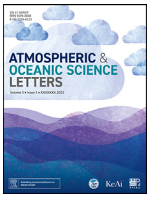
Contents lists available at ScienceDirect

Atmospheric and Oceanic Science Letters

journal homepage: http://www.keaipublishing.com/en/journals/atmospheric-and-oceanic-science-letters/



Computational uncertainty and optimal grid size and time step of the Lax–Friedrichs scheme for the 1D advection equation

Jing Cao[a], Jianping Li[b,c,*], Yanjie Li[d]

[a] College of Science, Tianjin University of Technology, Tianjin, China
[b] Frontiers Science Center for Deep Ocean Multispheres and Earth System/Key Laboratory of Physical Oceanography/Academy of the Future Ocean/ Innovation Center for Ocean Carbon Neutrality, Ocean University of China, Qingdao, China
[c] Laboratory for Ocean Dynamics and Climate, Pilot Qingdao National Laboratory for Marine Science and Technology, Qingdao, China
[d] State Key Laboratory of Numerical Modeling for Atmospheric Sciences and Geophysical Fluid Dynamics, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Keywords:
Lax–Friedrichs scheme
Computational uncertainty principle
Optimal grid size
Optimal time step
Universal relation
关键词:
Lax-Friedrichs 格式
数值计算的不确定性原理
最优格距
最优时间步长
普适关系

ABSTRACT

This paper examines truncation and round-off errors in the numerical solution of the 1D advection equation with the Lax–Friedrichs scheme, and accumulation of the errors as they are propagated to high temporal layers. The authors obtain a new theoretical approximation formula for the upper bound of the total error of the numerical solution, as well as theoretical formulae for the optimal grid size and time step. The reliability of the obtained formulae is demonstrated with numerical experimental examples. Next, the ratio of the optimal time steps under two different machine precisions is found to satisfy a universal relation that depends only on the machine precision involved. Finally, theoretical verification suggests that this problem satisfies the computational uncertainty principle when the grid ratio is fixed, demonstrating the inevitable existence of an optimal time step size under a finite machine precision.

摘要

本文对于应用Lax- Friedrichs 格式数值求解一维平流方程, 研究数值求解过程中产生的截断误差与舍入误差, 以及两种误差逐层向高时间层传播的累积, 得到新的数值解总误差上界的理论近似公式, 以及最优格距和最优时间步长的理论公式. 通过数值算例验证了所得公式的可靠性. 然后, 发现了两种不同机器精度下最优时间步长之比满足的一个仅与机器精度有关的普适关系. 最后, 理论验证了在网格比固定的情况下, 此问题满足数值计算的不确定性原理, 以及在机器有限精度下最优时间步长的必然存在.

1. Introduction

In April 2022, the State Department promulgated the Outline of High-Quality Meteorological Development (2022–2035), which proposed the establishment of an accurate weather forecasting system featuring five "ones": early warning of strong local weather one hour in advance; forecasting of hourly weather one day in advance; forecasting of disastrous weather one week in advance; forecasting of major weather processes one month in advance; and forecasting of global climate anomalies one year in advance. This proposal signifies China's need for numerical forecasting systems with high spatiotemporal resolution. The determination of temporal and spatial resolutions is an extremely important and critical component of studies on numerical weather and climate forecasting technologies. Although the use of a high-resolution model can improve the forecasting performance (van Roosmalen et al., 2010; Zhang et al., 2010; Kendon et al., 2012), applying such a model operationally is usually costly due to the high computational requirements, electrical power demand, data storage, etc. More importantly, studies have shown that not all simulations of high-resolution versions of all models are superior to those of low-resolution versions (Deng et al., 2012; Bacmeister et al., 2014; Falasca and Curci, 2018; Maurya et al., 2018; Shi et al., 2021). Yu et al. (2018) compared the effects of different resolutions of the mesoscale version of the global/regional assimilation and prediction system (GRAPES-MESO) model on summer precipitation forecasting in China. It was found that the refined spatial resolution simulation failed to significantly improve the precipitation location forecast, despite a better improvement in the precipitation maximum forecast. Liu et al. (2015) studied the effect of the spatiotemporal resolution of this model on the prediction skill and found that it is not in direct proportion to the spatial resolution. The question arises, therefore, as to whether there might be a way to determine the model resolution reasonably, efficiently, and inexpensively. One of the most fundamental and important steps in solving this problem is to select the optimal spatiotemporal resolution, i.e., grid size (spatial resolution) and time step

[*] Corresponding author.
E-mail address: ljp@ouc.edu.cn (J. Li).

https://doi.org/10.1016/j.aosl.2023.100331
Received 16 December 2022; Revised 8 January 2023; Accepted 28 January 2023
Available online 1 February 2023

(temporal resolution), in the numerical calculation of the partial differential equations (PDEs) to obtain the optimal numerical solution.

In previous theoretical studies on the optimal grid size and time step for the numerical solution of PDEs, truncation error was often the principal consideration, while round-off error received less attention (Faragó and Horváth, 1999; Guo, 2021). However, an originally small round-off error may greatly affect the accuracy of the numerical solution after a long numerical calculation time. For numerical solutions of nonlinear ordinary differential equations (ODEs), Li et al. (2000) proposed the computational uncertainty principle (CUP) based on an argument that the presence of round-off errors may bring about uncertainties in numerical computation, and then the existence of the optimal time step (Li et al., 2001) and effective time step interval (Cao et al., 2017) was theoretically demonstrated. Such uncertainty in numerical computation has also been brought to light in numerical solutions of PDEs. Wang et al. (2007) and Wang and Zhang (2008) studied the effect of computational uncertainty on the climate simulations of atmospheric circulation models and found that round-off errors were an important source of random error in the models. For GRAPES-MESO, Liu et al. (2015) verified the applicability of the theory of an optimal time step in the CUP for complex PDEs through numerical experiments. In a study on the heat diffusion equation for deep soil in a land surface model, Harvey and Verseghy (2016) reached a conclusion consistent with the CUP; that is, because of the effect of round-off errors, the model accuracy deteriorated when using smaller time steps. Li and Wang (2008) stated that the CUP is necessarily satisfied in numerical solutions of nonlinear systems, and studying round-off errors and the optimal grid size and time step in the numerical solutions of PDEs is an important issue.

As mentioned above, the CUP has been demonstrated by numerical experiments in the numerical solutions of PDEs, but relevant theoretical studies are still rare. This issue is studied in the present paper with the Lax–Friedrichs scheme of 1D advection equations. A theoretical analysis of round-off errors and their propagation laws during the numerical solution is conducted, and the CUP is subjected to preliminary theoretical verification in the PDEs. Then, theoretical formulae for the optimal grid size and time step are presented, providing theoretical support for further identification of the optimal resolutions in more complex numerical schemes and atmospheric numerical models.

2. Problem description

The initial boundary value problem of the 1D advection equation is investigated:

$$\begin{cases} u_t + a u_x = f(x,t),\ a > 0,\ 0 < x < l,\ t > 0, \\ u(0,t) = \omega_1(t),\ t > 0, \\ u(x,0) = \varphi(x),\ 0 \le x \le l, \end{cases} \tag{1}$$

where a is parameter of the advection equation, ω_1 and φ are boundary and initial values, respectively. The solution area is defined as $\{(x,t)|\ x \in [0,l],\ t \in [0,\rho]\}$, where l and ρ are the upper bounds of x and t, respectively. We divided the solution area into a $J \times N$ rectangular grid. The grid size is h and the time step is τ, while $\lambda = \frac{\tau}{h}$ denotes the time–space grid ratio. Each grid node is expressed as (x_j, t_n), where $x_j = jh,\ j = 0,\ 1,\ \cdots,\ J$ and $t_n = n\tau,\ n = 0,\ 1,\ \cdots,\ N$, and the exact solution to Eq. (1) at node (x_j, t_n) is prescribed as $u(x_j, t_n)$. The equation is discretized using the Lax–Friedrichs scheme, while the upwind scheme is used at the right-end boundary. We denote $u_{j,n}$ as the approximate solution of Eq. (1) after discretization, and $f_{j,n}$ as the value of $f(x_j, t_n)$. The computational scheme is as follows:

$$\begin{cases} u_{j,n} = \frac{1}{2}(1 + a\lambda)u_{j-1,n-1} + \frac{1}{2}(1 - a\lambda)u_{j+1,n-1} + \tau f_{j,n-1},\ j = 1,\ 2,\ \cdots,\ J-1, \\ n = 1,\ 2,\ \cdots,\ N, \\ u_{J,n} = (1 - a\lambda)u_{J,n-1} + a\lambda u_{J-1,n-1} + \tau f_{J,n-1},\ n = 1,\ 2,\ \cdots,\ N, \\ u_{0,n} = \omega_{1,n} = \omega_1\left(t_n\right),\ n = 1,\ 2,\ \cdots,\ N, \\ u_{j,0} = \varphi_j = \varphi\left(x_j\right),\ j = 0,\ 1,\ \cdots,\ J. \end{cases} \tag{2}$$

The left-end boundary condition is assumed to be accurate. The truncation error generated from the discretization in Eq. (2) is expressed as $T_{j,n} = u_{j,n} - u(x_j, t_n)$, the vector is assumed to be $\bar{\boldsymbol{T}}_n = (T_{1,n}, T_{2,n},\ \cdots, T_{J,n})^{\mathrm{T}}$, and the truncation error at the nth temporal layer is measured as $\|\bar{\boldsymbol{T}}_n\|_\infty$. We use the widely used maximum principle analysis (William, 1977) to estimate $\|\bar{\boldsymbol{T}}_N\|_\infty$, denoted as $\|T(h,\tau)\|_\infty$, i.e., the truncation error at the top temporal layer. The numerical calculation of Eq. (2) by a computer will also produce round-off errors; the new round-off error generated at the nth temporal layer is denoted as $R_{j,n}$, and it is assumed that $\bar{\boldsymbol{R}}_n = (R_{1,n}, R_{2,n}, \cdots, R_{J,n})^{\mathrm{T}}$ where $\|\bar{\boldsymbol{R}}_n\|_\infty$ is used to measure the newly generated round-off error at the nth temporal layer. We calculate the sum of the newly generated round-off errors propagated from each temporal layer to the top layer, namely $\sum_{n=0}^{N} \|\bar{\boldsymbol{R}}_n\|_\infty$ as the round-off errors at the top temporal layer prescribed as $\|R(h,\tau)\|_\infty$. By denoting the sum of the truncation error and round-off error of the top temporal layer as $\|E(h,\tau)\|_\infty$, we obtain

$$\|E(h,\tau)\|_\infty = \|T(h,\tau)\|_\infty + \|R(h,\tau)\|_\infty. \tag{3}$$

This paper will theoretically estimate the upper bound $\sup\|E(h,\tau)\|_\infty$ of $\|E(h,\tau)\|_\infty$ and present a method for determining the optimal grid size H and optimal time step Γ.

3. Theoretical analysis of the total error

3.1. Propagation of round-off error

It is assumed that only the initial temporal layer produces round-off errors while no other layers produce such errors; the numerical solution for this special case is denoted by $\dot{u}_{j,n}$, which should satisfy

$$\begin{cases} \dot{u}_{j,n} = \frac{1}{2}(1 + a\lambda)\dot{u}_{j-1,n-1} + \frac{1}{2}(1 - a\lambda)\dot{u}_{j+1,n-1} + \tau f_{j,n-1},\ j = 1,\ 2,\ \cdots, \\ J-1,\ n = 1,\ 2,\ \cdots,\ N, \\ \dot{u}_{J,n} = (1 - a\lambda)\dot{u}_{J,n-1} + a\lambda\dot{u}_{J-1,n-1} + \tau f_{J,n-1},\ n = 1,\ 2,\ \cdots,\ N, \\ \dot{u}_{0,n} = \omega_{1,n},\ n = 1,\ 2,\ \cdots,\ N, \\ \dot{u}_{j,0} = \varphi_j + R_{j,0},\ j = 0,\ 1,\ \cdots,\ J. \end{cases} \tag{4}$$

The error propagated from the initial temporal layer to the nth temporal layer in this particular case is denoted by $d_{j,n} = \dot{u}_{j,n} - u_{j,n}$. From Eqs. (2) and (4), $d_{j,n}$ should satisfy

$$\begin{cases} d_{j,n} = \frac{1}{2}(1 + a\lambda)d_{j-1,n-1} + \frac{1}{2}(1 - a\lambda)d_{j+1,n-1},\ j = 1,\ 2,\ \cdots,\ J-1, \\ n = 1,\ 2,\ \cdots,\ N, \\ d_{J,n} = (1 - a\lambda)d_{J,n-1} + a\lambda d_{J-1,n-1},\ n = 1,\ 2,\ \cdots,\ N, \\ d_{0,n} = 0,\ n = 1,\ 2,\ \cdots,\ N, \\ d_{j,0} = R_{j,0}, j = 0,\ 1,\ \cdots,\ J. \end{cases} \tag{5}$$

Vector $\bar{\boldsymbol{D}}_n = (d_{1,n},\ d_{2,n},\ \cdots,\ d_{J,n})^{\mathrm{T}}$ is defined, and the matrix method (Mitchell and Griffiths, 1980) is used below to analyze the propagation law of the error $\|\overline{\boldsymbol{D}}_n\|_\infty$ during the numerical solution. Eq. (5) leads to $\overline{\boldsymbol{D}}_n = \boldsymbol{Q}\overline{\boldsymbol{D}}_{n-1} = \boldsymbol{Q}^n\overline{\boldsymbol{D}}_0 = \boldsymbol{Q}^n\overline{\boldsymbol{R}}_0$, where

$$\boldsymbol{Q} = \begin{bmatrix} 0 & \frac{1}{2}(1 - a\lambda) & 0 & & \\ \frac{1}{2}(1 + a\lambda) & 0 & \frac{1}{2}(1 - a\lambda) & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{1}{2}(1 + a\lambda) & 0 & \frac{1}{2}(1 - a\lambda) \\ & & & a\lambda & 1 - a\lambda \end{bmatrix},$$

and thus we have

$$\|\overline{\mathbf{D}}_n\|_\infty \le \|\boldsymbol{Q}\|_\infty^n \|\overline{\mathbf{R}}_0\|_\infty\ . \tag{6}$$

Next, the stability of Eq. (4) (and that of Eq. (2)) is analyzed. Regardless of whether the initial value problem corresponding to Eq. (1) is solved by the Lax–Friedrichs scheme or the upwind scheme, the stability condition is $a\lambda \le 1$, which is the prerequisite for stability in solving the initial boundary value problem Eq. (1) (Thomas, 1995). Furthermore, the condition for the satisfaction of $\|\boldsymbol{Q}\|_\infty \le 1$ is also $a\lambda \le 1$. In summary, the condition for the stability of numerical Eq. (2) is $a\lambda \le 1$.

From Eq. (6), when Eq. (2) is stable, the round-off error propagated from the initial temporal layer to the Nth top temporal layer satisfies $\|\bar{\boldsymbol{D}}_N\|_\infty \le \|\bar{\boldsymbol{R}}_0\|_\infty$. Generally, if the round-off error is only produced as $R_{j,n}$ at the nth ($\le N$) temporal layer, the error propagated to the Nth temporal layer satisfies the following when the equation is stable:

$$\|\overline{\boldsymbol{D}}_N\|_\infty \le \|\overline{\boldsymbol{R}}_n\|_\infty. \tag{7}$$

3.2. Round-off error analysis

We denote the machine-computed value of $u_{j,n}$ as $(u_{j,n})^*$. The round-off error $R_{j,0} = \varphi_j^* - \varphi_j$ at the initial temporal layer satisfies (Kincaid and Cheney, 2002)

$$|R_{j,0}| \le c_{r0,j,0}\mu \text{ and } c_{r0,j,0} = |\varphi_j| \ (j = 0,\ 1,\ \cdots,\ J), \tag{8}$$

where $r0$ is a symbol as round-off error at initial temporal layer, $\mu = 2^{-q}$ is the machine unit round-off error, and q is the number of binary significant digits.

For the nth (> 0) temporal layer, when $j = 1,\ 2,\ \cdots,\ J-1$, Eq. (2) is implemented using a computer to yield the following equation:

$$(u_{j,n})^* = \left\{ a\tau\frac{(u_{j-1,n-1})^* - (u_{j+1,n-1})^*}{2h} + \frac{1}{2}\left[(u_{j-1,n-1})^* + (u_{j+1,n-1})^*\right] \right.$$
$$\left. +\tau\left(f_{j,n-1}\right)^* \right\}^*. \tag{9}$$

The round-off error of Eq. (9) is

$$(u_{j,n})^* - u_{j,n} = \pi_{1,j,n} + \pi_{2,j,n} + \pi_{3,j,n} + \pi_{4,j,n} \qquad (j = 1,\ 2,\ \cdots,\ J-1), \tag{10}$$

where $\pi_{1,j,n}$ is the round-off error resulting from the first-order numerical differential calculation of the variable x, $\pi_{2,j,n}$ represents the round-off error computed with numerical solution $\frac{(u_{j-1,n-1}+u_{j+1,n-1})}{2}$ at the immediately lower temporal layer, $\pi_{3,j,n}$ denotes the round-off error of term $f_{j,n-1}$, and $\pi_{4,j,n}$ represents the round-off error produced by the addition operation in Eq. (9). Their forms are

$$\pi_{1,j,n} = a\tau\frac{(u_{j-1,n-1})^* - (u_{j+1,n-1})^*}{2h} - a\tau\frac{u_{j-1,n-1} - u_{j+1,n-1}}{2h},$$

$$\pi_{2,j,n} = \frac{1}{2}\left[(u_{j-1,n-1})^* + (u_{j+1,n-1})^*\right] - \frac{1}{2}\left(u_{j-1,n-1} + u_{j+1,n-1}\right),$$

$$\pi_{3,j,n} = \tau\left(f_{j,n-1}\right)^* - \tau f_{j,n-1},$$

$$\pi_{4,j,n} = \left[a\tau\frac{(u_{j-1,n-1})^* - (u_{j+1,n-1})^*}{2h} + \frac{1}{2}\left[(u_{j-1,n-1})^* + (u_{j+1,n-1})^*\right]\right.$$
$$\left. + \tau\left(f_{j,n-1}\right)^*\right]^* - \left[a\tau\frac{(u_{j-1,n-1})^* - (u_{j+1,n-1})^*}{2h} + \frac{1}{2}\left[(u_{j-1,n-1})^* + (u_{j+1,n-1})^*\right]\right.$$
$$\left. + \tau\left(f_{j,n-1}\right)^*\right].$$

There are four round-off errors in $(u_{j,n})^* - u_{j,n}$ in Eq. (10), but they are not all new errors generated at the nth layer, where

$$\pi_{1,j,n} + \pi_{2,j,n} = \frac{1}{2}(1 + a\lambda)\left[(u_{j-1,n-1})^* - u_{j-1,n-1}\right] + \frac{1}{2}(1 - a\lambda)$$
$$\left[(u_{j+1,n-1})^* - u_{j+1,n-1}\right] \qquad (j = 1,\ 2,\ \cdots,\ J-1),$$

perfectly fits the form of error propagation in Eq. (5) when $j = 1,\ 2,\ \cdots,\ J-1$, which implies that two such errors are propagated from the lower temporal layers. $\pi_{3,j,n}$ and $\pi_{4,j,n}$ are round-off errors newly generated by the computation at the nth layer, so

$$\left|R_{j,n}\right| \le \left|\pi_{3,j,n}\right| + \left|\pi_{4,j,n}\right|,$$

where (Kincaid and Cheney, 2002)

$$\left|\pi_{3,j,n}\right| \le \tau\left|f_{j,n-1}\right|\mu,$$

$$\left|\pi_{4,j,n}\right| \le \left(a\left|\frac{u_{j+1,n-1} - u_{j-1,n-1}}{2h}\right|\tau + \frac{1}{2}\left|u_{j-1,n-1} + u_{j+1,n-1}\right| + \left|f_{j,n-1}\right|\tau\right)\mu$$
$$\approx \left(a\left|\frac{\partial u\left(x_j, t_{n-1}\right)}{\partial x}\right|\tau + \left|u_{j,n-1}\right| + \left|f_{j,n-1}\right|\tau\right)\mu.$$

Hence,

$$|R_{j,n}| \le c_{r1,j,n}\mu + ac_{r2,j,n}\tau\mu + c_{r3,j,n}\tau\mu \quad (j = 1,\ 2,\ \cdots,\ J-1,$$
$$n = 1,\ 2,\ \cdots,\ N), \tag{11}$$

where

$$c_{r1,j,n} = \left|u\left(x_j, t_{n-1}\right)\right|,\ c_{r2,j,n} = \left|\frac{\partial u\left(x_j, t_{n-1}\right)}{\partial x}\right|,\ c_{r3,j,n} = 2\left|f\left(x_j, t_{n-1}\right)\right|,$$

where $r1$, $r2$, and $r3$ are symbols of different parts in $|R_{j,n}|$.

Similarly, it can be deduced that the newly generated round-off error at the right-end boundary point $j = J$ is

$$|R_{J,n}| \le c_{r1,J,n}\mu + ac_{r2,J,n}\tau\mu + c_{r3,J,n}\tau\mu \ (n = 1,\ 2,\ \cdots,\ N). \tag{12}$$

Setting

$$C_{r0} = \max_{x\in[0,l]}|\varphi(x)|,\ \hat{C}_{r1} = \max_{x\in[0,l],\ t\in[0,\rho]}|u(x,t)|,$$

$$\hat{C}_{r2} = \max_{x\in[0,l],\ t\in[0,\rho]}\left|\frac{\partial u(x,t)}{\partial x}\right|, \hat{C}_{r3} = \max_{x\in[0,l],\ t\in[0,\rho]}2|f(x,t)|,$$

and with Eqs. (8), (11), and (12), we have

$$\|\overline{\boldsymbol{R}}_0\|_\infty \le C_{r0}\mu,\ \|\overline{\boldsymbol{R}}_n\|_\infty \le \left[\hat{C}_{r1} + \left(a\hat{C}_{r2} + \hat{C}_{r3}\right)\tau\right]\mu \quad (n = 1,\ 2,\ \cdots,\ N). \tag{13}$$

Define C_k as $\rho\hat{C}_k$, and k can take $r1$, $r2$, and $r3$. When Eq. (2) is stable, from Eqs. (7) and (13) it follows that

$$\|R(h,\tau)\|_\infty = \sum\nolimits_{n=0}^{N}\|\bar{\boldsymbol{R}}_n\|_\infty \le \left(C_{r0} + C_{r1}\tau^{-1} + aC_{r2} + C_{r3}\right)\mu. \tag{14}$$

3.3. Truncation error

Maximum principle analysis is applied to estimate the truncation error. The truncation error of Eq. (2) satisfies (Strikwerda, 2004)

$$T_{j,n} = \begin{cases} \frac{1}{2}(1 + a\lambda)T_{j-1,n-1} + \frac{1}{2}(1 - a\lambda)T_{j+1,n-1} + \tau\Lambda_{1,j,n},\ j = 1,\ 2,\ \cdots, \\ J-1, \\ (1 - a\lambda)T_{j,n-1} + a\lambda T_{j-1,n-1} + \tau\Lambda_{2,j,n},\ j = J, \end{cases} \tag{15}$$

where

$$\Lambda_{1,j,n} = \left[\frac{ah^2}{6}\frac{\partial^3 u}{\partial x^3} + \frac{\tau}{2}\frac{\partial^2 u}{\partial t^2} - \frac{h^2}{2\tau}\frac{\partial^2 u}{\partial x^2}\right]_{j,n-1},\ \Lambda_{2,j,n} = \left[\frac{ah}{2}\frac{\partial^2 u}{\partial x^2} + \frac{\tau}{2}\frac{\partial^2 u}{\partial t^2}\right]_{j,n-1}.$$

Eq. (15) leads to

$$\overline{\boldsymbol{T}}_n = \boldsymbol{Q}\overline{\boldsymbol{T}}_{n-1} + \tau\boldsymbol{\Pi}_n\ , \tag{16}$$

where $\boldsymbol{\Pi}_n = [\Lambda_{1,1,n}, \Lambda_{1,2,n}, \cdots, \Lambda_{1,J-1,n}, \Lambda_{2,J,n}]^{\mathrm{T}}$. We let

$$\hat{C}_{t1} = \max_{x\in[0,l],\ t\in[0,\rho]}\frac{1}{6}\left|\frac{\partial^3 u(x,t)}{\partial x^3}\right|,\ \hat{C}_{t2} = \max_{x\in[0,l],\ t\in[0,\rho]}\frac{1}{2}\left|\frac{\partial^2 u(x,t)}{\partial x^2}\right|,$$

$$\hat{C}_{t3} = \max_{x\in[0,l],\ t\in[0,\rho]}\frac{1}{2}\left|\frac{\partial^2 u(x,t)}{\partial t^2}\right|,$$

and then one has

$$\max_{n=1,2,\cdots,N}\|\boldsymbol{\Pi}_n\|_\infty \le a\hat{C}_{t1}h^2 + \hat{C}_{t2}\left(a + h\tau^{-1}\right)h + \hat{C}_{t3}\tau, \tag{17}$$

where $t1$, $t2$, and $t3$ are symbols of different parts in the upper bound of $\max_{n=1,2,\cdots,N}||\boldsymbol{\Pi}_n||_\infty$.

For the definition $C_k = \rho\hat{C}_k$, k can also take $t1$, $t2$, and $t3$. When Eq. (2) is stable, from Eqs. (16) and (17) the truncation error propagated to the top temporal layer satisfies

$$\|T(h,\tau)\|_\infty = \|\overline{\boldsymbol{T}}_N\|_\infty \le \|\boldsymbol{Q}\|_\infty\|\overline{\boldsymbol{T}}_{N-1}\|_\infty + \tau\|\boldsymbol{\Pi}_N\|_\infty \le \|\overline{\boldsymbol{T}}_{N-1}\|_\infty + \tau\|\boldsymbol{\Pi}_N\|_\infty$$
$$\le \|\overline{\boldsymbol{T}}_0\|_\infty + N\tau\max_{n=1,2,\cdots,N}\|\boldsymbol{\Pi}_n\|_\infty = \rho\max_{n=1,2,\cdots,N}\|\boldsymbol{\Pi}_n\|_\infty \le aC_{t1}h^2 + C_{t2}\left(a + h\tau^{-1}\right)h + C_{t3}\tau. \tag{18}$$

3.4. Theoretical estimation of the upper bound of total error

Combining Eqs. (3), (14), and (18), when Eq. (2) is stable, the upper bound of the total error is estimated as

$$\|E(h,\tau)\|_\infty \le \sup\|E(h,\tau)\|_\infty = aC_{t1}h^2 + C_{t2}\left(a + h\tau^{-1}\right)h + C_{t3}\tau + \left(C_{r0} + C_{r1}\tau^{-1} + aC_{r2} + C_{r3}\right)\mu. \tag{19}$$

4. Study on the optimal grid size and time step

The optimal grid size is firstly analyzed. According to Eq. (19), $\sup\|E(h,\tau)\|_\infty$ decreases as h decreases, indicating that the selected optimal grid size should be as small as possible while ensuring that Eq. (2) is stable; that is, it should be selected at the critical point of scheme stability, leading to Theorem 1.

Theorem 1. If the time step is τ, when Eq. (2) is used to solve the initial boundary value problem Eq. (1) of the 1D advection equation, then the optimal grid size is given by

$$H = a\tau. \tag{20}$$

Next, a study on choosing the optimal time step is performed, where the grid size takes its optimal value H. From Eqs. (19) and (20), one has

$$\sup\|E(H,\tau)\|_\infty = a^3C_{t1}\tau^2 + 2a^2C_{t2}\tau + C_{t3}\tau + \left(C_{r0} + C_{r1}\tau^{-1} + aC_{r2} + C_{r3}\right)\mu. \tag{21}$$

Differentiating Eq. (21) with respect to τ yields

$$\frac{\partial \sup\|E(H,\tau)\|_\infty}{\partial\tau} = 0 \Rightarrow 2a^3C_{t1}\tau + 2a^2C_{t2} + C_{t3} - C_{r1}\tau^{-2}\mu = 0, \tag{22}$$

which is difficult to solve. For simplicity, we ignore the high-order term of τ since τ is normally very small, and Eq. (22) becomes

$$2a^2C_{t2} + C_{t3} - C_{r1}\tau^{-2}\mu = 0,$$

which leads to the following theorem.

Theorem 2. When Eq. (2) is used to solve the initial boundary value problem Eq. (1) of the 1D advection equation, and if the grid size takes its optimal value H, the optimal time step is given by

$$\Gamma = \left(\frac{C_{r1}\mu}{2a^2C_{t2} + C_{t3}}\right)^{\frac{1}{2}}. \tag{23}$$

In practical computation, a method for determining the optimal step is provided: first, the optimal time step Γ is determined by Theorem 2, and then the optimal grid size $H = a\Gamma$ can be determined by Theorem 1.

Theorem 3 is established by Theorem 2.

Theorem 3. When the numerical solution is performed under two different machine precisions $\mu_1 = 2^{-q_1}$ and $\mu_2 = 2^{-q_2}$ with q_1 and q_2 ($q_1 < q_2$) binary significant digits, respectively, then the ratio L of the optimal time steps Γ_1 and Γ_2 under the two machine precisions, i.e., L, satisfies

$$L = \frac{\Gamma_1}{\Gamma_2} = 2^{\frac{(q_2-q_1)}{2}}. \tag{24}$$

Theorem 3 indicates that the ratio L of the optimal time steps under two different machine precisions satisfies a universal relation; that is, L is independent of the differential equation, the initial boundary values and the free terms of Eq. (1), and it is exclusively associated with the difference between the numbers of binary significant digits of the two machine precisions involved. This universal relation is similar to that discovered by Li et al. (2000, 2001) in the numerical solution of nonlinear ODEs. With this universal relation, the optimal time step at any machine precision can be immediately determined if the optimal time step at another machine precision is available; therefore, it is very convenient for practical calculation.

5. Theoretical verification of the CUP

When Eq. (2) is stable and the grid ratio is fixed to λ_0, we use Eq. (19) to obtain

$$\sup\|E(h,\tau)\|_\infty = aC_{t1}\lambda_0^{-2}\tau^2 + \left[C_{t2}\left(a + \lambda_0^{-1}\right)\lambda_0^{-1} + C_{t3}\right]\tau + \left(C_{r0} + C_{r1}\tau^{-1} + aC_{r2} + C_{r3}\right)\mu. \tag{25}$$

It should be noted that the grid size h and time step τ vary simultaneously when the grid ratio is fixed, and this situation is different from that in Section 4, in which the grid size is fixed as its optimal value H.

The upper bound $\sup\|E(h,\tau)\|_\infty$ of the total error in Eq. (25) is expressed as $\tilde{E} = \tilde{T} + \tilde{R}$, where the truncation error term $\tilde{T} = aC_{t1}\lambda_0^{-2}\tau^2 + [C_{t2}(a + \lambda_0^{-1})\lambda_0^{-1} + C_{t3}]\tau$ represents the uncertainty caused by the numerical method, the round-off error term $\tilde{R} = (C_{r0} + C_{r1}\tau^{-1} + aC_{r2} + C_{r3})\mu$ represents the uncertainty due to the finite accuracy of the computer, and $\tilde{E}$ is the sum of these two uncertainties.

Theorem 4. When the precision of the machine is finite and the grid ratio is fixed to λ_0, we have

$$\tilde{T} + \tilde{R} > 2\eta_\mu^{\frac{1}{2}} + \sigma_\mu > 0, \tag{26}$$

$$\tilde{T} \cdot \tilde{R} \approx \eta_\mu > 0, \tag{27}$$

where $\eta_\mu = [C_{t2}(a + \lambda_0^{-1})\lambda_0^{-1} + C_{t3}]C_{r1}\mu$ and $\sigma_\mu = (C_{r0} + aC_{r2} + C_{r3})\mu$.

According to the CUP, "the global discretization error due to numerical method and the accumulated round-off error due to calculation machine are two "adjoint" variables; they cannot decrease to zero simultaneously, and the smaller one of the two uncertainties, the greater will be the uncertainty of the other adjoint variable" (Li et al., 2001). From Eqs. (26) and (27), $\tilde{T}$ and $\tilde{R}$ are exactly the two "adjoint" variables in the CUP. The truncation error $\tilde{T}$ is proportional to the time step τ, while the round-off error term $\tilde{R}$ contains a term inversely proportional to the time step τ. They trade off and cannot be reduced to zero at the same time, which demonstrates that solving the initial boundary value problem Eq. (1) with the Lax–Friedrichs scheme satisfies the CUP. If a floating-point computer is used for the numerical solution, round-off error is inevitable, and the CUP always holds true. When the time step τ decreases, the truncation error decreases while the round-off error increases, and the total error decreases initially and then increases under the combined action of both errors. The time step at which the error shifts from decreasing to increasing is the optimal time step. Once the machine accuracy is determined, there will inevitably be an optimal time step in the numerical computation.

6. Numerical experiments

The following two initial boundary value problems of the 1D advection equation are considered:

$$\text{(I):}\ \begin{cases} u_t + au_x = m\cos(x)\cos(mt) - a\sin(x)\sin(mt),\ 0 < x < l,\ t > 0, \\ u(0,t) = \sin(mt),\ t > 0, \\ u(x,0) = 0,\ 0 \le x \le l, \end{cases}$$

which has the exact solution $u(x,t) = \cos(x)\sin(mt)$, and

$$\text{(II):}\ \begin{cases} u_t + au_x = m(t+2)^{m-1}\mathrm{e}^x + a(t+2)^m\mathrm{e}^x,\ 0 < x < l,\ t > 0, \\ u(0,t) = (t+2)^m,\ t > 0, \\ u(x,0) = 2^m\mathrm{e}^x,\ 0 \le x \le l, \end{cases}$$

which has the exact solution $u(x,t) = (t+2)^m\mathrm{e}^x$.

We let $(\mathrm{e}_{j,N})^* = u(x_j, t_N) - (u_{j,N})^*$ be the error of the numerical experiments at the top temporal layer, and set $\boldsymbol{E}^* = ((\mathrm{e}_{1,N})^*, (\mathrm{e}_{2,N})^*, \cdots, (\mathrm{e}_{J,N})^*)^{\mathrm{T}}$. Fig. 1 shows the variations of $\|\boldsymbol{E}^*\|_\infty$ (i.e., the error calculated in the numerical experiments) and

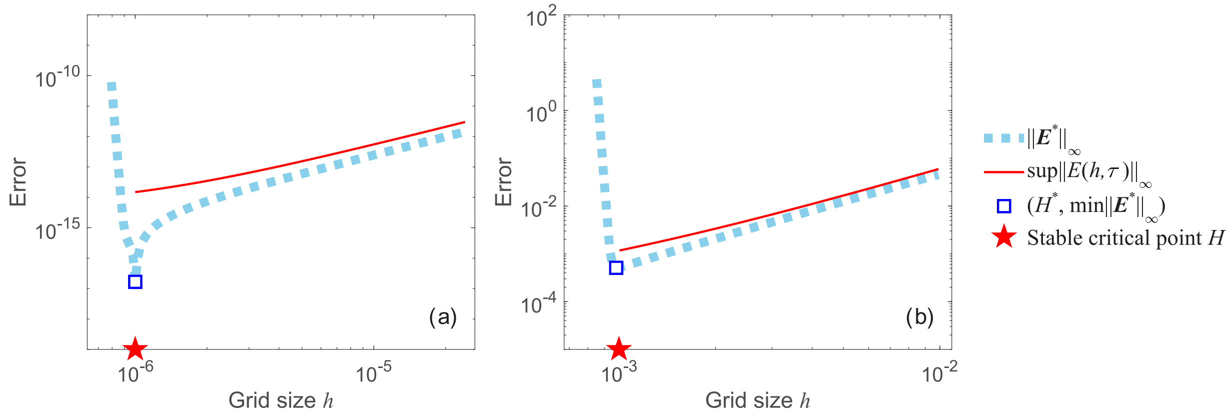


Fig. 1. Variations of the numerical results $\|\boldsymbol{E}^*\|_\infty$ and theoretical bound $\sup\|E(h,\tau)\|_\infty$ with the grid size h, and comparison of the theoretical optimal grid size H and numerical optimal grid size H^*, with $N = 100$: (a) machine double precision was used to solve equation (I), with $a = 1$, $m = 1$, $l = 0.1$, and $\tau = 10^{-6}$; (b) machine single precision was used to solve equation (II), with $a = 10$, $m = 2$, $l = 1$, and $\tau = 10^{-4}$.

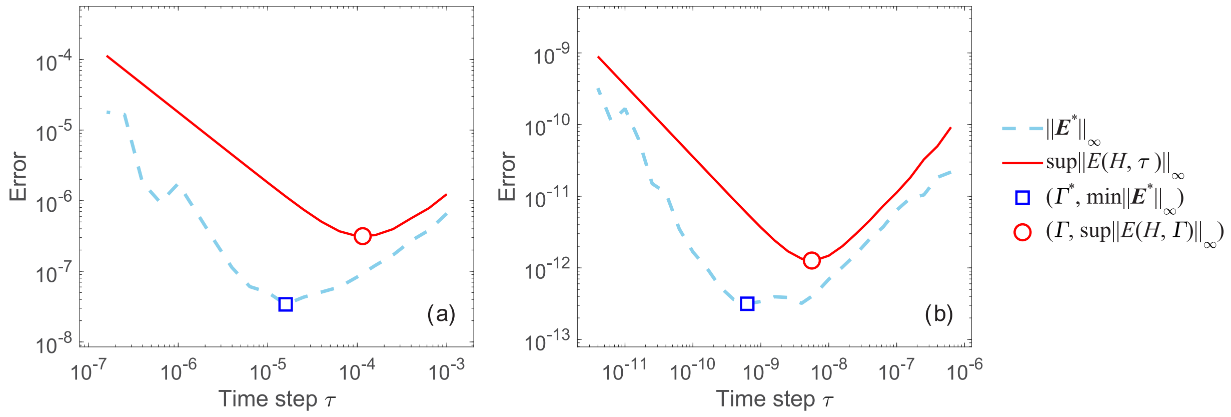


Fig. 2. Variations of the numerical results $\|\boldsymbol{E}^*\|_\infty$ and theoretical bound $\sup\|E(h,\tau)\|_\infty$ with the time step τ when the grid size h takes the optimal value H, and comparison of the theoretical optimal time step Γ and numerical optimal time step Γ^*: (a) machine single precision was used to solve equation (I), with $a = 0.1$, $m = 3$, $l = 10^{-2}$, and $\rho = 10^{-2}$; (b) machine double precision was used to solve equation (II), with $a = 1$, $m = 5$, $l = 10^{-6}$, and $\rho = 10^{-6}$.

$\sup\|E(h,\tau)\|_\infty$ (i.e., the estimated upper bound of the error obtained by theoretical derivation) with the grid size h. When the grid size takes its optimal value H, the variations of the numerical experimental result $\|\boldsymbol{E}^*\|_\infty$ and the theoretical upper bound $\sup\|E(H,\tau)\|_\infty$ with the time step τ are shown in Fig. 2. It can be seen from Figs. 1 and 2 that $\sup\|E(h,\tau)\|_\infty$ is reliable as the upper bound estimation of $\|\boldsymbol{E}^*\|_\infty$. The optimal grid size and the time step determined by the minimum error of the numerical solution in the numerical experiments are denoted as H^* and Γ^*, respectively. As shown in Fig. 1, H accurately estimates H^*. As illustrated in Fig. 2, Γ is slightly larger than Γ^*, but the error does not exceed an order of magnitude; thus, Γ is considered reliable. Moreover, the error increases rapidly when the time step is less than the optimal value, which fully demonstrates the necessity of determining the optimal time step.

7. Conclusions and prospects

When using the Lax–Friedrichs scheme to solve the initial boundary problem of the 1D advection equation, the truncation and round-off errors in the numerical solution were analyzed, and the propagation law of errors from low to high temporal layers was studied. A theoretical estimation for the upper bound of the total error was obtained, with its reliability verified by numerical experimental results. The relationships of the upper bound of the total error with the grid size and time step were analyzed to determine the theoretical formulae for the optimal grid size and time step, and the accuracy of these formulas was demonstrated by the numerical experimental results. On this basis, a universal relation satisfied by the ratio of the optimal time steps under any two different machine precisions was presented, which can be conveniently applied in practical calculation. The CUP for the numerical solution of PDEs was preliminarily and theoretically verified when the grid ratio is fixed. This study could be generalized to more complex numerical schemes for PDEs and atmospheric numerical models for determining optimal spatiotemporal resolutions.

There are two areas for future work. One is the theoretical verification of the CUP when the grid ratio is not necessarily fixed. Another is that the coefficient $\hat{C}_k$ involves the derivatives of $u(x,t)$, which is unknown when the function $u(x,t)$ is undetermined. This problem may be solved by the following method. When the initial boundary value problem is solved by using a certain grid size and time step, a temporary numerical solution of $u(x,t)$ is obtained. With the values of $u(x,t)$ on the grid nodes, numerical differentiation can be implemented to estimate those derivatives of $u(x,t)$ involved in the coefficients $\hat{C}_k$. Then, we can calculate the optimal grid size and time step to obtain the optimal numerical solution. By using this method, we will improve the practicability of the estimation for the optimal grid size and time step in future research work.

Funding

This research was funded by the National Natural Science Foundation of China [grant number 41905092].

References

Bacmeister, J.T., Wehner, M.F., Neale, R.B., Gettelman, A., Hannay, C., Lauritzen, P.H., Caron, J.M., Truesdale, J.E., 2014. Exploratory high-resolution climate simulations using the Community Atmosphere Model (CAM). J. Clim. 27 (9), 3073–3099. doi:10.1175/JCLI-D-13-00387.1.

Cao, J., Li, J.P., Zhang, X.Y., 2017. Interval of effective time-step size for the numerical computation of nonlinear ordinary differential equations. Atmos. Ocean. Sci. Lett. 10 (1), 17–20. doi:10.1080/16742834.2017.1248220.

Deng, L.T., Shi, X.L., Yan, Z.H., 2012. Mesoscale simulation of a heavy rainfall in the Huaihe river valley in July 2003: effects of different horizontal resolutions. J. Trop. Meteorol. 28 (02), 167–176. doi:10.3969/j.issn.1004-4965.2012.02.003. (in Chinese).

Falasca, S., Curci, G., 2018. High-resolution air quality modeling: sensitivity tests to horizontal resolution and urban canopy with WRF-CHIMERE. Atmos. Environ. 187, 241–254. doi:10.1016/j.atmosenv.2018.05.048.

Faragó, I., Horváth, R., 1999. An optimal mesh choice in the numerical solution of the heat equation. Comput. Math. Appl. 38 (9-10), 79–85. doi:10.1016/S0898-1221(99)00263-1.

Guo, D.L., 2021. Study on optimal discretization step size of parabolic partial differential equation. J. Jilin. Norm. Univ. (Nat. Sci. Ed.). 42 (03), 47–51. doi:10.16862/j.cnki.issn1674-3873.2021.03.008. (in Chinese).

Harvey, R., Verseghy, D.L., 2016. The reliability of single precision computations in the simulation of deep soil heat diffusion in a land surface model. Clim. Dyn. 46 (11), 3865–3882. doi:10.1007/s00382-015-2809-5.

Kendon, E.J., Roberts, N.M., Senior, C.A., Roberts, M.J., 2012. Realism of rainfall in a very high-resolution regional climate model. J. Clim. 25 (17), 5791–5806. doi:10.1175/JCLI-D-11-00562.1.

Kincaid, D.R., Cheney, E.W., 2002. Numerical analysis: mathematics of scientific computing. Brooks/Cole, Pacific Grove, CA, p. 48.

Li, J.P., Wang, S.H., 2008. Some mathematical and numerical issues in geophysical fluid dynamics and climate dynamics. Commun. Comput. Phys. 3, 759–793. http://www.global-sci.com/intro/article_detail/cicp/7874.html.

Li, J.P., Zeng, Q.C., Chou, J.F., 2000. Computational uncertainty principle in nonlinear ordinary differential equations (I)——numerical results. Sci. China Ser. E: Eng. Mater. Sci. 43 (05), 449–460. https://link.springer.com/article/10.1007/BF02916726.

Li, J.P., Zeng, Q.C., Chou, J.F., 2001. Computational uncertainty principle in nonlinear ordinary differential equations——II.Theoretical analysis. Sci. China Ser. E: Eng. Mater. Sci. 44 (01), 55–74. doi:10.1007/bf02916726.

Liu, D.Q., Feng, J., Li, J.P., Wang, J.C., 2015. The impacts of time-step size and spatial resolution on the prediction skill of the GRAPES-MESO forecast system. Chin. J. Atmos. Sci. 39 (6), 1165–1178. doi:10.3878/j.issn.1006-9895.1501.14307. (in Chinese).

Maurya, R.K.S., Sinha, P., Mohanty, M.R., Mohanty, U.C., 2018. RegCM4 model sensitivity to horizontal resolution and domain size in simulating the Indian summer monsoon. Atmos. Res. 210, 15–33. doi:10.1016/j.atmosres.2018.04.010.

Mitchell, A.R., Griffiths, D.F., 1980. The finite difference method in partial differential equations. John Wiley & Sons., pp. 40–43.

Shi, Y., Wu, J., Xu, Y., 2021. Role of horizontal resolution in regional climate simulations over the Huang-Huai-Hai River basin. Adv. Water Sci. 32 (6), 843–854. doi:10.14042/j.cnki.32.1309.2021.06.004. (in Chinese).

Strikwerda, J.C., 2004. Finite difference schemes and partial differential equations. SIAM, Philadelphia, pp. 25–26.

Thomas, J.W., 1995. Numerical partial differential equations: finite difference methods, first ed. Springer Science & Business Media, New York, p. 112.

van Roosmalen, L., Christensen, J.H., Butts, M.B., Jensen, K.H., Refsgaard, J.C., 2010. An intercomparison of regional climate model data for hydrological impact studies in Denmark. J. Hydrol. 380 (3-4), 406–419. doi:10.1016/j.jhydrol.2009.11.014.

Wang, P.F., Wang, Z.Z., Huang, G., 2007. The influence of round-off error on the atmospheric general circulation model. Chin. J. Atmos. Sci. 31 (05), 815–825. doi:10.3878/j.issn.1006-9895.2007.05.06. (in Chinese).

Wang, P.F., Zhang, F., 2008. A study on the influence of the computational uncertainty on the IAP-AGCM model. Front. Data. Comput. 03, 32–43 (in Chinese).

William, F.A., 1977. Numerical methods for partial differential equations. Academic press, Orlando, Florida, pp. 43–46.

Yu, F., Huang, L.P., Deng, L.T., 2018. Impacts of different GRAPES-MESO model spatial resolutions on summer rainfall forecast in China. Chin. J. Atmos. Sci. 42 (05), 1146–1156. doi:10.3878/j.issn.1006-9895.1710.17221. (in Chinese).

Zhang, Y., Guo, Z.H., Zhang, W.Y., Huang, H., 2010. Analysis of mesoscale numerical model's ability in atmospheric multi scale characteristics simulation for different resolutions. Chin. J. Atmos. Sci. 34 (03), 653–660. doi:10.3878/j.issn.1006-9895.2010.03.16. (in Chinese).