Int. J. Heat Mass Transfer. Vol. 25, No. 8, pp. 1137–1146, 1982
Printed in Great Britain

0017–9310/82/081137–10 $03.00/0

A MIXED FINITE DIFFERENCE–GALERKIN PROCEDURE FOR TWO-DIMENSIONAL CONVECTION IN A SQUARE BOX

J. M. McDONOUGH and I. CATTON
School of Engineering and Applied Science, University of California Los Angeles, Los Angeles, CA 90024, U.S.A.

(*Received for publication* 22 *January* 1982)

Abstract—A mixed finite difference–Galerkin method is used to solve the problem of thermal convection in a two-dimensional horizontal square box heated from below. The Galerkin procedure is applied in the horizontal direction; finite differencing is used in the vertical direction. The numerical features of such an approach are compared, theoretically, with those of usual finite difference and Galerkin methods. Specific numerical analytical performance data are given for the mixed finite difference–Galerkin procedure for several values of Prandtl number, and over a range of Rayleigh numbers.

NOMENCLATURE

e_m,	error at mth iteration;
g,	gravitational constant;
h,	finite difference step size;
K,	number of terms in Galerkin representation;
L,	length of side of convection box;
N,	number of points in finite difference grid;
Nu,	Nusselt number;
p,	pressure;
Pr,	Prandtl number;
Ra,	Rayleigh number;
T,	temperature;
v, w,	velocity components;
y, z,	spatial coordinates.

Greek symbols

α,	coefficient of volumetric expansion;
δ,	damping factor in iteration scheme;
Δ,	two-dimensional Laplacian, $\partial_y^2 + \partial_z^2$;
ε,	Newton–Kantorovich convergence tolerance;
κ,	thermal diffusivity;
ν,	kinematic viscosity;
ψ,	stream function.

Subscripts

C,	cold;
H,	hot;
i,	grid point index;
j, k, m,	summation indices;
opt,	optimum value;
n, y, z,	partial differentiation.

Superscripts

*,	perturbation quantity;
(0),	initial guess.

INTRODUCTION

THE PROBLEM which is considered here is fundamental in studies of thermal convection in enclosures. The physical situation is shown in Fig. 1. It consists of a square box of fluid with insulating sidewalls oriented perpendicularly with respect to the local gravitational field, heated on the bottom and cooled on the top. The top and bottom walls are assumed to be perfectly conducting in the present treatment. It is well known, both from theory [1] and from experiment [2, 3] that until a certain critical temperature difference is exceeded, heat is transferred vertically from bottom to top only by conduction. But after the critical temperature difference is surpassed, convection begins, with a significant increase in the overall heat transfer.

The basic problem we study here has been considered by numerous investigators. Finite difference solutions have been given, for example, by de Vahl Davis [4] and Wilkes and Churchill [5]. The Galerkin procedure was used by Catton *et al.* [6] among others. In addition, Denny and Clever [7] have provided a comparison of Galerkin and finite differencing for a quite similar problem. We are using this widely-studied problem as a model in presenting a somewhat different numerical approach, a mixed finite difference–Galerkin procedure.

This method has been used previously by Mc-

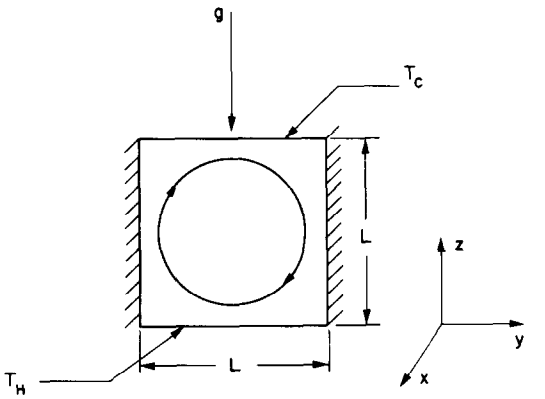

FIG. 1. Two-dimensional convection in a square box.

Donough and Catton [8] and Roberts [9] for similar problems posed on horizontally unbounded domains. We will give details of applying the method to a problem in a bounded, two-dimensional region. We will be concerned only with the supercritical, convecting case, which we treat as fully nonlinear (cf. [7] for example). Since in our treatment the Galerkin procedure is applied in only one direction, an opportunity is provided to give a very simple and complete exposition of this technique, which is not generally possible in two- and three-dimensional studies.

GOVERNING EQUATIONS

The system of equations customarily used in treating problems in thermal convection consists of the Oberbeck–Boussinesq approximation [10] to the Navier–Stokes equations, and the thermal energy equation. The steady two-dimensional form of these equations suitably scaled is as follows:

$$v_y + w_z = 0 \tag{1a}$$

$$\Delta v - p_y = \frac{1}{Pr}(vv_y + wv_z) \tag{1b}$$

$$\Delta w - p_z + RaT^* = \frac{1}{Pr}(vw_y + ww_z) \tag{1c}$$

$$\Delta T = vT_y + wT_z. \tag{1d}$$

Here the asterisk denotes a fluctuation, or perturbation, quantity, i.e. a departure from the mean temperature profile. The two dimensionless parameters which characterize the solutions to equation (1) are the Rayleigh and Prandtl numbers

$$Ra = \frac{\alpha g(T_H - T_C)L^3}{\nu\kappa} \quad \text{and} \quad Pr = \nu/\kappa.$$

The boundary conditions are

$$\begin{aligned} v(y,0) = v(y,1) = w(0,z) = w(1,z) = 0 \\ v(0,z) = v(1,z) = w(y,0) = w(y,1) = 0 \end{aligned} \tag{2a}$$

$$T(y,0) = 1,\ T(y,1) = 0,\ T_y(0,z) = T_y(1,z) = 0. \tag{2b}$$

Introducing the stream function ψ

$$v = \psi_z,\ w = -\psi_y$$

with cross-differentiation of equations (1b) and (1c), and subtracting (1b) from (1c) leads to

$$\Delta^2\psi = RaT^*_y + \frac{1}{Pr}[\psi_z\Delta\psi_y - \psi_y\Delta\psi_z] \tag{3}$$

$$\Delta T = \psi_z T_y - \psi_y T_z. \tag{4}$$

The boundary conditions employed for solving (3) are $\psi = \psi_n = 0$ on all boundaries. Those used with (4) are the temperature conditions given by (2b).

THE MIXED FINITE DIFFERENCE–GALERKIN PROCEDURE

The method to be used is a combination of finite differencing and the Galerkin procedure. To better accommodate boundary layer structure on the horizontal surfaces and take advantage of the less severe parameter variation across the region, finite differencing is done in the vertical direction and a Galerkin procedure is used for the horizontal direction. The solutions are represented by

$$\psi(y,z) = \sum_{k=1}^{K} \psi_k(z)D_k(y) \tag{5}$$

$$T(y,z) = \sum_{k=0}^{K} T_k(z)\cos a_k y, \quad a_k = k\pi \tag{6}$$

where

$$D_k(y) = \begin{cases} C_{k/2}(y) & k \text{ even} \\ S_{(k+1)/2}(y) & k \text{ odd} \end{cases}$$

with C_k and S_k being the "beam functions" developed by Harris and Reid [11]. Both sets contain odd and even subsets that will be an aid in eliminating much unnecessary computation and storage in the Galerkin procedure. The $2K + 1$ functions $\psi_k(z)$ and $T_k(z)$ are obtained from the finite difference approximation given below.

Substituting equations (5) and (6) into (3) and (4), respectively, and forming the inner product of (3) and $D_k(y)$ and (4) with $\cos a_k y$ yields for $k = 1, 2, \ldots, K$

$$\begin{aligned} \psi_k'''' + b_k^4\psi_k = -Ra\sum_j a_j A_{jk}^{(1)}T_j + \frac{1}{Pr} \\ \times\left[\sum_{j,m} B_{jmk}^{(2)}\psi_j\psi_m' - (B_{jmk}^{(2)} + B_{kmj}^{(2)})\psi_j'\psi_m \right. \\ \left. + B_{jmk}^{(1)}\psi_j'\psi_m'' + B_{mjk}^{(1)}\psi_j\psi_m'''\right]. \end{aligned} \tag{7}$$

$$T_0'' = -\sum_{j,m} a_m A_{mj}^{(1)}[T_m\psi_j' + T_m'\psi_j], \quad \text{for } k = 0 \tag{8}$$

and for $k = 1, \ldots, K$

$$\begin{aligned} T_k'' - a_k^2 T_k = -a_k T_0' \sum_j A_{kj}^{(1)}\psi_j \\ - \sum_{j,m}[a_m A_{jmk}^{(2)}T_m\psi_j' + (a_m A_{jmk}^{(2)} + a_k A_{jkm}^{(2)})T_m'\psi_j]. \end{aligned} \tag{9}$$

In equation (7) the b_k are defined as

$$b_k \equiv \begin{cases} \lambda_k & k \text{ even} \\ \mu_k & k \text{ odd} \end{cases}$$

with λ_k and μ_k as tabulated in [11]. The inner products are represented by

$$A_{jk}^{(1)} \equiv \langle \sin a_j y,\ D_k \rangle \tag{10a}$$

$$A_{jmk}^{(2)} \equiv \langle D_j \sin a_m y,\ \cos a_k y \rangle \tag{10b}$$

$$B_{jmk}^{(1)} \equiv \langle D_j D_{m,y},\ D_k \rangle \tag{10c}$$

$$B_{jmk}^{(2)} \equiv \langle D_{j,y} D_{m,yy},\ D_k \rangle \tag{10d}$$

where the subscript y denotes differentiation with respect to y.

In calculation of the inner products (10), use is made of the parities of the basis functions (even or odd) to determine which sets of indices lead to zero values.

Consider first $A^{(1)}_{jm}$. When j is even, $\sin a_j y$ is odd, and when m is even D_m is even. Therefore $A^{(1)}_{jm}$ is nonzero only when j and m are of opposite parity. As a result only half the elements of $A^{(1)}_{jm}$ need be calculated and stored.

For inner products having three indices, it is useful to construct parity tables. The table constructed for $B^{(2)}_{kjm}$ is shown in Table 1. The upper half of the table contains the various possible combinations of parity, and the lower half indicates the parity of each of the functions in the inner product. The product must be even for the integral to be nonzero. Cases where this holds are denoted by an asterisk.

From Table 1 and similar information for $A^{(2)}_{jmk}$ and $B^{(1)}_{jmk}$, it follows that all such inner products are zero, except for the following two cases: (i) $j + m$ is even with k odd; and (ii) $j + m$ is odd with k even. Thus, each of the inner products fills only one-half of a three-dimensional array. This reduction in required storage was not used in this work; however, only the nonzero inner products were calculated.

Table 1. Parity table for $B^{(2)}_{kjm}$

Index								
j	e	e	e	o	o	o	e	o
m	e	e	o	o	o	e	o	e
k	e	o	o	o	e	e	e	o
Function								
$D_{j,y}$	o	o	o	e	e	e	o	e
$D_{m,yy}$	e	e	o	o	o	e	o	e
D_k	e	o	o	o	e	e	e	o
		*		*		*	*	

e, even; o, odd.

Table 2. Dependence of Nu on y-mesh for $\Delta z = 0.01$, $K = 10$, $\varepsilon = 0.001$, $Ra = 10\,000$, $Pr = 6.7$

Δy	0.02	0.01	0.005
Nu	1.923 603	1.923 602	1.923 602

Analytical integration of the inner products is possible; but in this work, numerical quadrature by Simpson's rule was used. Different mesh sizes were tested, and the effect on Nusselt number for $K = 10$ is shown in Table 2. For larger K the sensitivity to mesh size will increase; for this study, $\Delta y = 0.01$ was sufficient.

Quasilinearization of equations (7)–(9) is accomplished by rewriting these in terms of a linear operator plus a nonlinear operator and linearizing the nonlinear operator under the assumption that it is twice Frechét differentiable [12]. When the linearized equations are differenced and the result put into matrix form, a sparse matrix containing

$$\left(N \cdot (2K+1)\right)^2$$

terms results. Unfortunately the structure is such that the sparsity is not easily exploited; and in double precision, storage requirements exceed the capacity of most modern computers. For this reason, a modal quasilinearization is used to diagonalize the matrix. (For details of this, see McDonough [13].) The storage difficulties are then completely eliminated; but the system of decoupled equations must be iterated to restore the original coupling effects. Since iterations are required, anyway, to account for nonlinearities the coupling is accomplished simultaneously with the Newton–Kantorovich iterations.

When the solution algorithm is built in this way, the overall iteration scheme is analogous to a Newton method in which only the diagonal of the Jacobian is used. Clearly, for systems having very strongly diagonally-dominant Jacobians (i.e. weak coupling), convergence rates are nearly quadratic. As the coupling becomes stronger, the convergence rate may deteriorate significantly. It will be shown for the problem considered here that convergence is slightly sublinear.

The governing equations become, after modal quasilinearization, for temperature when $k = 0$

$$T_0'' = -\sum_{j,m}^{K} a_m A^{(1)}_{mj}\left[T^{(0)}_m \psi_j'^{(0)} + T_m'^{(0)} \psi^{(0)}_j\right] \tag{11}$$

and when $k = 1, 2, \ldots, K$,

$$\begin{aligned} T_k'' + 2a_k\left(\sum_{j=1}^{K} A^{(2)}_{jkk}\psi^{(0)}_j\right)T_k' + a_k\left[\sum_{j=1}^{K} A^{(2)}_{jkk}\psi_j'^{(0)} - a_k\right]T_k \\ = -a_k T_0'^{(0)} \sum_{j=1}^{K} A^{(1)}_{jk}\psi^{(0)}_j - \sum_{\substack{j=1 \\ j\neq k}}^{K}\left[a_j A^{(2)}_{kjk} T^{(0)}_j \psi_k'^{(0)} + (a_j A^{(2)}_{kjk} + a_k A^{(2)}_{kkj}) T_j'^{(0)} \psi^{(0)}_k\right] \\ - \sum_{\substack{j,m=1 \\ j,m\neq k}}^{K}\left[a_m A^{(2)}_{jmk} T^{(0)}_m \psi^{(0)}_j + (a_m A^{(2)}_{jmk} + a_k A^{(2)}_{jkm}) T_m'^{(0)} \psi^{(0)}_j\right]. \end{aligned} \tag{12}$$

For the stream function

$$\begin{aligned} \psi_k'''' - \left(\frac{\partial N_{2,k}}{\partial \psi_k''''}\right)^{(0)}\psi_k''' - \left(\frac{\partial N_{2,k}}{\psi_k''}\right)^{(0)}\psi_k'' - \left(\frac{\partial N_{2,k}}{\partial \psi_k'}\right)^{(0)}\psi_k' + \left[b_k^4 - \left(\frac{\partial N_{2,k}}{\partial \psi_k}\right)^{(0)}\right]\psi_k \\ = -N^{(0)}_{2,k} + \left(\frac{\partial N_{2,k}}{\partial \psi_k}\right)^{(0)}\psi^{(0)}_k + \left(\frac{\partial N_{2,k}}{\partial \psi_k'}\right)^{(0)}\psi_k'^{(0)} + \left(\frac{\partial N_{2,k}}{\partial \psi_k''}\right)^{(0)}\psi_k''^{(0)} + \left(\frac{\partial N_{2,k}}{\partial \psi_k''''}\right)^{(0)}\psi_k''' \end{aligned} \tag{13}$$

for all $k = 1, 2, \ldots, K$, where

$$N_{2,k}^{(0)} = Ra \sum_{j=1}^{K} a_j A_{kj}^{(1)} T_j^{(0)} - \frac{1}{Pr}\Bigg\{ -B_{kkk}^{(2)}\psi_k^{(0)}\psi_k'^{(0)} + B_{kkk}^{(1)}(\psi_k'^{(0)}\psi_k''^{(0)} + \psi_k^{(0)}\psi_k'''^{(0)})$$

$$+ \sum_{\substack{j=1 \\ j\neq k}}^{K} \{[(B_{kjk}^{(2)} - B_{jkk}^{(2)} - B_{kkj}^{(2)})\psi_j'^{(0)} + B_{jkk}^{(1)}\psi_j'''^{(0)}]\psi_k^{(0)}$$

$$+ [(B_{jkk}^{(2)} - 2B_{kjk}^{(2)})\psi_j^{(0)} + B_{kjk}^{(1)}\psi_j'^{(0)}]\psi_k'^{(0)} + B_{jkk}^{(1)}\psi_j'^{(0)}\psi_k''^{(0)} + B_{kjk}^{(1)}\psi_j^{(0)}\psi_k'''^{(0)}\}$$

$$+ \sum_{\substack{j,m=1 \\ j,m\neq k}}^{K} [B_{jmk}^{(2)}\psi_j^{(0)}\psi_m'^{(0)} - (B_{jmk}^{(2)} + B_{kmj}^{(2)})\psi_j'^{(0)}\psi_m^{(0)} + B_{jmk}^{(1)}\psi_j'^{(0)}\psi_m''^{(0)} + B_{mjk}^{(1)}\psi_j^{(0)}\psi_m'''^{(0)}]\Bigg\}$$

$$\left(\frac{\partial N_{2,k}}{\partial \psi_k}\right)^{(0)} = -\sum_{j=1}^{K} [(B_{kjk}^{(2)} - B_{jkk}^{(2)} - B_{kkj}^{(2)})\psi_j'^{(0)} + B_{jkk}^{(1)}\psi_j'''^{(0)}]$$

$$\left(\frac{\partial N_{2,k}}{\partial \psi_k'}\right)^{(0)} = -\sum_{j=1}^{K} [(B_{jkk}^{(2)} - 2B_{kjk}^{(2)})\psi_j^{(0)} + B_{kjk}^{(1)}\psi_j''^{(0)}]$$

$$\left(\frac{\partial N_{2,k}}{\partial \psi_k''}\right)^{(0)} = -\sum_{j=1}^{K} B_{jkk}^{(2)}\psi_j'^{(0)}$$

$$\left(\frac{\partial N_{2,k}}{\partial \psi_k'''}\right)^{(0)} = -\sum_{j=1}^{K} B_{kjk}^{(1)}\psi_j^{(0)}, \quad \text{for } k = 1, 2, \ldots, K.$$

The (0) superscripts denote initially guessed values which are updated at each iteration.

The boundary value problem corresponding to equations (11)–(13) is solved by a finite difference method, as discussed in Keller [14]. All derivatives are approximated by centered differences. The differencing of (11) and (12) is trivial and will not be discussed further. It is important however to note that damping (underrelaxation) is required for the temperature equation. Thus, at each iteration, the updated value at each mesh point is given by

$$T_{k,i}^{(m+1)} = (1 - \delta)\, T_{k,i}^{(m)} + \delta T_{k,i}^{(m+1)*}$$

where m, k, and i are the iteration counter, mode number, and mesh point number. The term δ is the damping factor, and $0 < \delta \leq 1$. The asterisk refers to the most recently computed value.

Successive application of the central difference operator to equation (13) yields

$$C_1\psi_{i-2} + C_2\psi_{i-1} + C_3\psi_i + C_4\psi_{i+1} + C_5\psi_{i+2} = h^4 G_i \tag{14}$$

where the modal index has been suppressed. Here

$$C_1 = 1 + \frac{1}{2}h\left(\frac{\partial N_{2,k}}{\partial \psi'''}\right)_i^{(0)}$$

$$C_2 = -\left[4 + h\left(\frac{\partial N_{2,k}}{\partial \psi_k'''}\right)_i^{(0)} + h^2\left(\frac{\partial N_{2,k}}{\partial \psi_k''}\right)_i^{(0)} - \frac{1}{2}h^3\left(\frac{\partial N_{2,k}}{\partial \psi_k'}\right)_i^{(0)}\right]$$

$$C_3 = 6 + 2h^2\left(\frac{\partial N_{2,k}}{\partial \psi_k''}\right)_i^{(0)} + h^4\left[b_k^4 - \left(\frac{\partial N_{2,k}}{\partial \psi_k}\right)_i^{(0)}\right]$$

$$C_4 = -\left[4 - h\left(\frac{\partial N_{2,k}}{\partial \psi_k'''}\right)_i^{(0)} + h^2\left(\frac{\partial N_{2,k}}{\partial \psi_k''}\right)_i^{(0)} + \frac{1}{2}h^3\left(\frac{\partial N_{2,k}}{\partial \psi_k'}\right)_i^{(0)}\right]$$

$$C_5 = 1 - \frac{1}{2}h\left(\frac{\partial N_{2,k}}{\partial \psi'''}\right)_i^{(0)}.$$

If the set of grid points is $\{ih\}_{i=0}^{N}$, the difference equation (15) must be solved on the subset $\{ih\}_{i=1}^{N-1}$. But for $i = 1$ and $i = N - 1$, equation (14) contains grid function values at points not included in the original grid point set. These are eliminated using the derivative boundary conditions on ψ, via the centered difference approximation, e.g. at $z = 0$

$$\frac{\psi_1 - \psi_{-1}}{2h} = 0$$

or $\psi_{-1} = \psi_1$. Thus, for $i = 1$, the difference equation (14) is replaced by

$$(C_1 + C_3)\psi_1 + C_4\psi_2 + C_5\psi_3 = h^4 G_1$$

with an analogous expression holding at $i = N - 1$.

The algebraic system which results from the difference equations (14) is pentadiagonal, while tridiagonal systems arise from the difference approximations to equations (11) and (12). A general band Gaussian elimination routine was used to solve each such system of equations. This approach is very efficient, requiring only $O(N)$ arithmetic operations. (The tridiagonal case requires exactly the same number of operations as does the Thomas algorithm [16]).

The overall algorithm employed in solving equation (3) and (4) is as follows:

0. Compute required Galerkin inner products $A^{(1)}_{jm}$, $A^{(2)}_{jmk}$, $B^{(1)}_{jm}$, $B^{(2)}_{jmk}$ using Simpson's rule quadrature.
1. Assign initially-guessed values to the grid of functions $T_{0,i}$, $T_{k,i}$, and $\psi_{k,i}$, $k = 1, 2, \ldots, K$ on $\{ih\}_{i=0}^{N}$.
2. Form centered difference approximations needed to calculate $N^{(0)}_0$, $N^{(0)}_{1,k}$, and $N^{(0)}_{2,k}$.
3. Set $m = 1$, and begin Newton–Kantorovich iterations.
4. Solve difference approximation to equation (11), and update values of T'_0.
5. For $k = 1, 2, \ldots, K$:
   (a) Solve difference approximation to equation (12), and update $T'^{(0)}_k$, and
   (b) Solve difference equations (14), and update $\psi'^{(0)}_k$, $\psi''^{(0)}_k$, $\psi'''^{(0)}_k$.
6. If $m > 1$ test convergence.
7. If solution is converged, stop; otherwise set $m = m + 1$, and go to 4.

This algorithm has been coded in double precision FORTRAN; all computed results presented herein were obtained using the IBM 3033 at UCLA.

NUMERICAL ANALYTICAL TESTS

Besides convergence of the numerical approximations to the Galerkin inner products, there are three limit processes whose convergence must be demonstrated before we may accept the computed sets of numbers as solutions to the problem being considered. They are: (1) convergence of the grid functions as $h \to 0$; (2) convergence of the Newton–Kantorovich iterations as $\varepsilon \to 0$; and (3) convergence of the Fourier series, equations (5) and (6) as $K \to \infty$.

Grid function convergence

Calculations were carried out with $Ra = 20000$, $Pr = 0.71$, $\Delta y = 0.01$ and $K = 4$ with the Newton–Kantorovich convergence tolerance $\varepsilon = 0.001$. The values of ψ_k and T_k correspond roughly to their maxima in magnitude over the z-grid except for T_0 whose maximum is the boundary value. Table 3 shows the computed results and the Nusselt number.

It can be easily seen that the rate of convergence is second order and that the Nusselt number for the finest mesh is accurate to about two decimal places (the extrapolated value is 1.880 777). Calculations reported in the following paragraphs used $h = 0.01$ even though $h = 0.0125$ is sufficient. The dominant error can be seen to be from truncation of the Fourier series.

Newton–Kantorovich iteration convergence

A precise theoretical rate of convergence cannot be easily given because of our use of modal decomposition. Further, it follows from the Newton–Kantorovich theorem [12] that convergence may be strongly influenced by Ra, Pr, δ, initial guesses, and the norm by which convergence is measured. Results will be given to show the influence of each of these parameters.

Since an exact solution is not known, the error is not known. For this work we use

$$e_m = \max_k \max_{z_i} \left(\left|\psi_k^{(m+1)} - \psi_k^{(m)}\right|, \left|T_k^{(m+1)} - T_k^{(m)}\right|\right) \quad (15)$$

where $z_i \in \{ih\}_{i=1}^{N-1}$. Figure 2 is a log–log plot of e_{m+1} vs e_m for $m = 1$–30. For this run, $Ra = 20000$, $Pr = 6.7$, $K = 4$, $\Delta y = 0.01$, $\delta = 0.5$ and $h = 0.01$. Points lie both above and below the line $e_{m+1} = e_m$, indicating that convergence is not monotone. The dashed line corresponds to

$$e_{m+1} = 0.489 e_m^{0.968}.$$

Since the exponent is less than one, convergence is slightly sublinear but the error is approximately halved at each iteration.

To measure convergence rate as a function of δ, the error tolerance, ε, was fixed at 10^{-3} and the number of iterations necessary for each value of δ was obtained. Results are shown in Fig. 3 for $Ra = 10000$ and 20000. It can be seen that the choice of δ is important; the number of iterations is very sensitive to δ when $\delta > \delta_{opt}$.

With $\delta = 0.4$, $K = 4$, $\Delta y = 0.01$ and $h = 0.01$ calculations were made at several values of Pr and Ra. The number of iterations needed for convergence is

Table 3. Grid function convergence

Grid function	$h = 0.1$	$h = 0.05$	$h = 0.025$	$h = 0.0125$
$T_0(0.3)$	5.542195-1	5.615725-1	5.636105-1	5.640501-1
$T_1(0.5)$	−1.350705-1	−1.304509-1	−1.292512-1	−1.289492-1
$T_2(0.2)$	4.067096-2	3.582110-2	3.458850-2	3.434991-2
$T_3(0.5)$	6.835981-3	6.943724-3	6.973215-3	6.980665-3
$T_4(0.3)$	−3.406716-3	−3.127656-3	−3.063978-3	−3.048376-3
$\psi_1(0.3)$	6.050797-1	5.249729-1	5.037742-1	5.000764-1
$\psi_2(0.5)$	3.097098-0	3.007878-0	2.982816-0	2.976444-0
$\psi_3(0.4)$	−3.690025-2	−3.189531-2	−3.060985-2	−3.033642-2
$\psi_4(0.2)$	1.368411-1	1.057800-1	9.837934-2	9.679529-2
Nu	2.104220	1.920818	1.888373	1.882676

Note: in this table numbers are represented in exponential form, but without the base; i.e. 5.542-1 ≡ 5.542×10^{-1}.

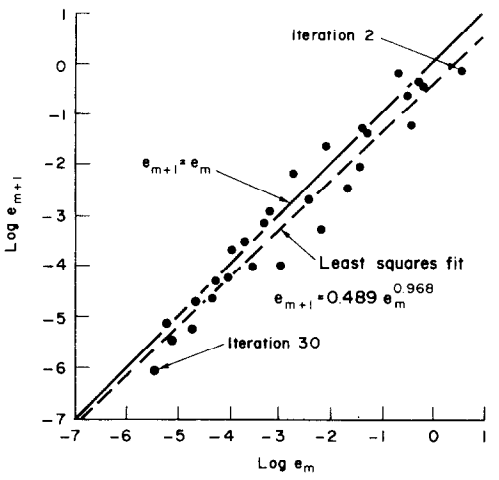

FIG. 2. Error at $m + 1$th iteration vs error at mth iteration.

shown in Table 4. It can be seen that the number of iterations is not very sensitive to either Ra or Pr, at least at moderate values of Ra, for $\delta < \delta_{opt}$.

The initial guess used in the calculations presented above was

$$\psi_k(z) \equiv 0 \tag{16a}$$

$$T_k(z) = \frac{1}{a_k}\sin(a_k z), \quad k = 1, 2, \ldots, K \tag{16b}$$

and

$$T_0(z) = 1 - z - \frac{\sin 2\pi z}{100}. \tag{16c}$$

Figure 4 provides a comparison of the initial guesses to T_k with the solution obtained at $Ra = 20000$, $Pr = 6.7$, and $K = 4$ after 25 iterations. It is clear that this guess was a poor one.

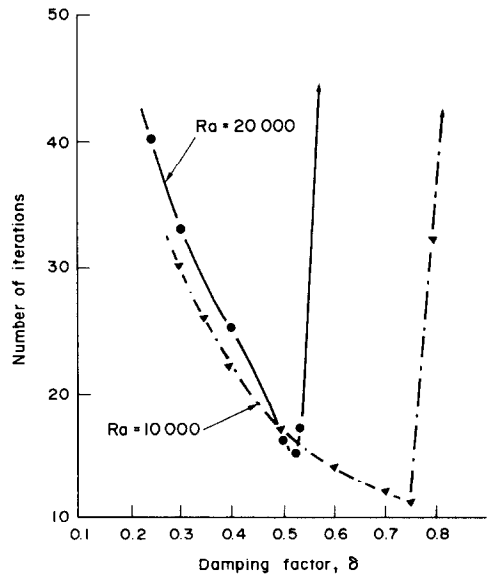

FIG. 3. Number of iterations vs damping factor.

Table 4. Effect of Pr on required number of iterations

	Number of iterations	
Pr	$Ra = 10000$	$Ra = 20000$
0.71	23	22
6.7	22	25
70.0	25	21
700.0	25	23

To demonstrate the impact of the initial guess, calculations were made for this same case using equation (16) to obtain a solution at $Ra = 10000$, and at 15000. These solutions were then used as the initial guess for a solution at $Ra = 20000$. Table 5 summarizes these results. It is clear that improved initial guesses accelerate convergence. Moreover, for $Ra \geq 40000$ it is not possible to obtain convergence using equation (16), and "continuation" is required.

Changing the norm by which convergence is measured will also alter convergence rate. The convergence criterion given by equation (15) requires that solutions to (3) and (4) be uniformly continuous on $[0, 1] \times [0, 1]$. A frequently used weaker form of convergence is convergence in Nusselt number (it can be shown that $(Nu - 1)^{1/2}$ provides a measure of convergence similar to the L^2-norm, which is natural to a Galerkin procedure [13]). A comparison of the number of iterations required to satisfy the same tolerance, $\varepsilon = 0.0001$, for the norm given by equation (15) and the convergence measure

$$\bar{e}_{m+1} = |Nu^{(m+1)} - Nu^{(m)}| \tag{17}$$

is given in Table 6 for cases with $Pr = 0.71$.

As Ra increases, the number of iterations for the pointwise norm becomes significantly greater than for the L^2-norm. This is not unexpected since L^2-solutions need not be continuous, while pointwise convergence implies uniformly continuous solutions. Higher Ra leads to a more complicated flow structure. At higher Pr, the Ra at which the two norms differ is higher. For example at $Pr = 6.7$, and $Ra = 35000$, 39 and 40 iterations were required for L^2 and pointwise convergence, respectively.

Fourier series convergence

To provide numerical evidence of existence (and uniqueness) of solutions to equations (3) and (4) it must be shown that convergence of the Fourier representations is absolute and uniform. Since the basis functions in (5) and (6) are uniformly bounded, we define the pointwise norm,

$$\|f\|_z = \max|f(y, z)|$$

and it follows that

$$\|\psi\|_z \leq C_1 \sum_{k=1}^{z} \|\psi_k\|_z \tag{18a}$$

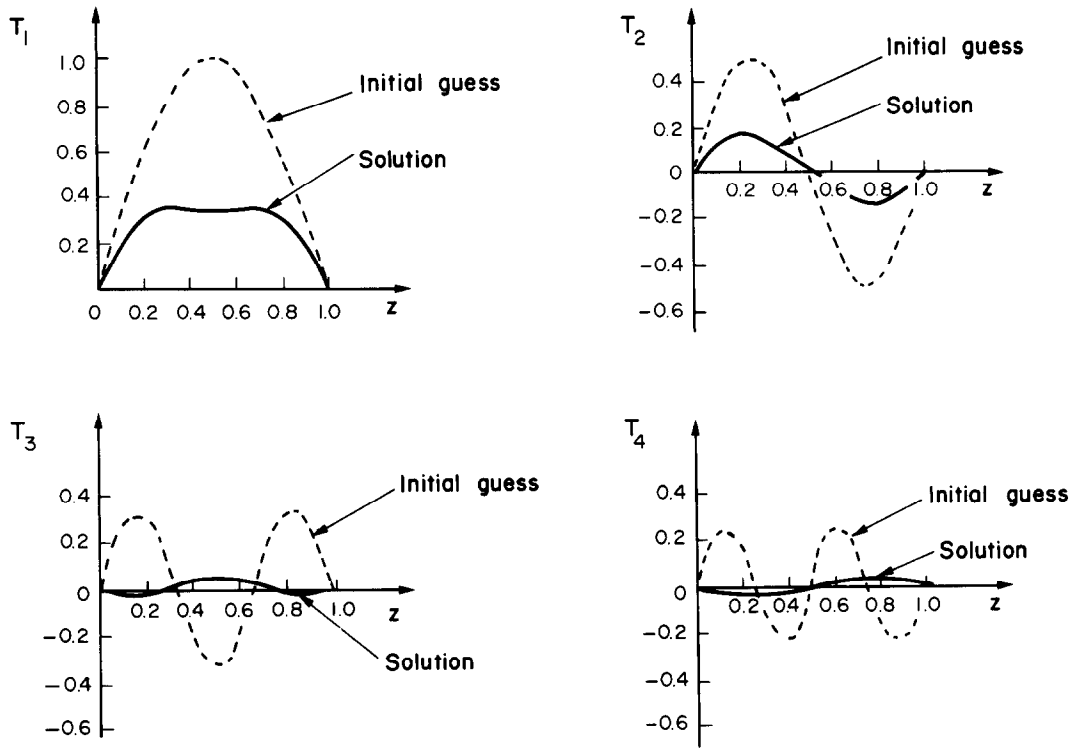

FIG. 4. Comparison of initial guess and solution for $T_k(z)$, $k = 1, 2, 3, 4$. Note: all data scaled by 0.31831, maximum value of T_1 initial guess.

$$\| T \|_\infty \leq C_2 \sum_{k=0}^{\infty} \| T_k \|_\infty \tag{18b}$$

where C_1 and C_2 are normalization constants.

If the series on the right-hand side of equation (18) converge, they do so absolutely and uniformly, by definition of $\| \cdot \|_\infty$. Thus, convergence is demonstrated by showing that $\| \psi \|_\infty$ and $\| T \|_\infty$ are of the order k^{-p} with $p > 1$. Calculations were carried out for K up to 10 in the Fourier representations. The results are shown in Fig. 5 for $Ra = 40000$ and $Pr = 0.71$. The series clearly appear to converge.

The points plotted in Fig. 5 are obtained by scaling the max-norm of each mode with the max-norm of the mode beyond which convergence is monotone. For the ψ_k, this was mode #1; and for the T_k, it was mode #5. The values of k in the figure are referenced to the mode used for scaling. We have plotted only results for odd modes because these converge more slowly. The ψ_k converge more rapidly than do the T_k; $\| \psi_4 \|_\infty \sim O(10^{-6})$ cannot be shown in the scale of the figure. Convergence of $\| T_k \|_\infty$ goes somewhat faster than $1/k$, so we expect that the series for temperature is absolutely convergent.

The absolute convergence of the series representations (5) and (6) shows that the rearrangements needed to obtain equations (7)–(9) are justified. Hence, the right-hand sides of these equations are well defined. This is necessary, but not sufficient, for uniqueness of solutions for (7)–(9), and thus also for (3) and (4). Rigorous sufficient conditions are somewhat difficult to obtain, but in fact have been given for the infinite domain problem for low values of Ra by Rabinowitz [17]. The basic approach was to show that the nonlinear integral equations corresponding to equations (7) and (9) provide a contractive mapping of a certain Sobolev space into itself. Another possible approach is to employ extensions of the maximum

Table 5. Effect of initial guess for solutions at $Ra = 20000$, $Pr = 6.7$

Initial guess	Number of iterations
Equation (16)	25
Solution at $Ra = 10000$	20
Solution at $Ra = 15000$	13

Table 6. Number of iterations needed for pointwise- and L^2-convergence, $Pr = 0.71$

	Number of iterations	
Ra	Pointwise	L^2
5000	27	26
10000	25	24
15000	31	29
20000	31	29
25000	31	31
30000	34	28
35000	42	37

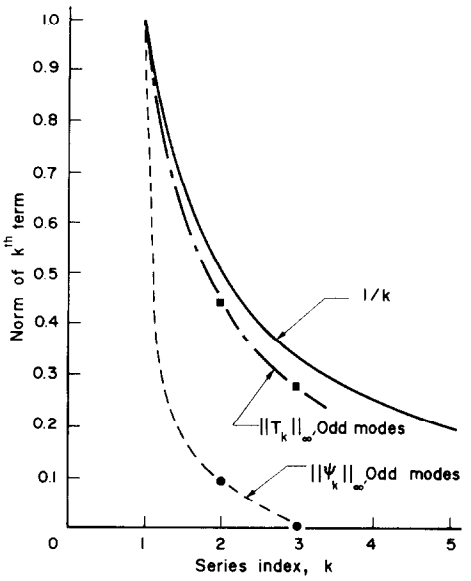

FIG. 5. Comparison of $\|T_k\|_\infty$, $\|\psi_k\|_\infty$ with $1/k$.

principle theories given in Protter and Weinberger [18]. We obtain numerical evidence of at least local uniqueness by computing the same solution with several different initial guesses.

If we accept the conclusions reached from Fig. 5, we can consider the error incurred by truncation of the series after K terms. Table 7 shows Nu as a function of K for several values of Pr at $Ra = 40\,000$ ($h = 0.01$, $\Delta y = 0.01$, $\varepsilon = 0.001$ and $\delta = 0.4$) with pointwise convergence required. Convergence is clearly demonstrated. Theory presented by Gottlieb and Orszag [19] predicts that the convergence should be infinite-order, and the table gives some indication of this.

Two interesting aspects of the results are that lower Pr does not converge as fast as high Pr, and convergence is not monotone. The behavior with Pr is not unexpected since the nonlinearity increases with decreasing Pr, as does the potential for turbulence. The lack of monotone convergence was unexpected, but also found by Catton *et al.* [6], and Denny and Clever [7]. It is worth noting that for $Pr = 6.7$, the value of Nu for $K = 2$ is within a fraction of a per cent of the converged value. Hence, if details of the flow and temperature field are not required, a two term approximation is sufficient. Further, for $Ra = 40\,000$ a four term approximation yields results well within experimental accuracy.

Table 7. Nu dependence on number of terms in Galerkin approximation, $Ra = 40\,000$

	Nu				
Pr	$K = 2$	$K = 4$	$K = 6$	$K = 8$	$K = 10$
0.71	1.8236	2.1364	2.1423	2.1471	2.1514
6.7	2.8387	2.8236	2.8356	2.8393	2.8419
70.0	2.9986	2.9975	2.9887	2.9926	2.9924
700.0	3.0128	2.9900	3.0028	3.0064	3.0062

COMPARISON WITH OTHER METHODS

A comparison between the mixed finite difference–Galerkin, Galerkin, and finite difference methods is made for computer storage requirements, arithmetic operation count and program set-up complexity. The comparisons are in many respects qualitative because only order of magnitude estimates are available in the first two categories, and the last is mostly subjective.

Storage requirements

For a pure Galerkin solution it is not uncommon to use on the order of 10^2 trial functions. This leads to about 10^6 words of storage for the inner products alone. For finite difference methods, in particular for methods using ADI, storage of the grid functions themselves usually determines the overall storage requirements. The matrices corresponding to the linear systems to be solved are band matrices. Hence, even for a $10^2 \times 10^2$ grid only about 10^4 words of storage are required. Storage requirements for the mixed method are 10^5–10^6 words if modal decomposition is not used. For the method used here, the main storage requirement comes from the combination of grid function solutions to the Galerkin ODEs and the Galerkin inner products. The first of these is $O(N \cdot K)$ for an N-point finite difference grid and a K-term Galerkin approximation, while the second is $O(K^3)$. For $K \lesssim 50$ and $N \lesssim 100$, storage requirements are of the same order as for a full finite difference grid.

Arithmetic operations

For the Galerkin procedure, the arithmetic comes from a calculation of inner products, construction of matrix elements and solution of the system of equations. The first of the operations is done only once, but requires at least $O(K^3)$ arithmetic operations. Set up of the matrices usually requires $O(K^2)$ arithmetic operations. Solution of the system of equations is carried out iteratively, and involves $O(K^3)$ operations per iteration. If good initial guesses are used, only a few iterations are needed in a Newton's method solution algorithm.

A finite difference method on an $M \times N$ grid using an ADI scheme and solving quasilinearized equations (equivalent to Newton's method) will require the solution of N linear systems having $O(M)$ elements, and M linear systems having $O(N)$ elements. Thus, $O(M \cdot N)$ arithmetic operations are needed to solve the linear systems, provided a direct band-elimination method is employed. In addition, $O(M \cdot N)$ operations are needed to update the linear systems at each iteration. Hence, the total operation count for a solution is $O(M \cdot N)$ if Newton's method converges rapidly.

A mixed finite difference–Galerkin procedure requires that $O(K)$ linear systems be solved on an N-point finite difference grid at each iteration. The system of equations is sparse and requires $O(N)$ arithmetic operations for solution, and $O(N \cdot K^2)$ arithmetic

Table 8. Summary of theoretical comparisons

Method	Storage	Arithmetic operations	Set-up time
Galerkin	$\sim O[(KL)^3]$	$\sim O[(KL^3)]$	8 days*
Finite difference (ADI)	$\sim O(MN)$	$\sim O(MN)$	5 days
Mixed finite difference–Galerkin	$\sim O(NK + K^3)$	$\sim O(NK + NK^3)$	3 days*

* Quadrature used to calculate inner products.

operations for set up for each K. In addition, $O(K^3)$ operations are needed to compute the inner products. Therefore, $O(NK + NK^3 + K^3)$ arithmetic operations are needed to obtain a solution.

Program complexity

To compare the three methods, all the equations needed to implement a computer code were derived, and put into a form ready for coding. Only the mixed finite difference–Galerkin procedure was coded and executed. The other two codes will be produced later for more detailed comparisons. Here we only point out the difficulties one might encounter with each of the methods.

The two major steps in the Galerkin method are calculation of inner products and solution of linear systems. The inner products are multiple integrals and only calculable analytically for very simple basis sets. The parity tables used here are very difficult to construct if more than one dimension must be considered. The matrices representing the linear systems are often non-sparse and frequently mildly ill-conditioned. Iterative methods cannot be applied, and roundoff error can be difficult to control. Pivoting strategies or iterative improvement should be employed. Preparation of the equations for coding required eight days.

The major difficulties in implementing a finite difference scheme are linearization of the nonlinear equations and treatment of the boundary conditions. Quasilinearization of partial differential equations requires very tedious algebraic manipulations. To set up the equations for an ADI finite difference scheme took five days.

The problems encountered in the mixed method include those found in both of the other methods. They are, however, easier to deal with because the space dimensions have been reduced. Determining null inner products from parity tables is easily accomplished. Quasilinearization for Galerkin ODEs is simpler than for partial differential equations. The form of the Galerkin ODEs is similar to the ADI equations resulting from the finite difference method, and the boundary conditions present similar difficulties. The mixed method required three days to prepare for coding.

Table 8 summarizes the findings of this section. It is not possible to give precise performance comparisons at this time. We will, however, use the results of [7] to infer that $KL \approx 60$ and $M = N = 50$ to achieve the three place accuracy in Nu obtained here for high Pr with $N \simeq 100$ and $K = 8$. For the Galerkin method, approximately 2×10^5 words of storage and arithmetic operations per iteration are needed. The ADI finite difference scheme requires only about 10^4 words of storage and 5×10^3 arithmetic operations. The mixed method requires about 8×10^3 words of storage and 10^4 arithmetic operations per iteration.

Table 9. Time per iteration vs number of modes for $Ra = 20\,000$, $Pr = 0.71$, $\Delta y = 0.01$ and $\varepsilon = 0.001$

Number of modes	2	4	6	8	10
Time (s)	0.059	0.171	0.380	0.756	1.332

In terms of storage requirements, the mixed method and the ADI finite-difference method compare favorably and are superior to the Galerkin procedure. The number of arithmetic operations for the mixed method is greater than for a finite difference scheme by a factor of two or so; both are considerably less than for a full Galerkin procedure. Thus, the performance of the mixed method appears to fall between that of the ADI finite difference scheme and a full Galerkin procedure. The mixed method is, however, the easiest of the three to implement. Further, acceptance of slightly less accuracy, $K = 2$ instead of 8, shifts the balance in favour of the mixed method.

Finally, in Table 9, some timing data are presented for direct comparison with other methods. An IBM Extended-H FORTRAN compiler in optimizing mode was used to generate the object code for these runs. The times listed in the table are average central processing seconds per iteration for runs on an IBM 3033, obtained by dividing total execution time by the number of iterations. With h (i.e. N) fixed the increase in the number of operations should increase like K^3, see Table 8. The data of Table 9 seem to indicate an $O(K^2)$ variation; but experience in the unbounded domain problem [13] shows that this holds only when K is small enough that the set up operations are still a significant fraction of the total operation count.

Acknowledgement—Financial support for this research was provided by the National Science Foundation, grant ENG 78-25273.

REFERENCES

1. I. Catton, Convection in a closed rectangular region: the onset of motion, *J. Heat Transfer* **92,** 186–188 (1970).
2. I. Catton and D. K. Edwards, Effect of side walls on natural convection between horizontal plates heated from below, *J. Heat Transfer* **89,** 295–299 (1967).
3. J. N. Arnold, I. Catton and D. K. Edwards, Experimental investigation of natural convection in inclined rectangular regions of differing aspect ratios. *J. Heat Transfer* **98,** 67–71 (1976).
4. G. de Vahl Davis, Laminar natural convection in an enclosed rectangular cavity, *Int. J. Heat Mass Transfer* **11,** 1–17 (1968).
5. J. O. Wilkes and S. W. Churchill, The finite-difference computation of natural convection in a rectangular enclosure, *AIChE Jl* **12,** 161–166 (1966).
6. I. Catton, P. S. Ayyaswamy and R. M. Clever, Natural convection flow in a finite rectangular slot arbitrarily oriented with respect to the gravity vector, *Int. J. Heat Mass Transfer* **17,** 173–183 (1974).
7. V. E. Denny and R. M. Clever, Comparisons of Galerkin and finite difference methods for solving highly nonlinear thermally driven flows, *J. Comp. Physics* **16,** 271–284 (1974).
8. J. M. McDonough and I. Catton, Wavenumber selection via thermodynamic stability for two-dimensional Bénard convection, ASME Paper 79-WA/HT-14, 1–7 (1979).
9. P. H. Roberts, Convection in horizontal layers with internal heat generation: theory, *J. Fluid Mech.* **30,** 33–49 (1967).
10. D. D. Joseph, *Stability of Fluid Motions II.* Springer, New York (1976).
11. D. L. Harris and W. H. Reid, On orthogonal functions which satisfy four boundary conditions: I. Tables for use in Fourier-type expansions, *Astrophys. J. Supp.* **3,** 429–452 (1958).
12. S. M. Roberts and J. S. Shipman, *Two-Point Boundary Value Problems: Shooting Methods.* Elsevier, New York (1972).
13. J. M. McDonough, The Rayleigh–Bénard problem on a horizontally unbounded domain: determination of the wavenumber of convection, Ph.D. dissertation, School of Engineering and Applied Science, University of California, Los Angeles (1980).
14. H. B. Keller, *Numerical Methods For Two-Point Boundary Value Problems.* Blaisdell, Waltham (1968).
15. J. J. Dongarra, C. B. Moler, J. R. Bunch and G. W. Stewart, *Linpack Users Guide.* SIAM, Philadelphia (1979).
16. W. F. Ames, *Numerical Methods for Partial Differential Equations,* 2nd edn. Academic Press, New York (1977).
17. P. H. Rabinowitz, Existence and non-uniqueness of rectangular solutions to the Bénard problem, *Arch. Rational Mech. Anal.* **29,** 32–49 (1968).
18. M. H. Protter and H. F. Weinberger, *Maximum Principles in Differential Equations.* Prentice-Hall, Englewood Cliffs (1967).
19. D. Gottlieb and S. A. Orszag, Numerical analysis of spectral methods: theory and applications, CBMS-NSF Regional Conference in Applied Mathematics no. 26, SIAM, Philadelphia (1977).

UNE PROCEDURE MIXTE GALERKIN–DIFFERENCE FINIE POUR LA CONVECTION BIDIMENSIONNELLE DANS UNE CAVITE CARREE

Résumé—Une méthode mixte Galerkin–différence finie est utilisée pour résoudre le problème de la convection thermique dans une cavité carrée, horizontale, bidimensionnelle et chauffée par le bas. La procédure Galerkin est appliquée dans la direction horizontale, les différences finies dans la direction verticale. Les résultats numériques de cette approche sont comparés théoriquement à ceux des méthodes usuelles des différences finies et de Galerkin. Des données spécifiques numériques et analytiques sont données pour cette procédure mixte dans le cas de plusieurs valeurs de nombre de Prandtl et pour un domaine de nombre de Rayleigh.

EIN GEMISCHTES DIFFERENZEN- UND GALERKIN-VERFAHREN FÜR ZWEIDIMENSIONALE KONVEKTION IN EINEM QUADRATISCHEN BEHÄLTER

Zusammenfassung—Mit einem gemischten Differenzen- und Galerkin-Verfahren wird der Fall der thermischen Konvektion in einem zweidimensionalen quadratischen Behälter, der von unten beheizt wird, berechnet. Das Galerkin-Verfahren wird in horizontaler Richtung angewandt, das Differenzenverfahren in vertikaler Richtung. Die numerischen Grundzüge eines solchen Ansatzes werden theoretisch mit denen üblicher Differenzen- bzw. Galerkin-Verfahren verglichen. Spezielle numerisch-analytische Rechenergebnisse aus dem gemischten Differenzen- und Galerkin-Verfahren werden für einige Prandtl-Zahlen und über einen gewissen Rayleigh-Zahlenbereich vergelegt.

ИСПОЛЬЗОВАНИЕ МЕТОДА КОНЕЧНЫХ РАЗНОСТЕЙ СОВМЕСТНО С МЕТОДОМ ГАЛЕРКИНА ДЛЯ РЕШЕНИЯ ЗАДАЧИ ДВУМЕРНОЙ КОНВЕКЦИИ В КВАДРАТНОЙ ОБЛАСТИ

Аннотация — Метод конечных разностей в сочетании с методом Галеркина используется для решения задачи тепловой конвекции в двумерной горизонтальной нагреваемой снизу квадратной области. Метод Галеркина применяется для горизонтального направления, а метод конечных разностей — для вертикального. Проведено численное сравнение такого подхода с каждым из указанных методов в отдельности. Представлены результаты решения для нескольких значений числа Прандтля и большого диапазона чисел Релея.