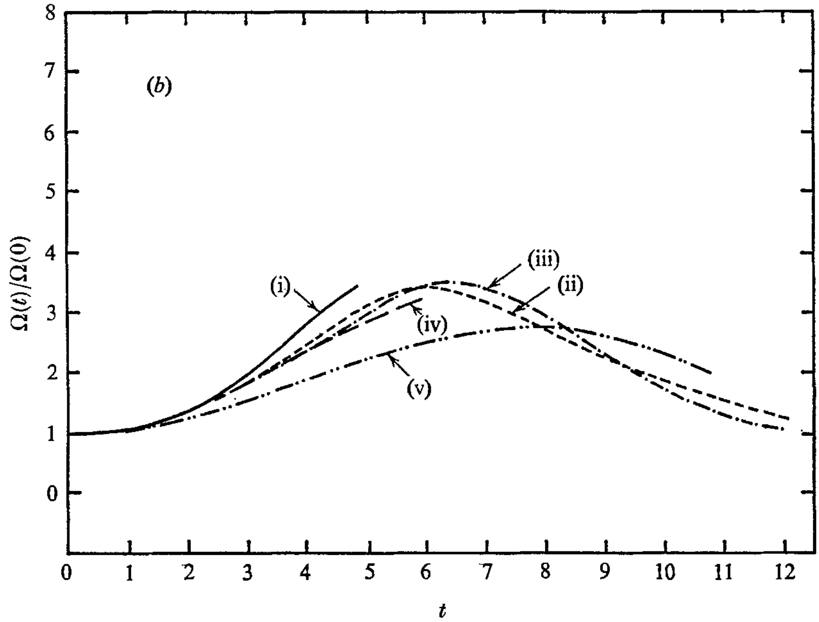


FIGURE 6. Enhancement of mean-square vorticity, $\Omega(t)/\Omega(0)$, *vs.* time for the Taylor–Green vortex. (*a*) $R = 100$: (i) Taylor & Green (1937), using perturbation expansions in powers of time. (ii) cut-off ($K = 8$ *and* $K = 16$) Fourier-expansion method; (iii) staggered-mesh scheme on a $16 \times 16 \times 16$ grid; (iv) centred $2\Delta x$-difference scheme on a $16 \times 16 \times 16$ grid. (*b*) $R = 200$: (i) Taylor & Green (1937), using perturbation expansions in powers of time; (ii), (iii) cut-off Fourier-expansion method with $K = 16$, 8, respectively; (iv), (v) staggered-mesh scheme on $32 \times 32 \times 32$ and $16 \times 16 \times 16$ space grids, respectively.

as a function of time in figures 6(*a*), (*b*). In the calculations summarized in figure 6(*a*), the evolution of the Taylor–Green vortex is calculated at a Reynolds number $R\ (=1/\nu)=100$, while in figure 6(*b*) the Reynolds number of the calculations is 200. The curves labelled (i) in both figures show the results obtained by Taylor & Green (1937), using perturbation expansions in power of t. For t small, the perturbation-theory results should be asymptotic, but they break down before the time that the curves (i) break off in the figures.

In figure 6(*a*), curve (ii) shows the results obtained at $R=100$ using the cut-off Fourier-expansion method with cut-offs $K=8$ and $K=16$. The results obtained with these two cut-offs are not distinguishable on the scale of the graph, so it is plausible to consider both sets of results close to exact. The calculation with $K=8$ has about as many degrees of freedom as a finite-difference calculation on a $16\times16\times16$ grid, while $K=16$ corresponds to a $32\times32\times32$ grid. Curve (iii) is plotted using the results obtained by the second-order staggered-mesh scheme on a $16\times16\times16$ grid, while curve (iv) involves the centred $2\Delta x$-difference scheme on a $16\times16\times16$ grid. The results plotted in figure 6(*a*) show that, on a $16\times16\times16$ grid at the relatively low Reynolds number $R=100$, the $2\Delta x$-difference scheme gives the energy dissipation rate incorrect by nearly 100 % when this rate is near its maximum, while the staggered-mesh scheme is off by about 25 %, and the Fourier-space calculations are nearly exact. Even at $R=50$, the $2\Delta x$-difference results on a $16\times16\times16$ grid are in substantial error.

In figure 6(*b*), curve (ii) is a plot of the results obtained using the cut-off Fourier-expansion scheme with $R=200$ and $K=16$, while curve (iii) is for the Fourier method with $K=8$. Curves (iv), (v) are for the staggered-mesh scheme on $32\times32\times32$ and $16\times16\times16$ grids, respectively. (Curve (iv) was obtained some time ago using the computer facility at the Goddard Institute for Space Studies, New York. The curve was not continued beyond the maximum of $\Omega(t)/\Omega(0)$ at that time.)

It seems safe to infer from the results presented in figures 6(*a*), (*b*) that: (*a*) for given numbers of degrees of freedom, the cut-off Fourier-expansion method is more accurate than finite-difference methods; (*b*) achievement of the accuracy of the Fourier method requires at least twice as many grid points in *each* space direction for the staggered-mesh scheme, and twice again as many for the $2\Delta x$-difference scheme; and (*c*) for the same cell size, the staggered mesh gives more accurate results than obtained by $2\Delta x$-differences.

Other numerical schemes have been tested and compared with the cut-off Fourier-expansion method. Instead of using a second-order form for the dissipation term $\nu\nabla^2\mathbf{v}$ in (3.1), it was thought that a fourth-order finite-difference approximation to $\nabla^2\mathbf{v}$, together with the second-order staggered-mesh approximation to the other terms in (3.1), would improve the results, especially for the small scales that dominate $\Omega(t)$. The results of numerical experiment showed only very slight improvement.

The results of §2 suggest that a fourth-order approximation to (3.1) would improve the results. As yet, there is no published version of the staggered-mesh scheme, which is fourth-order and also semi-conserves the quadratic energy integral. On the other hand, it is not difficult to construct centred difference approximations to $-\mathbf{v}.\nabla\mathbf{v}-\nabla p$ that are fourth order and energetically conservative. One way to achieve this is to rewrite (3.1) in rotation form as

$$\partial\mathbf{v}(\mathbf{x},t)/\partial t = \mathbf{v}(\mathbf{x},t)\times\boldsymbol{\omega}(\mathbf{x},t)-\nabla\Pi(\mathbf{x},t)+\nu\nabla^2\mathbf{v}(\mathbf{x},t), \tag{3.4}$$

where $\Pi = p+\frac{1}{2}v^2$ is the total head, and $\boldsymbol{\omega} = \nabla\times\mathbf{v}$ is the vorticity. This form of the Navier–Stokes equations was apparently first suggested for numerical work by Shuman (1969). Energy is semiconserved as long as the finite-difference approximation to the right-hand side of (3.4) is consistent with that to (3.2) in the sense that $\int\mathbf{v}.\nabla\Pi\,d\mathbf{x} = \int\nabla.(\mathbf{v}\Pi)\,d\mathbf{x} = 0$ even in finite-difference form and $\mathbf{v}\times\boldsymbol{\omega}$ is approximated as the pointwise cross-product of $\mathbf{v}$ with a finite-difference approximation to $\boldsymbol{\omega}$. For $\mathbf{v}.(\mathbf{v}\times\boldsymbol{\omega}) = 0$, the rotation form of the advection term in (3.1) conserves energy, but not momentum, pointwise. It is straightforward to make a fourth-order centred finite-difference approximation to (3.4). A numerical experiment using such a scheme gave results whose accuracy was roughly midway between the second-order $2\Delta x$-difference and staggered-mesh schemes. The increased accuracy due to the compactness of the staggered mesh evidently outweighs the advantages of fourth-order over second-order schemes, at least in the present application. It may be that fourth-order schemes are inappropriate for solution of incompressible flow problems with energy cascade. Although fourth-order schemes decrease first- and second-differencing errors (§4), they may increase aliasing errors just because first derivatives are evaluated more exactly than in a second-order scheme (Grammeltvedt 1969; Lilly 1965).

4. Analysis of errors in simulation

In §4 we attempt to explain why the cut-off Fourier-expansion method gives results that are more accurate than those obtained by finite-difference methods on similar grids. There are at least five errors of simulation that may be isolated (following, in part, Lilly 1965). They are as follows:

(i) *First-differencing* (*phase*) *errors*

The approximation of $\partial A/\partial x$ by the finite-difference approximation,

$$\partial A/\partial x \approx [A(x+\Delta x)-A(x-\Delta x)]/(2\Delta x), \tag{4.1}$$

is not exact. For example, if $A(x) = \exp(ikx)$, then the ratio of the right-hand side of (4.1) to the left-hand side is $\sin(k\Delta x)/(k\Delta x)$, which approaches 1 only for $k\Delta x\to 0$. For large k and finite Δx, the errors in (4.1) are appreciable.

The effect of first-differencing errors on the solution of a partial differential equation is usually illustrated by the one-dimensional advection equation,

$$\partial A(x,t)/\partial t+U\,\partial A(x,t)/\partial x = 0, \tag{4.2}$$

where U is a constant uniform convecting velocity (Thompson 1961). A solution to (4.2) is

$$A(x,t) = \exp[ik(x-Ut)]. \tag{4.3}$$

In fact, if the values of $A(x,t)$ are given on N discrete points equally spaced by Δx, the one-dimensional analogs of (2.17), (2.18) show that an arbitrary excitation on these N points may be represented as a linear combination of the N solutions (4.3) with $k\Delta x = 2\pi m/N(-\frac{1}{2}N \leqslant m < \frac{1}{2}N)$, m integral (using $\cos(k(x-Ut))$ instead of (4.3) if $m = -\frac{1}{2}N$).

The leapfrog centred $2\Delta x$-difference approximation to (4.2) is

$$A_j^{n+1} = A_j^{n-1} - \alpha(A_{j+1}^n - A_{j-1}^n), \tag{4.4}$$

where $A_j^n = A(j\Delta x, n\Delta t)$, Δt is the time increment, and $\alpha = U\Delta t/\Delta x$. The finite-difference equation (4.4) is satisfied by

$$A_j^n = \exp[ikj\Delta x - in\theta], \tag{4.5}$$

where $\sin\theta = \alpha\sin(k\Delta x)$. If $|\alpha| \leqslant 1$, then θ is real for all k. In each time step, the exact change of phase of $A(x,t)$ given by (4.3) is $-kU\Delta t = -k\alpha\Delta x$, while the change of phase of (4.5) is $-\theta = -\sin^{-1}[\alpha\sin(k\Delta x)]$. Since $|\theta/(k\alpha\Delta x)| < 1$ when $|\alpha| < 1$ and $|k\Delta x| \leqslant \pi$, it follows that the waves (4.5) *lag* the true waves (4.3).[†]

$k\Delta x$	Second-order Arakawa scheme	Fourth-order Arakawa scheme	Cut-off Fourier-expansion scheme
15°	0·989	1·000	1·000
30°	0·955	0·998	1·000
45°	0·901	0·989	1·001
60°	0·828	0·966	1·002
90°	0·637	0·851	1·004
120°	0·414	0·622	1·007
150°	0·191	0·310	1·012
180°	0	0	1·017‡

† The values of $\delta = \theta/(k\alpha\Delta x)$ are given for $\alpha = U\Delta t/\Delta x = 0{\cdot}1$.
‡ For $k\Delta x = 180°$, $\delta = 0$. The number listed is $\lim_{k\Delta x\uparrow\pi}\delta$. See text.

TABLE 2. Phase errors (δ)†

Each of the three schemes of §2 when applied to (4.2) with leapfrog time differencing is neutrally stable, in the sense that solutions of the form (4.5) exist with θ real (provided $|\alpha|$ is small enough). The phase error, $\delta = \theta/(k\alpha\Delta x)$, for each of the three schemes of §2 is listed in table 2 as a function of $k\Delta x$ in the special case $\alpha = 0{\cdot}1$. For example, the cut-off Fourier-expansion method applied to (4.2) gives

$$A(k, t+\Delta t) = A(k, t-\Delta t) - \begin{cases} 2ikU\Delta t A(k,t) & \text{if} \quad |k\Delta x| < \pi, \\ 0 & \text{if} \quad k\Delta x = -\pi \end{cases} \tag{4.6}$$

so that $\sin\theta = k\alpha\Delta x$ for $|k\Delta x| < \pi$. The values of δ listed in table 2 show that the fourth-order scheme has smaller phase errors than the second-order scheme, and that the cut-off Fourier-expansion scheme has phase errors of less than 2% for all waves satisfying $|k\Delta x| < \pi$ when $\alpha = 0{\cdot}1$. The so-called '$2\Delta x$' wave (*viz.* $k\Delta x = \pm\pi$ so that the wavelength is $2\Delta x$) is stationary ($\delta = 0$) in all the methods. The $2\Delta x$ wave is stationary even in the Fourier method, because the expansion function for $k\Delta x = \pm\pi$ is $\cos(kx)$ not $\exp(ikx)$.

It should be noticed that the second-order scheme has lagging phase error ($\delta < 1$), while the fourth-order scheme has mostly lagging errors (there is a

[†] See also Orszag (1971*c*).

region of leading phase error $\delta > 1$ for small $k\Delta x$ for the fourth-order scheme). As noted in §2, the lagging phase errors provide a partial explanation for the wakes of bad numbers given by finite-difference approximations.

The cut-off Fourier-expansion scheme has only leading phase errors for finite Δt, and even these disappear in the limit $\Delta t \to 0$ (if $|k\Delta x| < \pi$). The latter property is, of course, intimately related to the fact that the cut-off Fourier-expansion method has spatial errors that decrease more rapidly than any finite power of Δx as $\Delta x \to 0$ (if the exact solution that is being approximated is infinitely differentiable). Finite-difference approximations have non-zero phase errors in the limit $\Delta t \to 0$.

It is well known that numerical stability of (4.4) requires $|\alpha| \leqslant 1$ (since there exist solutions (4.5) with k real and $\mathrm{Im}\,(\theta) > 0$ when $|\alpha| > 1$). Similarly, numerical stability of centred (or Arakawa) difference approximations with spatial error of order $(\Delta x)^n$ may be shown to require $|\alpha| \leqslant 1$ ($n = 2$), 0·73 ($n = 4$), 0·63 ($n = 6$), 0·59 ($n = 8$). As $n \to \infty$, the stability limit approaches $|\alpha| \leqslant 1/\pi$ (slowly like $n^{-\frac{1}{2}}$), which is in fact the stability limit for the Fourier method. Thus, for stable numerical simulation of convective processes using leapfrog time differencing, the Fourier method requires a time step π times smaller than a second-order scheme with the same spatial resolution.† This situation is improved somewhat in the presence of dissipation. The most unstable waves for difference approximations with spatial error of order $(\Delta x)^n$ are $k\Delta x \simeq 1{\cdot}57$ ($n = 2$), 1·80 ($n = 4$), 1·94 ($n = 6$), 2·01 ($n = 8$), while the most unstable mode for the Fourier method is $k\Delta x \simeq 3{\cdot}14$. Since dissipation selectively damps high-frequency modes, the 'most dangerous' modes are preferentially damped in the Fourier method (being careful to treat the dissipation terms implicitly, so that they cannot cause numerical instabilities).

(ii) *Second-differencing errors*

These errors are associated with incorrect evaluation of $\nu\nabla^2\mathbf{v}$ by finite differences. Reduction of these errors is possible by use of a fourth-order finite-difference approximation to $\nabla^2\mathbf{v}$, though, as noted in §3, little quantitative benefit ensues.

There are no second-differencing errors in the Fourier-expansion method as the transform of $\nu\nabla^2\mathbf{v}$ is exactly $-\nu k^2\mathbf{u}(\mathbf{k})$. The diagonal algebraic form $-\nu k^2\mathbf{u}(\mathbf{k})$ makes it a simple matter to treat viscous dissipation implicitly in time in the Fourier method. However, this latter advantage is slight, because high-Reynolds-number flow simulations are usually limited by convective rather than diffusive stability criteria. Some comments on the treatment of non-constant diffusion coefficients (e.g. eddy viscosity coefficients) are made in §6.

(iii) *Incompressibility errors*

These errors are due to incorrect imposition of the supplementary constraint (3.2). In all the simulations reported here, the finite-difference form of (3.2) is

† *Note added in proof.* Implicit time-differencing methods suitable for Galerkin approximations will be discussed in a later paper. The crucial fact is that convective numerical instabilities originate from the convection of small scales by large ones, an effect that is conveniently isolated and treated implicitly in the Galerkin (collective co-ordinate) framework.

imposed exactly (except for round-off error) by use of the fast Fourier transform, to solve the finite-difference Poisson equation for p. Likewise, the cut-off Fourier-expansion simulations impose incompressibility exactly. Round-off error never accumulates appreciably in the simulations reported here, because of the high digital accuracy of the CDC 6600, so that the methods developed by Harlow & Welch (1965) and Piacsek & Williams (1970) to reduce incompressibility errors need not be applied. Incompressibility errors never exceed 1 part in 10^{11} in single-precision calculations.

(iv) *Boundary errors*

Incorrect imposition of boundary conditions is a frequent cause of poor simulation. In the simulations reported in this paper, only periodic boundary conditions are applied and these are applied exactly. However, the spirit of the Galerkin method is to ensure that boundary conditions are imposed correctly (Orszag 1971*a*, *b*), which is not always possible with finite-difference methods. The problems are particularly acute if the equations are essentially hyperbolic so that boundary errors are not damped as they propagate into the domain of flow. Clearly, the nature of boundary errors in both finite-difference and Galerkin schemes requires much further investigation.

(v) *Aliasing errors and blocking*

The nature of these errors is most elusive and, in the author's opinion, has not been given adequate explanation in the literature. It is hoped that the discussion below will clarify some of the salient features of aliasing, though it is clear that this discussion will not be the last word on the subject.

The values of the arbitrary functions $f(x)$, $g(x)$ at N equally spaced grid points $x_n = 2\pi n/N$ $(n = 0, 1, \ldots, N-1)$ are expansible in the discrete Fourier series (cf. (2.17)),

$$\left.\begin{aligned} f(x_n) \equiv f_n &= \sum_{|k|\leqslant K} u(k)\exp(ikx_n), \\ g(x_n) \equiv g_n &= \sum_{|k|\leqslant K} v(k)\exp(ikx_n), \end{aligned}\right\} \qquad (4.7)$$

where k is an integer satisfying $-K \leqslant k < K$ with $N = 2K$ or $-K \leqslant k \leqslant K$ with $N = 2K+1$. The product function $h(x) = f(x)g(x)$ has the value $h_n = f_n g_n$ at x_n, and the expansion

$$h_n = \sum_{|k|\leqslant K} w(k)\exp(ikx_n) \quad (n = 0, \ldots, N-1), \qquad (4.8)$$

where
$$w(k) = \sum_{\substack{p+q=k \\ |p|,\,|q|\leqslant K}} u(p)\,v(q) + \sum_{\substack{p+q=k+N \\ |p|,\,|q|\leqslant K}} u(p)\,v(q) + \sum_{\substack{p+q=k-N \\ |p|,\,|q|\leqslant K}} u(p)\,v(q). \qquad (4.9)$$

The last two terms in (4.9) arise because $\exp(ik'x_n) = \exp(ikx_n)$ for all $n = 0, \ldots, N-1$ if $k' \equiv k$ (modulo N).

Aliasing is usually explained by noting that $f(x)$, $g(x)$, $h(x)$ have the *exact* Fourier expansions

$$(f(x), g(x), h(x)) = \Sigma(\overline{u}(k), \overline{v}(k), \overline{w}(k))\exp(ikx), \qquad (4.10)$$

where the sum is taken over all integers k, $0 \leqslant x < 2\pi$, and

$$\overline{w}(k) = \sum_{p+q=k} \overline{u}(p)\,\overline{v}(q). \qquad (4.11)$$

(It should be noted that $\bar{u}$, $\bar{v}$, $\bar{w}$ are, in general, not equal to u, v, w, respectively.) It is apparent from (4.11) that the exact $\bar{w}(k)$ includes no counterpart of the last two terms in the expression (4.9) for $w(k)$. The false interactions included in the last two terms in (4.9) are called aliasing interactions; $k+N$ and $k-N$ are aliases of k on the discrete grid x_n, since $\exp[i(k\pm N)x_n] = \exp[ikx_n]$ for $n = 0, \ldots, N-1$. A finite-difference approximation to the Navier–Stokes equations (3.1) may lead to difficulty, because of aliasing errors in the discrete-grid approximation of the (scalar) product of $\mathbf{v}$ with $\nabla\mathbf{v}$. Phillips (1959) gives a simple example where aliasing terms in the representation of a product induce an instability not present without aliasing.

The foregoing discussion of aliasing is deficient in several respects, notably the reason for believing that aliasing is an error. There is nothing in the above argument that explains why the last two terms in (4.9) lead to inaccuracies of a finite-difference scheme, and why the simulation would be more accurate if these terms did not appear. In fact, if the differential equation to be solved is

$$\partial v(x,t)/\partial t = -[v(x,t)]^2, \tag{4.12}$$

in which x is just a parameter, the presence of aliasing 'errors' ensures that a finite-difference approximation gives exact results except for time-differencing errors. The finite-difference approximation,

$$\partial v(x_n,t)/\partial t = -[v(x_n,t)]^2,$$

is obviously exact, since there is no propagation from point to point. If the aliasing terms were dropped in the respresentation (4.9) of the Fourier coefficients of $[v(x_n,t)]^2$, the numerical approximation to the solution of (4.12) would not be exact in the limit $\Delta t \to 0$. In other words, significant errors are possible in an alias-free numerical scheme. While Phillips's example shows that aliasing interactions may induce instability, it is also possible to construct an example of a bounded system in which leaving out the aliasing interactions induces instability!

A more convincing explanation of the error involved in aliasing is based on an extension of an argument due to Platzman (1961). We consider the 'best' approximation to $h(x) = f(x)g(x)$ given only certain finite amount of information about $f(x)$ and $g(x)$. If the function values $f(x_n)$ and $g(x_n)$ on a discrete grid are known *exactly*, then the best approximation to $h(x)$ on the same discrete grid is obviously $h(x_n) = f(x_n)g(x_n)$. However, in the simulation of solutions to partial differential equations in which information can propagate from point to point (as in (3.1), because of the presence of spatial derivatives), the values of $f(x_n)$ and $g(x_n)$ determined by a finite-difference scheme are usually not exact. In this case, the 'best' approximation to $h(x)$ at the grid points x_n need not be $f(x_n)g(x_n)$.

When the function values $f(x_n)$ and $g(x_n)$ may be inexact, it is reasonable to seek the 'best' approximation to $h(x) = f(x)g(x)$ as the best approximation in mean square to $h(x)$. By (4.7), the function values $f(x_n)$, $g(x_n)$ on the discrete grid determine values of the Fourier coefficients $u(k)$, $v(k)$ for $|k| \leqslant K$. The best mean-square approximation to $h(x)$ is not determined unless something is known about the Fourier coefficients $u(k)$, $v(k)$ for $|k| > K$. It is convenient to postulate that the finite-difference (or Galerkin) method determines $u(k)$, $v(k)$ for $|k| \leqslant K$ (which need not be the exact $\bar{u}(k)$, $\bar{v}(k)$), but that the Fourier coefficients for $|k| > K$ are random. It is assumed that $u(k)$, $v(k)$ for $|k| > K$ have zero mean and are distributed independently for distinct k subject only to the reality constraint $u(k) = [u(-k)]^*$. The latter statement about the distribution of $u(k)$, $v(k)$ for $|k| > K$ is reasonable in view of the assumption that no information other than $u(k)$, $v(k)$ for $|k| \leqslant K$ is known about $f(x)$, $g(x)$ from the numerical simulation.

With these assumptions, the best average approximation $\hat{h}(x)$ to $f(x)\,g(x)$ is found by minimizing the quantity,

$$I = \left\langle \int |\hat{h}(x) - (\Sigma u(p) \exp(ipx))\,(\Sigma v(q) \exp(iqx))|^2\, dx \right\rangle, \tag{4.13}$$

where $\langle \ldots \rangle$ denotes an average over $u(k)$, $v(k)$ for $|k| > K$. Setting

$$\hat{h}(x) = \Sigma \hat{w}(k) \exp(ikx),$$

it follows that

$$\begin{aligned} I &= \Sigma \langle |\hat{w}(k) - \sum_{p+q=k} u(p)\,v(q)|^2 \rangle \\ &= \Sigma \{ |\hat{w}(k) - \sum_{\substack{p+q=k \\ |p|,\, |q| \leqslant K}} u(p)\,v(q)|^2 \\ &\quad - 2\,\mathrm{Re}\, [(\hat{w}(k) - \sum_{\substack{p+q=k \\ |p|,\, |q| \leqslant K}} u(p)\,v(q))^* \langle \sum_{\substack{p+q=k \\ |p| \text{ or } |q| > K}} u(p)\,v(q) \rangle] + \langle | \sum_{\substack{p+q=k \\ |p| \text{ or } |q| > K}} u(p)\,v(q)|^2 \rangle \}, \\ &= \Sigma [|\hat{w}(k) - \sum_{\substack{p+q=k \\ |p|,\, |q| \leqslant K}} u(p)\,v(q)|^2 + \langle | \sum_{\substack{p+q=k \\ |p| \text{ or } |q| > K}} u(p)\,v(q)|^2 \rangle], \end{aligned} \tag{4.14}$$

using the assumed properties of $u(k)$, $v(k)$. It is apparent from (4.14) that the $\hat{h}(x)$ that minimizes I has

$$\hat{w}(k) = \sum_{\substack{p+q=k \\ |p|,\, |q| \leqslant K}} u(p)\,v(q) \tag{4.15}$$

for $|k| \leqslant K$, i.e. the alias-free form of the Fourier coefficients (4.9). (Note that $\hat{w}(k)$ is given by (4.15) for all k, so that $\hat{h}(x_n) = \hat{f}(x_n)\,\hat{g}(x_n)$ where

$$\hat{f}(x_n) = \Sigma_{|k| \leqslant K}\, u(k) \exp(ikx_n),$$

etc. However, $\hat{f}(x_n) \neq f(x_n)$, in general.) If only the N data $w(k)$ ($|k| \leqslant K$) are admissible, then the alias-free convolution sum (4.15) gives the best average representation of $f(x)\,g(x)$.

In summary, we find that the alias-free sum (4.15) is, in the special sense outlined above, the best approximation to the Fourier coefficients of the product $f(x)\,g(x)$. In other words, the best mean-square approximation to a product is not necessarily the product of the best mean-square approximations, but rather the alias-free product.

The error induced by aliasing is illustrated by the cone problem studied in § 2. The 'fully aliased' Fourier coefficients of the Jacobian $J(\psi, A)$ are given in analogy to (4.9) by replacing the interaction coefficients $I(\mathbf{k}|\mathbf{p}, \mathbf{q})$ that appear in (2.21) by their aliased counterparts

$$I_A(\mathbf{k}|\mathbf{p}, \mathbf{q}) = \sum_{\substack{m_1 = 0,\, \pm 1 \\ m_2 = 0,\, \pm 1}} I(\mathbf{k} + \mathbf{m}N|\mathbf{p}, \mathbf{q}), \tag{4.16}$$

where $N = 2K$, and $\mathbf{m} = (m_1, m_2)$. The contours obtained by this fully-aliased Fourier-space calculation are shown after $\frac{1}{4}$ revolution in figure 7. The errors at the leading edge of the cone are due to aliasing errors, but may be confused with leading phase errors (which do not exist for this calculation because there are no first-differencing errors). Between $\frac{1}{4}$ and $\frac{1}{2}$ revolution, the fully aliased Fourier-space calculations are subject to non-linear instability, and cannot be continued. Comparison of figure 7 with figure 2(*e*), obtained by the alias-free equation (2.21) with $I(\mathbf{k}|\mathbf{p}, \mathbf{q})$, shows that aliasing is indeed an error.

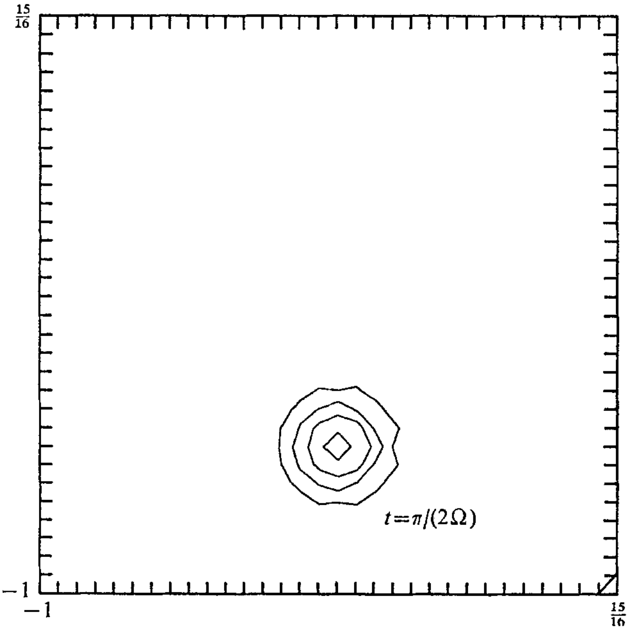


FIGURE 7. Contours of $A(\mathbf{x}, t)$ obtained after $\frac{1}{4}$ revolution using the fully aliased Fourier-space equations on a 32×32 grid (cut-off $K = 16$). Initially, $r = \frac{1}{4}$.

The fully aliased equations may be useful for Navier–Stokes flow simulations. If a fully aliased approximation to the Navier–Stokes equations is obtained by computing $\boldsymbol{\omega}(\mathbf{x}, t)$ using a discrete Fourier series and local physical-space multiplication in $\mathbf{v}(\mathbf{x}, t) \times \boldsymbol{\omega}(\mathbf{x}, t)$ in (3.4), energy is semi-conserved. These fully aliased Fourier-expansion equations have aliasing interactions included at full strength, but there are no first- or second-differencing errors or non-linear instabilities attributable to aliasing.

The energy-conserving finite-difference schemes discussed in §§2, 3 have aliasing errors (Lilly 1965; Grammeltvedt 1969), but they are not susceptible to aliasing instability. As mentioned in §3, aliasing errors usually increase with increasing order of the scheme, just because first derivatives are evaluated more accurately and hence aliasing interactions among high-frequency modes have larger interaction coefficients. The fully aliased interactions in (4.16) are the limiting case of generalizing finite-difference schemes to Nth order on an $N \times \ldots \times N$ space grid.

The alias-free version of the Fourier-expansion method has none of the errors (i)–(iv) listed above, nor does it have aliasing errors, as only convolution sums of the form (4.15) are involved. The only error of the alias-free scheme may be called blocking or damming-up. This is the error involved in truncating the representation (2.14) at K, and disallowing all interactions with wave vectors beyond the cut-off. Evidently, the results of §§2, 3 show that blocking is a less serious error than those suffered by other numerical schemes.

Aliasing and blocking errors limit the Reynolds number at which accurate simulation of flows with energy cascade is possible (cf. e.g. Orszag 1969, §IV*c*). As the Reynolds number increases, aliasing and blocking errors increase, as shown quantitatively for the Taylor–Green problem by comparison of figures 6 (*a*) and 6 (*b*). In order to simulate numerically high-Reynolds-number flows with energy cascade, it is necessary to account for energy transfer to scales smaller than those retained in the simulation ('sub-grid-scale turbulence'), perhaps by use of an eddy viscosity coefficient.

5. Theoretical investigation of accuracy: passive scalar convection

It is possible to give a rather complete mathematical explanation of the improved accuracy of spectral methods for the passive scalar problem (2.1) with the special choice of time-independent convecting velocity field,

$$v_1(x_1, x_2) = 1, \quad v_2(x_1, x_2) = 1 + f(x_1), \tag{5.1}$$

where $f(x+2\pi) = f(x)$ and

$$\int_0^{2\pi} f(x)\, dx = 0 \tag{5.2}$$

(P. D. Lax 1970, private communication). The choice of non-zero constant terms in (5.1) is not crucial to what follows. Since (2.1) states that $A(\mathbf{x}, t)$ is constant on particle orbits, and (5.1), (5.2) imply that the orbit through (x_1, x_2, t) also passes through $(x_1 + 2\pi, x_2 + 2\pi, t + 2\pi)$, it follows that

$$A(x_1 + 2\pi, x_2 + 2\pi, t + 2\pi) = A(x_1, x_2, t). \tag{5.3}$$

If the initial $A(\mathbf{x}, 0)$ field is periodic with period 2π in x_1 and x_2, then (2.1), with the periodic velocity field (5.1), implies that $A(\mathbf{x}, t)$ is periodic in $\mathbf{x}$ for all t. Consequently, (5.3) implies

$$A(\mathbf{x}, t + 2\pi) = A(\mathbf{x}, t) \tag{5.4}$$

for all spatially periodic fields. That is, the initial scalar distribution $A(\mathbf{x}, 0)$ recurs after a time of 2π.

If (2.1) is rewritten in the formal operator form,

$$\partial A(\mathbf{x}, t)/\partial t = -iL(\mathbf{x})\, A(\mathbf{x}, t), \tag{5.5}$$

where the linear operator $L(\mathbf{x})$ accounts for the convective terms, then it follows from (5.4) that $L(\mathbf{x})$ possesses a complete set of eigenfunctions $A_{np}(\mathbf{x})$ satisfying

$$L(\mathbf{x})\, A_{np}(\mathbf{x}) = nA_{np}(\mathbf{x}), \tag{5.6}$$

where the eigenvalue n and label p are both integers. The general periodic solution to the initial-value problem (2.1) with $\mathbf{v}$ given by (5.1) is

$$A(\mathbf{x}, t) = \sum_{n,p} a_{np} A_{np}(\mathbf{x}) \exp(-int), \tag{5.7}$$

where the expansion coefficients a_{np} are determined by $A(\mathbf{x}, 0)$.

Since the convecting velocity (5.1) is independent of x_2, the eigenfunctions $A_{np}(\mathbf{x})$ are of the form,

$$A_{np}(\mathbf{x}) = A_{np}(x_1) \exp(ipx_2), \tag{5.8}$$

where the conflicting notation A_{np} should cause no confusion. The explicit eigenfunction determined by (5.6), (5.8) is

$$A_{np}(x_1) = \exp\left[i(n-p)\,x_1 - ip\int^{x_1} f(y)\,dy\right]. \tag{5.9}$$

Finite-difference and spectral methods for solution of (2.1) with (5.1) may be put in the form (5.5), with $L(\mathbf{x})$ replaced by an approximation $\bar{L}$. While $L(\mathbf{x})$ is clearly an unbounded (differential) operator, each of the approximations $\bar{L}$ are finite-dimensional linear-operator approximations to L. If there are $N_1 N_2$ independent degrees of freedom in the numerical simulation of (2.1) (e.g. N_1 grid points along the x_1 axis and N_2 along the x_2 axis), then the corresponding $\bar{L}$ may be realized as an $(N_1 N_2) \times (N_1 N_2)$ matrix. Each of these matrices has at most $N_1 N_2$ eigenvalues and eigenvectors, in terms of which the general solution to the semi-discrete (i.e. spatially but not temporally discretized) initial-value problem may be expanded, as in (5.7). The faithfulness of these eigenvectors and eigenvalues to those of the exact problem (in the corresponding grid, Fourier, representation) gives a direct measure of the errors in simulation (aside from time-differencing errors). The eigenvalues and eigenvectors of matrix representations of various $\bar{L}$ were determined by the Q–R algorithm (Wilkinson 1965), in order to reach quantitative conclusions of the relative accuracy of various numerical methods.

The homogeneity in x_2 of the convecting velocity field (5.1) gives approximations $\bar{L}$ with eigenvectors whose x_2 dependence (in $\mathbf{x}$ representation) is pure complex exponential as in (5.8). Consequently, the eigenvalues of $\bar{L}$ may be determined independently for independent integral values of the x_2 wavenumber p. This factorization of the matrix $\bar{L}$ is important to obtain matrices of tractable size (viz. $N_1 \times N_1$ instead of $(N_1 N_2) \times (N_1 N_2)$) before application of the Q–R algorithm.

A convenient, though arbitrary, definition of eigenvalues and eigenvectors 'faithful' to those of the exact problem is: (a) an eigenvector is faithful to an eigenvector $A_{np}(\mathbf{x})$ of the exact problem, if the approximate eigenvector lies within 45° of the projection of $A_{np}(\mathbf{x})$ into the space of approximating functions; and (b) the approximating eigenvalue λ_{np} is faithful for q recurrence times of the system if $2\pi q|n-\lambda_{np}| < 1$. The latter definition is reasonable, since q recurrences of the system require a time $2\pi q$, and $2\pi q|n-\lambda_{np}| < 1$ is the condition that the phase error in the mode not exceed 1 radian over this time interval, so that the mode is faithfully accounted for in the approximate form of (5.7).

In figure 8 (*a*), we plot those modes (labelled by n, p) given by the second-order staggered-mesh approximation (2.13) to (2.1) that are faithful to the exact modes with $f(x) = \sin(x)$ in (5.1). All modes lying inside the curves labelled 1 and 10 in figure 8 (*a*) are faithful for $q = 1$ and $q = 10$ recurrences, respectively. These curves were determined from a compilation of eigenvalues and eigenvectors for the numerical scheme applied on 32×32 ($K = 16$), 64×64 ($K = 32$), and 100×100 ($K = 50$) spatial grids, where K is the cut-off frequency in the discrete Fourier representation of the A(x) field. The region of figure 8 (*a*) containing modes faithful for 1 recurrence is approximately given by

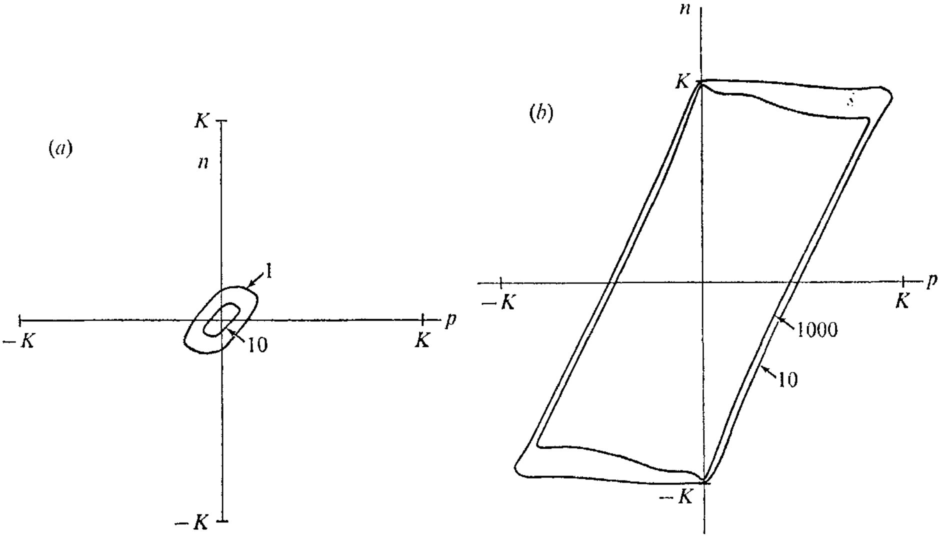


FIGURE 8. Eigenvalues of (*a*) second-order staggered-mesh, (*b*) cut-off Fourier-expansion scheme for passive scalar convection by the periodic convecting velocity (5.1) with $(x) = \sin(x)$. Modes with label (n, p) lying within curve marked 1 in (*a*) are faithful to the exact eigenmodes for $q = 1$ recurrence; that marked 10 in (*a*), (*b*) are faithful for $q = 10$; marked 1000 in (*b*) are faithful for $q = 1000$.

$$|n-p| + |p| \leqslant \tfrac{1}{5}K. \qquad (5.10)$$

For fourth-order schemes, the coefficient $\frac{1}{5}$ in (5.10) is increased to about $\frac{1}{3}$.

The corresponding faithful eigenmodes given by the cut-off Fourier-expansion method are plotted in figure 8 (*b*). The n, p values lying within the curves labelled 10 and 1000 give modes faithful to the exact modes for 100 and 1000 recurrences, respectively (in the absence of time-differencing errors). These regions of faithful modes are quite closely given by

$$|n-p| + |p| \leqslant K. \qquad (5.11)$$

The theoretical basis for (5.11) is explained as follows. With $f(x) = \sin(x)$, the exact eigenmodes are

$$A_{np}(x_1) = \exp[i(n-p)x_1 + ip\cos(x_1)]$$

by (5.9). The Fourier coefficients of $A_{np}(x_1)$ are given exactly in terms of Bessel functions as

$$A_{np}(k) \equiv \frac{1}{2\pi}\int_0^{2\pi} A_{np}(y)\exp(-iky)\,dy = i^{n-p-k}J_{n-p-k}(p). \tag{5.12}$$

Since the cut-off Fourier-expansion method includes only modes up to the cut-off K, accurate representation of the eigenmode $A_{np}(x_1)$ requires that $A_{np}(k)$ be rapidly decreasing and small for $|k| \geqslant K$. However, the Fourier coefficients (5.12) decrease rapidly with $|k|$ for $|k+p-n| \geqslant |p|$ provided $|k| \geqslant |n-p|$. It is easy to see that this condition for rapid decrease of the Fourier coefficients of $A_{np}(x_1)$ is satisfied if (5.11) holds. Consequently, the eigenvectors and eigenvalues should be accurately described by the cut-off Fourier-expansion scheme if (5.11) is satisfied.

The results just described for $f(x) = \sin(x)$ are representative of those for rather general $f(x)$, despite an apparent bias of the choice $\sin(x)$ towards the Fourier method. Results not differing substantially from those stated above were obtained for the choice $f(x) = |x|$ $(-\pi \leqslant x < \pi)$, $f(x+2\pi) = f(x)$, which has a rather slowly converging Fourier series.

The results (5.10), (5.11) imply that second- (fourth-) order finite-difference schemes require at least 5(3) times as many degrees of freedom in *each* space direction as the Fourier-expansion method to achieve the same reasonable accuracy.† The estimated factor 5 is somewhat larger than found in §§2, 3, indicating the conservative nature of these latter estimates. Also, very few modes lie in the region between the $q = 1$ (not drawn) and $q = 1000$ curves in figure 8(*b*). Consequently, aside from time-differencing errors, very few modes are 'lost' by the Fourier method between 1 and 1000 recurrences. About half the total $4K^2$ modes lie within the region (5.11), so that between $t = 0$ and the first recurrence about half the modes (those outside the region (5.11)) are 'lost' due to phase mixing of incorrect frequencies. After that, modes are lost very slowly, so that the calculation hardly deteriorates at all between, say, 1 and 1000 recurrences (except for the increasingly disruptive effect of time-differencing errors that probably make accurate simulation of 1000 recurrences quite impractical). On the other hand, about 98 % of all modes lies outside the curve labelled 1 in figure 8(*a*), so that 98 % of all modes are lost in the second-order simulations before the first recurrence. Thereafter, the noticeable difference between the $q = 1$ and $q = 10$ curves in figure 8(*a*) ensures that modes are continually lost by phase errors in the finite-difference schemes and long-time simulations are ruled out.

The above conclusions seem to hold true for quite general velocity fields without stagnation points (e.g. (5.1)). For convecting velocity fields with stagnation points, the exact eigenvalue spectrum is continuous and the analysis considerably more complicated. The eigenvalue analysis of passive convection by velocity fields with stagnation points will be presented in a later publication.

We have also repeated the 'cone' experiments of §2 using the convecting velocity field (5.1) with $f(x) = \sin(x)$. The cone (2.5) was chosen centred at $x_0 = -\frac{1}{2}\pi$ with base radius $r = \frac{1}{4}\pi$, as shown in figures 1, 2 (when the co-ordinate axes are scaled by π). The effect of the velocity field (5.1) is to convect, squeeze,

† See also Orszag (1971*c*).

and stretch the cone, while (5.4) implies that the circular cone should be reformed at intervals of $t = 2\pi$. The results for $A(\mathbf{x}, 2\pi)$ at a time of 1 recurrence using the second-order staggered-mesh scheme (2.13) on a 32×32 grid are contoured in figure 9(*a*). Quite clearly, the squeezing and stretching has deformed the initial cone almost out of recognition; the grid-point maximum of $A(\mathbf{x}, 2\pi)$ is 0·44 and the maximum is lagging its correct position by $\Delta x_1 = -0{\cdot}6$ (with periodicity interval 2π). At the same time, the cone given by the fourth-order Arakawa scheme on a 32×32 grid is utterly unrecognizable.

The results of $A(\mathbf{x}, 10\pi)$ and $A(\mathbf{x}, \frac{21}{2}\pi)$ after 5 recurrences as determined by the cut-off Fourier-expansion scheme on a 32×32 space grid (cut-off $K = 16$) are contoured in figure 9(*b*). After one recurrence, the grid-point maximum is 0·92; after five, the grid-point maximum is 0·91. There is virtually no lag of the maximum, and the cone has remained quite well localized in space.

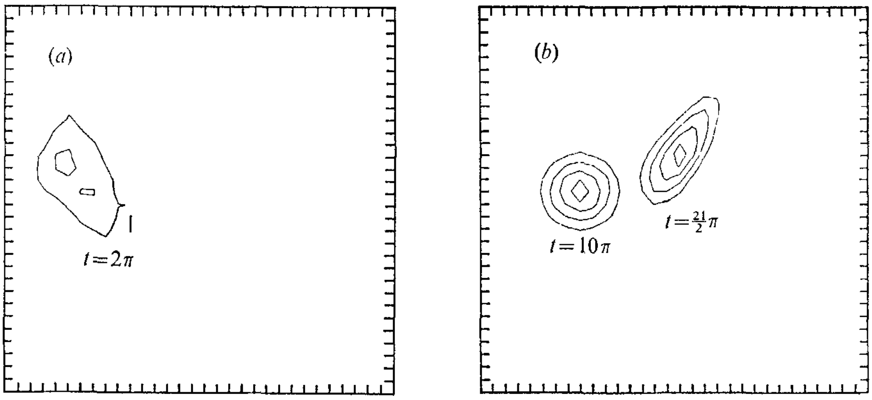


FIGURE 9. Contours of $A(\mathbf{x}, t)$ obtained after (*a*) 1 recurrence using the second-order staggered-mesh scheme on a 32×32 grid (contours are those of $A = 0{\cdot}2, 0{\cdot}4$); (*b*) 5 recurrences ($t = 10\pi$) and $5\frac{1}{4}$ recurrences ($t = \frac{21}{2}\pi$), using cut-off Fourier-expansion scheme on 32×32 grid. Periodic convecting velocity is (5·1) with $f(x) = \sin(x)$; initial radius of the cone is $r = \frac{1}{4}\pi$.

6. Comments and conclusions

In most cases, simulations using the cut-off Fourier-expansion method require more computer time than finite-difference simulations involving the same numbers of independent degrees of freedom. However, in those cases where efficient transform methods have been invented, the relative slowness of the Fourier-expansion simulations is much less than the relative inaccuracy of the finite-difference simulations. In order to achieve a reasonable standard of accuracy, the cut-off Fourier-expansion method requires considerably less resolution (and, hence, computer memory), and somewhat less computer time, than do finite-difference simulations.

For example, the two-dimensional scalar-convection simulations reported in §2 required roughly 0·06 s per time step for the second-order staggered-mesh scheme, 0·08 s per step for the second-order Arakawa scheme, 0·13 s per step for the fourth-order Arakawa scheme, and 0·3 s per step for the cut-off Fourier-expansion scheme, all on 32×32 ($K = 16$) grids. These calculations were performed on the CDC 6600 at the National Center for Atmospheric Research; all the codes were written in Fortran with comparable care. If a machine-language fast-Fourier-transform routine is used, the speed of the cut-off Fourier-expansion method is increased by about a factor 2.

The Taylor–Green vortex-decay calculations reported in §4 are actually performed more efficiently using the Fourier-expansion scheme than by finite-difference methods on the same space grid, because of the symmetries of the vortex (cf. Orszag 1971*a*, §5). On a $32 \times 32 \times 32$ grid (cut-off $K = 16$), the cut-off Fourier-expansion scheme requires just 4·2 s per time step (using a Fortran fast-Fourier-transform routine). However, neglecting the speed-up attributable to the symmetries, the cut-off Fourier-expansion method for solution of the Navier–Stokes equations is about a factor $2\frac{1}{2}$ less efficient than the second-order staggered-mesh scheme for three-dimensional simulations involving the same numbers of degrees of freedom, and about a factor 2 less efficient for two-dimensional simulations involving given numbers of degrees of freedom. The most efficient three-dimensional simulations use the isotropic-truncation transform method of (1971*a*, appendix IV; see also Patterson & Orszag 1971).

The relative inefficiency of the cut-off Fourier-expansion method on grids of the same *a priori* resolution as those used for finite-difference simulations is *not* disqualifying. First, the computation times should be compared on grids such that the overall accuracy of the various simulations are comparable. In this case, the cut-off Fourier-expansion method is decidedly superior. For example, using Fortran programs on the CDC 6600, it takes about 5 s per time step with the cut-off Fourier-expansion method on a $16 \times 16 \times 16$ grid (cut-off $K = 8$), and about 2 s per time step for a $16 \times 16 \times 16$ staggered-mesh calculation, in both cases neglecting any speed-up attributable to symmetries or machine-language fast Fourier transforms. However, the $K = 8$ Fourier-space calculation is at least equivalent in terms of accuracy with a $32 \times 32 \times 32$ staggered-mesh calculation. The latter calculation would require at least $8 \times 2 = 16$ s per time step, which is a very optimistic estimate, because the calculation can no longer be done within the high-speed memory (capacity $\sim 5 \times 10^4$ words) and peripheral devices must be employed to store some of the dynamical variables. Using a machine-language fast Fourier transform, reasonably accurate simulations require nearly an order-of-magnitude less computer time and memory using the Fourier method than using finite-difference methods.

Secondly, the next generation of computers will be much faster than those presently available, but present indications are that the amount of addressable high-speed storage will not be increased very much. Since the amount of high-speed storage, not computer speed, is usually the most limiting factor on computers such as the CDC 6600, it seems that memory requirements will remain critical in the near future. In this case, the cut-off Fourier-expansion method offers the significant advantage that it gives the most accuracy for a given number of degrees of freedom. Thirdly, the fact that the Fourier method gives infinite-order accurate approximations to infinitely differentiable solutions of the equations of motion implies that very accurate (or moderately accurate long-time) simulations are more easily obtained than with finite-difference methods.

Transform methods to speed the evaluation of Galerkin equations are known to apply in the following cases:

(i) Periodic and free-slip boundary conditions on flows within rectangular boundaries, using Fourier expansions (Orszag 1969, 1971*a*; Patterson & Orszag 1971).

(ii) Rigid no-slip or free-slip boundary conditions on flows within slab, spherical, or cylindrical geometries using expansions in Chebyshev polynomials (Orszag 1971*b*). Galerkin (Chebyshev) simulations are also efficiently implementable with stretched (boundary-layer) co-ordinate systems. Flows within rather arbitrary geometries may be efficiently and accurately simulated using Chebyshev expansions after mapping the domain of flow into a simple standard domain.

(iii) Flows on the surface of a sphere using Galerkin approximations obtained from truncated expansions in surface harmonics (Orszag 1970).

(iv) Simulation of sub-grid-scale turbulence using an eddy viscosity coefficient (Smagorinsky 1963) is efficiently accomplished within spectral methods by first using collocation to evaluate the eddy viscosity at selected points in physical space, then using transform methods to evaluate the dissipation term with a non-constant viscosity coefficient. In general, the method of 'pseudo-spectral' approximation advocated in Orszag (1971*a*, §8) consists mainly of the idea of maintaining flexibility between spectral and grid representations. Complicated, highly non-linear, but physically local terms are evaluated locally in physical space, while differentiations are performed locally in spectral representation in order to minimize phase errors.

Another important application of the Fourier-expansion method is also under investigation (Fox, Fulker & Orszag, manuscript in preparation). In this work, the Navier–Stokes equations are solved in slab geometries for a variety of boundary conditions, using a mixed Galerkin-finite-difference method. The horizontal (x_1, x_2) dependence of the motion is Fourier transformed and a Galerkin (Fourier) approximation is made. The vertical (x_3) dependence is treated using a staggered-mesh finite-difference approximation. The resulting numerical scheme has a number of attractive features, including its suitability for buffered simulations in which calculations are performed on one (x_1, x_2) plane at a time, and the property that it semi-conserves energy. The mixed Galerkin-finite-difference method is well suited for major thermal convection and shear flow calculations.

In conclusion, we have demonstrated that Galerkin methods, in particular those using cut-off Fourier-expansions, are an attractive alternative to finite-difference methods for numerical simulation of many of the flows of current interest. The Galerkin equations, coupled with transform methods for their efficient evaluation, offer the advantages of improved accuracy and efficiency in comparison with finite-difference methods.

The author benefited greatly from discussions with Drs J. W. Deardorff, P. D. Lax, C. E. Leith, and D. K. Lilly. The model problem discussed in §5 was suggested by Professor Lax. The author would also like to thank T. J. Wright for assistance with the machine calculations. The computations were performed on the National Center for Atmospheric Research Control Data 6600 computer. This work was partially supported by the Air Force Office of Scientific Research under Contract Number F44620-67-C-0007 at M.I.T. The work was begun at the National Center for Atmospheric Research and was completed there under tenure of an Alfred P. Sloan Research Fellowship. The National Center for Atmospheric Research is sponsored by the National Science Foundation.

Appendix

In the appendix, we indicate the minor modifications in the transform methods of Orszag (1971*a*) that are necessary to evaluate the right-hand side of (2.21). We indicate the modifications only for the transform method given in (1971*a*, appendix III); analogous modifications apply to the transform methods of (1971*a*, §3). Detailed proofs of the algorithms stated in this appendix will be given elsewhere.

The object is to evaluate the right-hand side of (2.21), viz.

$$\overline{w}(\mathbf{k}) = \sum_{\|\mathbf{p}\|\leqslant K} \sum_{\|\mathbf{q}\|\leqslant K} I(\mathbf{k}|\mathbf{p},\mathbf{q})\,\psi(\mathbf{p})\,A(\mathbf{q}) \quad (\|\mathbf{k}\| \leqslant K), \tag{A 1}$$

where we write $I(\mathbf{k}|\mathbf{p},\mathbf{q})$ given by (2.23) in the alternative form

$$I(\mathbf{k}|\mathbf{p},\mathbf{q}) = \begin{Bmatrix} 0 \quad \text{if} \quad \bar{\mathbf{k}} \neq \bar{\mathbf{p}}+\bar{\mathbf{q}} \\ \pi^2 n(p_1)\,n(p_2)\,n(q_1)\,n(q_2)\,[\bar{k}_1\bar{p}_2-\bar{k}_2\bar{p}_1] \quad \text{if} \quad \bar{\mathbf{k}} = \bar{\mathbf{p}}+\bar{\mathbf{q}}, \end{Bmatrix} \tag{A 2}$$

in order to effect some economy in the algorithm. The notation is explained following (2.20) and (2.23). The evaluation of (A 1) essentially involves the evaluation of two convolution sums, one between $i\pi p_2\psi(\mathbf{p})$ and $A(\mathbf{q})$, the other between $i\pi p_1\psi(\mathbf{p})$ and $A(\mathbf{q})$, with the results multiplied by $i\pi\bar{k}_1$ and $i\pi\bar{k}_2$, respectively, to give $\overline{w}(\mathbf{k})$. $I(\mathbf{k}|\mathbf{p},\mathbf{q})$ is rewritten in the form (A 2) so that only one set of transforms of $A(\mathbf{q})$ need be performed. The form (A 2) corresponds to rewriting the dynamical equation (2.3) in conservation form.

The basic result is that the additional terms appearing in the algorithm of Orszag (1971*a*, appendix III) due to wave vectors with components equal to $-K$ should be treated as if the components are equally $+K$, $-K$. Thus, we introduce the three sets of 9 discrete Fourier transforms on $K \times K$ points

$$\begin{Bmatrix} A^{\mathbf{s}}(\mathbf{j}) \\ U^{\mathbf{s}}(\mathbf{j}) \\ V^{\mathbf{s}}(\mathbf{j}) \end{Bmatrix} = \sum_{0\leqslant k\leqslant K} \begin{Bmatrix} a^{\mathbf{s}}(\mathbf{k}) \\ u^{\mathbf{s}}(\mathbf{k}) \\ v^{\mathbf{s}}(\mathbf{k}) \end{Bmatrix} \exp(2\pi i\mathbf{j}\,.\,\mathbf{k}/K) \quad (0 \leqslant \mathbf{j} < K), \tag{A 3}$$

where $0 \leqslant \mathbf{k} < K$ means $0 \leqslant k_\alpha < K$ for $\alpha = 1, 2$, and the nine vectors $\mathbf{s} = (s_1, s_2)$ have components $s_1, s_2 = 0, 1, 2$. The fields $a^{\mathbf{s}}(\mathbf{k})$, $u^{\mathbf{s}}(\mathbf{k})$, $v^{\mathbf{s}}(\mathbf{k})$ $(0 \leqslant \mathbf{k} < K)$ are defined by

$$a^{\mathbf{s}}(\mathbf{k}) = \sum_{\mathbf{r}} \begin{Bmatrix} \exp(-2\pi i r_1 s_1/3) \\ \mathrm{Re}\,[\exp(-2\pi i r_1 s_1/3)] \end{Bmatrix} \begin{Bmatrix} \exp(-2\pi i r_2 s_2/3) \\ \mathrm{Re}\,[\exp(-2\pi i r_2 s_2/3)] \end{Bmatrix} \\ \times A(\mathbf{k}-\mathbf{r}K)\exp[2\pi i\mathbf{s}\,.\,\mathbf{k}/(3K)], \tag{A 4}$$

$$u^{\mathbf{s}}(\mathbf{k}) = \sum_{\mathbf{r}} \left\{\begin{array}{l} i\pi(k_1-r_1K)\exp\left(-2\pi i r_1 s_1/3\right) \\ \mathrm{Re}\,[\ldots] \end{array}\right\} \left\{\begin{array}{l} \exp\left(-2\pi i r_2 s_2/3\right) \\ \mathrm{Re}\,[\ldots] \end{array}\right\}$$

$$\times\, \psi(\mathbf{k}-\mathbf{r}K)\exp\left[2\pi i\mathbf{s}\,.\,\mathbf{k}/(3K)\right], \quad \text{(A 5)}$$

$$v^{\mathbf{s}}(\mathbf{k}) = \sum_{\mathbf{r}} \left\{\begin{array}{l} \exp\left(-2\pi i r_1 s_1/3\right) \\ \mathrm{Re}\,[\ldots] \end{array}\right\} \left\{\begin{array}{l} i\pi(k_2-r_2K)\exp\left(-2\pi i r_2 s_2/3\right) \\ \mathrm{Re}\,[\ldots] \end{array}\right\}$$

$$\times\, \psi(\mathbf{k}-\mathbf{r}K)\exp\left[2\pi i\mathbf{s}\,.\,\mathbf{k}/(3K)\right], \quad \text{(A 6)}$$

where the four vectors $\mathbf{r} = (r_1, r_2)$ have components $r_1, r_2 = 0, 1$. Here the lower term in each of the two curly brackets in (A 4)–(A 6) applies only if the corresponding r_i *and* k_i satisfy $r_i = 1, k_i = 0$. For example, if $\mathbf{s} = (0,0)$ and $0 < k_2 < K$, then $u^{\mathbf{s}}(0, k_2) = 0$; if $\mathbf{s} = (1,1)$ and $0 < k_2 < K$, then

$$u^{\mathbf{s}}(0,k_2) = -K\sin(2\pi/3)\exp\left[2\pi i k_2/(3K)\right]\sum_{r_2=0}^{1}\exp\left(-2\pi i r_2/3\right)\psi(-K, k_2-r_2K).$$

It should be noted that the 27 fields $a^{\mathbf{s}}(\mathbf{k})$, $u^{\mathbf{s}}(\mathbf{k})$, $v^{\mathbf{s}}(\mathbf{k})$ are each half-complex in the sense of Orszag (1971*a*, (III, 5)), since ψ, A satisfy (2.16). Therefore, the 27 discrete Fourier transforms $A^{\mathbf{s}}(\mathbf{j})$, $U^{\mathbf{s}}(\mathbf{j})$, $V^{\mathbf{s}}(\mathbf{j})$ defined by (A 3) are each real.

Finally, it may be shown by straightforward, though lengthy, algebra that

$$\overline{w}(\mathbf{k}) = [9K^2 n(k_1)\,n(k_2)]^{-1}\left(\sum_{\mathbf{s}}\left\{\begin{array}{l} \exp\left[-2\pi i s_1 k_1/(3K)\right] \\ \mathrm{Re}\,[\ldots] \end{array}\right\}\left\{\begin{array}{l} i\pi k_2\exp\left[-2\pi i s_2 k_2/(3K)\right] \\ \mathrm{Re}\,[\ldots] \end{array}\right\}\right.$$

$$\times \sum_{0\leqslant\mathbf{j}<K} A^{\mathbf{s}}(\mathbf{j})\,U^{\mathbf{s}}(\mathbf{j})\exp\left(-2\pi i\mathbf{j}\,.\,\mathbf{k}'/K\right) - \sum_{\mathbf{s}}\left\{\begin{array}{l} i\pi k_1\exp\left[2\pi i s_1 k_1/(3K)\right] \\ \mathrm{Re}\,[\ldots] \end{array}\right\}$$

$$\left.\times\left\{\begin{array}{l} \exp\left[-2\pi i s_2 k_2/(3K)\right] \\ \mathrm{Re}\,[\ldots] \end{array}\right\}\sum_{0\leqslant j<K} A^{\mathbf{s}}(\mathbf{j})\,V^{\mathbf{s}}(\mathbf{j})\exp\left(-2\pi i\mathbf{j}\,.\,\mathbf{k}'/K\right)\right), \quad \text{(A 7)}$$

where $k_i' = k_i$ if $0 \leqslant k_i < K$, $k_i' = k_i + K$ if $-K \leqslant k_i < 0$ $(i = 1, 2)$, and the lower quantities in curly brackets equal the real part of the upper quantity in the corresponding bracket and are to be taken only if the corresponding value of k_i $(i = 1, 2)$ is $-K$. Note that (A 7) involves a total of 18 discrete Fourier transforms of $K \times K$ real data.

Equations (A 3)–(A 7) are the generalization of the transform method of Orszag (1971*a*, appendix III) to the evaluation of (A 1). The modifications of the original algorithm contained in (A 3)–(A 7) are: (*a*) if some components of $\mathbf{k}$ equal $-K$, the phase shifts in (A 4)–(A 7) are taken for those components with real parts only (in order to maintain conditions analogous to (2.16)); and (*b*) if some components of $\mathbf{k}$ equal $-K$, the result for $\overline{w}(\mathbf{k})$ is multiplied by a corresponding factor 2 (as accounted for by the factors $n(k_i)$ in (A 7)). The modifications made in this way also convert the algorithm of Orszag (1971*a*, §3) to the evaluation of (A 1), and they work for quite general convolution sums of the sort encountered in (1971*a*, §§2–5, 7).

The algorithm (A 3)–(A 7) involves 45 real (or half-complex) discrete Fourier transforms on $K \times K$ points in order to evaluate exactly $\overline{w}(\mathbf{k})$ for $\|\mathbf{k}\| \leqslant K$. If the transforms $U^{\mathbf{s}}(\mathbf{k})$, $V^{\mathbf{s}}(\mathbf{k})$ are computed and stored before the start of the evolution calculation using (2.21) (as would be convenient if the convecting velocity field is time independent), then only 27 transforms on $K \times K$ points need be performed each time step. Under the same conditions, the transform method of Orszag (1971*a*, §3) requires 12 real (or half-complex) Fourier transforms on $2K \times 2K$ points, i.e. roughly $\frac{16}{9}$ as many calculations as in the algorithm described here.

REFERENCES

Arakawa, A. 1966 Computational design for long-term numerical integration of the equations of motion: two-dimensional incompressible flow. Part 1. *J. Comp. Phys.* **1**, 119–143.

Arakawa, A. 1970 Numerical simulation of large-scale atmospheric motions. *Numerical Solution of Field Problems in Continuum Physics Proc. SIAM-AMS*, vol. 2, 24–40. Providence: American Mathematical Society.

Burstein, S. Z. & Mirin, A. A. 1970 Third-order difference methods for hyperbolic equations. *J. Comp. Phys.* **5**, 547–571.

Crank, J. & Nicolson, P. 1947 A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. *Proc. Camb. Phil. Soc.* **43**, 50–67.

Crowley, W. P. 1968 Numerical advection experiments. *Mon. Weath. Rev.* **96**, 1–11.

Deardorff, J. W. 1970 A numerical study of three-dimensional turbulent channel flow at large Reynolds numbers. *J. Fluid Mech.* **41**, 453–480.

Deardorff, J. W. 1971 On the magnitude of the subgrid scale eddy coefficient. *J. Comp. Phys.* **7**, 120–133.

Ellsaesser, H. W. 1966 Evaluation of spectral versus grid methods of hemispheric numerical weather prediction. *J. Appl. Meteor.* **5**, 246–262.

Fromm, J. E. 1963 A method for computing nonsteady incompressible, viscous fluid flows. *Los Alamos Sci. Lab. Rep.* LA-2910.

Fromm, J. E. 1969 Practical investigation of convective difference approximations. *Phys. Fluids* (suppl. 2) **12**, 3–12.

Fromm, J. E. & Harlow, F. H. 1963 Numerical solution of the problem of vortex street development. *Phys. Fluids*, **6**, 975–982.

Gaunt, J. A. 1927 The deferred approach to the limit. Part 2. Interpenetrating lattices. *Phil. Trans.* A **226**, 350–361.

Goldstein, S. 1940 Three-dimensional vortex motion in a viscous fluid. *Phil. Mag.* **30**, 85–102.

Grammeltvedt, A. 1969 A survey of finite-difference schemes for the primitive equations for a barotropic fluid. *Mon. Weath. Rev.* **97**, 384–404.

Harlow, F. H. & Welch, J. E. 1965 Numerical calculation of time-dependent viscous incompressible flow of fluid with free surface. *Phys. Fluids*, **8**, 2182–2189.

Lilly, D. K. 1964 Numerical solutions for the shape-preserving two-dimensional thermal convection element. *J. Atmos. Sci.* **21**, 83–98.

Lilly, D. K. 1965 On the computational stability of numerical solutions of time-dependent non-linear geophysical fluid dynamics problems. *Mon. Weath. Rev.* **93**, 11–26.

Lilly, D. K. 1967 The representation of small-scale turbulence in numerical simulation experiments. *Proc. IBM Scientific Computing Symposium on Environmental Sciences*, 195–210. White Plains, N.Y.: IBM.

Molenkamp, C. R. 1968 Accuracy of finite-difference methods applied to the advection equation. *J. Appl. Meteor.* **7**, 160–167.

Orszag, S. A. 1969 Numerical methods for the simulation of turbulence. *Phys. Fluids* (suppl. 2) **12**, 250–257.

Orszag, S. A. 1970 Transform method for the calculation of vector-coupled sums: application to the spectral form of the vorticity equation. *J. Atmos. Sci.* **27**, 890–895.

ORSZAG, S. A. 1971*a* Numerical simulation of incompressible flows within simple boundaries. 1. Galerkin (spectral) representations. *Stud. in Appl. Math.* To be published.

ORSZAG, S. A. 1971*b* Galerkin approximations to flows within slabs, spheres, and cylinders. *Phys. Rev. Letters*, **26**, 1100–1133.

ORSZAG, S. A. 1971*c* On the resolution requirements of finite-difference schemes. *Stud. in Appl. Math.* To be published.

PATTERSON, G. S. & ORSZAG, S. A. 1971 Spectral calculations of isotropic turbulence: efficient removal of aliasing interactions. *Phys. Fluids.* To be published.

PHILLIPS, N. A. 1959 An example of non-linear computational instability. *The Atmosphere and the Sea in Motion*, 501–504. New York: Rockefeller Institute Press.

PIACSEK, S. A. & WILLIAMS, G. P. 1970 Conservation properties of convection difference schemes. *J. Comp. Phys.* **6**, 392–405.

PLATZMAN, G. W. 1961 An approximation to the product of discrete functions. *J. Meteor.* **18**, 31–37.

PRICE, H. S. & VARGA, R. S. 1970 Error bounds for semidiscrete Galerkin approximations of parabolic problems with applications to petroleum reservoir mechanics. *Numerical Solution of Field Problems in Continuum Physics, Proc. SIAM-AMS* vol. 2, 74–94. Providence: American Mathematical Society.

RICHARDSON, L. F. 1910 The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam. *Phil. Trans.* A **210**, 307–357. (Also *Proc. Roy. Soc.* A **83**, 335–336.)

RICHARDSON, L. F. 1927 The deferred approach to the limit. Part 1. Single lattice. *Phil. Trans.* A **226**, 299–349.

ROBERTS, K. V. & WEISS, N. O. 1966 Convective difference schemes. *Math. Comput.* **20**, 272–299.

SHUMAN, F. G. & STACKPOLE, J. D. 1969 The currently operational NMC model, and results of a recent simple numerical experiment. *Proc. WMO/IUGG Symp. on Numerical Weather Prediction*, vol. 2, 85–98. Tokyo: Japan Meteorological Agency.

SMAGORINSKY, J. 1963 General circulation experiments with the primitive equations. Part 1. The basic experiment. *Mon. Weath. Rev.* **91**, 99–164.

TAYLOR, G. I. & GREEN, A. E. 1937 Mechanism of the production of small eddies from large ones. *Proc. Roy. Soc.* A **158**, 499–521.

THOMPSON, P. D. 1961 *Numerical Weather Analysis and Prediction.* Macmillan.

WILKINSON, J. H. 1965 *The Algebraic Eigenvalue Problem.* Oxford University Press.

WILLIAMS, G. P. 1969 Numerical integration of the three-dimensional Navier–Stokes equations for incompressible flow. *J. Fluid Mech.* **37**, 727–750.