Journal of Computational Physics 250 (2013) 511–526

Contents lists available at SciVerse ScienceDirect

Journal of Computational Physics

journal homepage: www.elsevier.com/locate/jcp

Optimized explicit finite-difference schemes for spatial derivatives using maximum norm ☆

Jin-Hai Zhang *, Zhen-Xing Yao

Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Article history:
Received 1 September 2012
Received in revised form 27 February 2013
Accepted 21 April 2013
Available online 10 May 2013

Keywords:
Optimized scheme
Explicit finite-difference
Numerical dispersion
Maximum norm
Simulated annealing algorithm

ABSTRACT

Conventional explicit finite-difference methods have difficulties in handling high-frequency components due to strong numerical dispersions. One can reduce the numerical dispersions by optimizing the constant coefficients of the finite-difference operator. Different from traditional optimized schemes that use the 2-norm and the least squares, we propose to construct the objective functions using the maximum norm and solve the objective functions using the simulated annealing algorithm. Both theoretical analyses and numerical experiments show that our optimized scheme is superior to traditional optimized schemes with regard to the following three aspects. First, it provides us with much more flexibility when designing the objective functions; thus we can use various possible forms and contents to make the objective functions more reasonable. Second, it allows for tighter error limitation, which is shown to be necessary to avoid rapid error accumulations for simulations on large-scale models with long travel times. Finally, it is powerful to obtain the optimized coefficients that are much closer to the theoretical limits, which means greater savings in computational efforts and memory demand.

1. Introduction

The explicit finite-difference (FD) scheme is one of the most popular approaches used in various numerical simulations, because it is simple in its numerical implementation and is powerful in handling complex media. However, the conventional explicit FD method has serious numerical artifacts in the presence of high-frequency components and/or coarse grids. This problem would dramatically increase both the demands on the memory and computational cost, especially for large-scale models [7], since a fine grid should be properly designed and a high-order FD operator should be applied. A popular way to avoid this problem is to manually decrease the dominant frequency. This method could result in an acceptable running time, but would result in very limited spatial resolutions, because high-frequency components are necessary for improving the final resolutions. Another way is to apply advanced methods that have less numerical dispersion, such as optimized explicit FD methods and implicit FD methods (either conventional or optimized). Compared with implicit FD methods, explicit FD methods usually have much less computational cost. Therefore, we prefer to develop the optimized scheme of explicit FD methods to further reduce the numerical dispersions while maintaining the computational cost.

Optimized schemes of FD methods appeared two decades ago [10,17]. It has been widely used to reduce the numerical dispersions in many practical applications, such as acoustics, seismology, and electromagnetics. The basic idea of the

* Corresponding author. Tel.: +86 1082998908.
E-mail address: zjh@mail.iggcas.ac.cn (J.-H. Zhang).

0021-9991/$ - see front matter © 2013 The Authors. Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jcp.2013.04.029

optimized scheme is to increase the accurate wave number coverage of the FD operator within a tolerable error range by modifying the constant coefficients. The main advantages of using optimized explicit FD methods are that we can significantly improve the numerical results by maintaining the algorithm structure, the source code and the computational efficiency. In addition, we can use a relatively coarser grid as well as larger time step, hence the memory demand and the total running time would further decrease.

Holberg [10] suggests using the group velocity, which is related to dispersion errors. Etgen [6] suggests employing the phase velocity rather than the group velocity, since the phase velocity is more straight-forward than the group velocity. Some works propose adding a weight function to the objective functions to enhance the influence of the wave number of interest [9,13,16]. However, almost all traditional optimized schemes use the least squares to minimize the objective functions. Thus the forms and the contents of the objective functions are fairly limited and inflexible. Unfortunately, the forms and the contents of the objective functions greatly affect the extent of the improvement in accuracy. In other words, the coverage of accurate wave numbers obtained by traditional optimized schemes is not wide enough because of the limitations created by the objective functions. Therefore there is still some development space for optimized schemes.

In addition, traditional optimized schemes do not pay enough attention to the proper selection of error limitation, which is found to be critical for solid accuracy improvement. Usually traditional optimized schemes tend to employ relatively large error limitation to obtain a wide-enough accurate wave number coverage. Typically, the error limitations shown in the literature range from 0.0003 to 0.03 [10]. Although we may have seen a large accurate wave number coverage range in theoretical analyses, we would ultimately find that an optimized FD operator using a large error limitation is actually apt to obtain less improvement or even worse results compared with un-optimized FD operators [26]. Therefore, we should try to obtain a reasonable accurate range that is as wide as possible; meanwhile we should carefully select the error limitation to guarantee that the accuracy improvements are tangible.

Almost all previous optimized schemes use the 2-norm to construct the objective functions, because such objective functions can be easily solved by the least squares (e.g., [2,3,13,16,23,27]). Whereas, the main task of optimization is to improve the accurate wave number coverage, as widely as possible; thus all our works should contribute to this task, including constructing the objective functions and solving the objective functions. We should not consider too much whether the objective function is easy to solve by some existing method; after all, either the 2-norm or the least squares is just one possible choice. Besides the 2-norm, we can use the 1-norm, the p-norm and the maximum norm. Besides the least squares we can use many other advanced optimization approaches, such as the simulated annealing algorithm [14], the genetic algorithm [11] and the particle swarm optimization [5]. Therefore we may obtain much greater accuracy improvement, or even reach the theoretical upper limit, if we adopt a reasonable objective function, a powerful solver and a proper error limitation.

In this paper, we employ the maximum norm to construct the objective functions and use the simulated annealing algorithm to minimize the objective functions. The maximum norm provides us with an intuitive and effective measure of the optimized FD operator. The simulated annealing algorithm provides us with a more powerful tool in searching for the optimal coefficients. Without being constrained by the solver as before, we can freely design the forms and contents of the objective functions and try to make the objective functions more reasonable. The maximum norm and the simulated annealing algorithm allow us to use much smaller error limitations to make the accuracy improvements more concrete.

Traditional optimization methods have difficulties in evaluating the error before performing the optimization. Thus, they empirically give a possible wave number range that the optimized operator may cover. In consequence, the maximum error will be very high if the expected input wave number is too big; otherwise, the accurate wave number range will be underestimated. In contrast, the new scheme proposed in this paper does not have such a problem. One can determine the preferred error limitation according to the modeling tasks by either the size of the model or the duration of the record. Then the program will try its best to find the optimal coefficients under the given error limitation.

2. Basic forms of optimized FD operator

According to the sampling theory of discrete signals, the band-limit continuous signal $f(x)$ can be recovered by a sinc interpolation of uniformly sampled signals f_n as follows

$$f(x) = \sum_{n=-\infty}^{\infty} \frac{\sin\left[\frac{\pi}{\Delta}(x - n\Delta)\right]}{\frac{\pi}{\Delta}(x - n\Delta)} f_n, \tag{1}$$

where Δ is the grid interval of spatial direction x, and $f_n \equiv f(n\Delta)$ are the sampled values on the discrete positions $n\Delta$. For simplicity, we define $\varphi \equiv \pi/\Delta$ and $\theta \equiv (x - n\Delta)\varphi$; hence the first four orders of spatial derivatives are

$$\frac{\partial f(x)}{\partial x} = \varphi \sum_{n=-\infty}^{\infty} \left(-\frac{\sin\theta}{\theta^2} + \frac{\cos\theta}{\theta} \right) f_n, \tag{2}$$

$$\frac{\partial^2 f(x)}{\partial x^2} = \varphi^2 \sum_{n=-\infty}^{\infty} \left(\frac{2\sin\theta}{\theta^3} - \frac{2\cos\theta}{\theta^2} - \frac{\sin\theta}{\theta} \right) f_n, \tag{3}$$

$$\frac{\partial^3 f(x)}{\partial x^3} = \varphi^3 \sum_{n=-\infty}^{\infty} \left(-\frac{6\sin\theta}{\theta^4} + \frac{6\cos\theta}{\theta^3} + \frac{3\sin\theta}{\theta^2} - \frac{\cos\theta}{\theta} \right) f_n, \tag{4}$$

$$\frac{\partial^4 f(x)}{\partial x^4} = \varphi^4 \sum_{n=-\infty}^{\infty} \left(\frac{24\sin\theta}{\theta^5} - \frac{24\cos\theta}{\theta^4} - \frac{12\sin\theta}{\theta^3} + \frac{4\cos\theta}{\theta^2} + \frac{\sin\theta}{\theta} \right) f_n. \tag{5}$$

These expansions at $x = 0$ are as follows

$$\left.\frac{\partial f(x)}{\partial x}\right|_{x=0} = \sum_{n=-\infty}^{\infty} \frac{\cos(n\pi)}{-n\Delta} f_n, \tag{6}$$

$$\left.\frac{\partial^2 f(x)}{\partial x^2}\right|_{x=0} = \sum_{n=-\infty}^{\infty} \frac{2\cos(n\pi)}{-n^2\Delta^2} f_n, \tag{7}$$

$$\left.\frac{\partial^3 f(x)}{\partial x^3}\right|_{x=0} = \sum_{n=-\infty}^{\infty} \left(\frac{\pi^2}{n\Delta^3} - \frac{6}{n^3\Delta^3} \right) \cos(n\pi) f_n, \tag{8}$$

$$\left.\frac{\partial^4 f(x)}{\partial x^4}\right|_{x=0} = \sum_{n=-\infty}^{\infty} \left(\frac{4\pi^2}{n^2\Delta^4} - \frac{24}{n^4\Delta^4} \right) \cos(n\pi) f_n, \tag{9}$$

respectively. We see that there is a singularity when $n = 0$. To avoid this singularity, we can also express the expansions according to the symmetry as follows [16]

$$\left.\frac{\partial^m f(x)}{\partial x^m}\right|_{x=0} = 2\sum_{n=1}^{\infty} a'_n (f_n - f_{-n}) \quad \text{for } m \text{ is odd}, \tag{10}$$

$$\left.\frac{\partial^m f(x)}{\partial x^m}\right|_{x=0} = a'_0 f_0 + 2\sum_{n=1}^{\infty} a'_n (f_n + f_{-n}) \quad \text{for } m \text{ is even}, \tag{11}$$

where m is the order of the derivatives, and a'_n are the coefficients of f_n in Eqs. (6)-(9), $n = 1,2,\ldots,\infty$.

The conventional explicit FD operators are actually truncated Nth-order expansions multiplied by a window function a_n [8,4], where N is an even integer and a_n is the constant coefficient defined by the binomial coefficient formula

$$a_n = \binom{N}{\frac{N}{2}+n} \bigg/ \binom{N}{\frac{N}{2}}. \tag{12}$$

For the optimized scheme, the basic aim is to search for a new group of coefficients that are different from the above expansions and have better numerical performance. The final form of the optimized FD operator can be expressed as follows:

$$\frac{\partial^m f(x)}{\partial x^m} \approx \frac{1}{\Delta^m} \sum_{n=-N/2}^{N/2} b_n f_n, \tag{13}$$

where b_n are the coefficients that are ready to be optimized.

According to the Fourier transform theory, the spatial derivatives can be equally expressed in the wave number domain as follows [15]

$$\frac{\partial^m f(x)}{\partial x^m} \rightleftharpoons (ik_x)^m F(k_x), \tag{14}$$

where k_x is the wave number, $F(k_x)$ is the forward Fourier transform of $f(x)$, and $i = \sqrt{-1}$. Eq. (14) is the analytical expression of the spatial derivatives in the wave number domain, which covers the whole Nyquist bandwidth. Thus we can examine the accuracy of our optimized FD operators by comparing their Fourier transforms with the analytical wave number $(ik_x)^m$. When $m = 1$, applying a spatial Fourier transformation to (13), we obtain the following relation

$$ik_x F(k_x) \approx \frac{i}{\Delta} \sum_{n=-N/2}^{N/2} b_n \sin(nk_x\Delta) F(k_x) \equiv ik_x^* F(k_x), \tag{15}$$

where k_x^* is defined as the wave number of the optimized FD operator, and k_x^* is an approximation of the analytical wave number k_x. When $m = 2$, we obtain the following relation

$$-k_x^2 F(k_x) \approx \frac{1}{\Delta^2} \sum_{n=-N/2}^{N/2} b_n \cos(nk_x\Delta) F(k_x) \equiv -(k_x^*)^2 F(k_x). \tag{16}$$

3. Objective functions using maximum norm

The 2-norm is the most popular criterion used to construct objective functions (e.g., [2,13,16,19,23,27])

$$E = \int_0^{k_c} |(ik_x)^m - (ik_x^*)^m|^2 w(k_x) dk, \tag{17}$$

where k_c is the maximum accurate wave number under a given error limitation, k_x^* is the approximated wave number, and $w(k_x)$ is some weight function. The least squares are usually used to find the optimized coefficients that minimize the objective functions. The optimized coefficients could be determined by setting

$$\frac{\partial E}{\partial b_n} = 0, \tag{18}$$

and solving the resulting system of linear algebraic equations. The advantage of using the 2-norm and the least squares is that we can obtain a unique group of optimized coefficients. However, the forms and the contents of the objective functions are somewhat limited; that is, one cannot design the objective functions arbitrarily since the designed objective functions should be solvable by the least squares. This limitation makes it difficult to find the optimal group of optimized coefficients, because the flexibility of designing the objective functions would severely influence the final accuracy.

In fact, the 2-norm is only one of the candidates for examining the optimized coefficients. We can also use the 1-norm or the maximum-norm (i.e. the infinite-norm). Following the theory proposed by Tam and Webb [23], Bogey and Bailly [3] minimize the relative difference rather than the traditional absolute difference. Using 1-norm and proper weight functions, obtain much higher accuracy than the standard explicit high-order methods. For all traditional optimization methods, however, it is difficult to provide a proper accurate wave number range, which is required before performing the optimization but is in fact critical for the success of optimization; in addition, they all fall into the least square when minimizing the objective function.

The p-norm is defined as

$$||y||_p = \left(\sum_{j\in\mathbb{N}} |y_j|^p\right)^{1/p}, \quad j = 1, 2, \ldots, J. \tag{19}$$

For $p = 2$, Eq. (19) denotes the 2-norm; for $p = \infty$, Eq. (19) denotes the maximum-norm, which can also be expressed as

$$\|y\|_\infty = \max(|y_1|, |y_2|, \ldots, |y_J|). \tag{20}$$

Recalling that

$$\|y\|_\infty \leqslant \|y\|_2, \tag{21}$$

we see that the maximum-norm is not so strict as the 2-norm. Generally, if $p > r > 0$, we have

$$\|y\|_p \leqslant \|y\|_r. \tag{22}$$

Inequality (22) indicates that the maximum-norm is the loosest among all p-norms. Fortunately, this loosest constraint would not seriously affect the accuracy since the value of $||y||_\infty$ is comparable to that of the 2-norm and 1-norm. The maximum-norm provides us with the largest number of possible solutions under a given error limitation [24]. This would greatly enhance the possibility of finding a group of optimized coefficients when scanning a vast solution set. On the other hand, checking the maximum deviation sounds more reasonable than checking the "distance" between the accurate and approximated wave numbers since it is not working in the space domain. Therefore, we chose the maximum-norm as our criterion for designing the objective functions to extend the accurate wave number coverage as widely as possible.

In this paper, we examine the absolute error between the analytical wave numbers and the approximated wave numbers using the maximum norm. For the optimized FD operators of the frequently-used first- and second-order derivatives, the objective functions are

$$E(k_c, \varepsilon) \equiv \max_{0\leqslant k_x\leqslant k_c} \left| k_x\Delta - \sum_{n=-N/2}^{N/2} b_n \sin(nk_x\Delta) \right| \leqslant \varepsilon \tag{23}$$

and

$$E(k_c, \varepsilon) \equiv \max_{0\leqslant k_x\leqslant k_c} \left| -k_x^2\Delta^2 - \sum_{n=-N/2}^{N/2} b_n \cos(nk_x\Delta) \right| \leqslant \varepsilon, \tag{24}$$

respectively, where k_c is the maximum accurate wave number range that the optimized FD operator can handle, and ε is the error limitation, also called the tolerant threshold. This is probably the most straightforward and simplest objective function that we can find in the literature. It is an intuitive and effective measure of the optimized FD operator.

4. Optimized scheme using simulated annealing

Despite being straightforward and simple, the maximum-norm is actually seldom used in designing the objective functions; in contrast the 2-norm is popular. The main reason is that the maximum-norm cannot be solved easily by the least squares. Holberg [10] presents the absolute error of the group velocity based on the maximum-norm; whereas he uses the 4-norm in practice in order to still use the least squares for determining the optimized coefficients. Lele [17] suggests using the relative error of the optimized FD operator based on the maximum-norm when designing a compact scheme (i.e., optimized Padé scheme); however he determines the optimized coefficients by solving the linear algebraic equations on several specified wave numbers.

Obviously, it is difficult to solve the maximum-norm problem; thus we have to employ a much more complex optimization approach. In this paper, we use the simulated annealing algorithm [14,22], as it has good flexibility in handling various optimization problems. The simulated annealing algorithm is also famous for searching global minima that are buried among many local minima. Therefore it is suitable for our purposes. Fig. 1 shows the flowchart of solving the objective functions based on the maximum norm using the simulated annealing algorithm. In fact, we would obtain many quite different groups of reasonable solutions under a given error limitation rather than only one group as with the least squares; thus we can further select the best one by some tradeoff between the accurate-wave number coverage and the total error (or the peak error).

The flowchart shown in Fig. 1 tries to find the best group of the optimized coefficients under the given error threshold T by continuously searching until it cannot get a better group within the given iteration number N. The best group of optimized coefficients means that it provides the widest accurate-wave number range of $[0, k_c]$. The temperature S basically controls the range of perturbation on the solution; for a high temperature, there is a high possibility to reach a wide range, and vice versa.

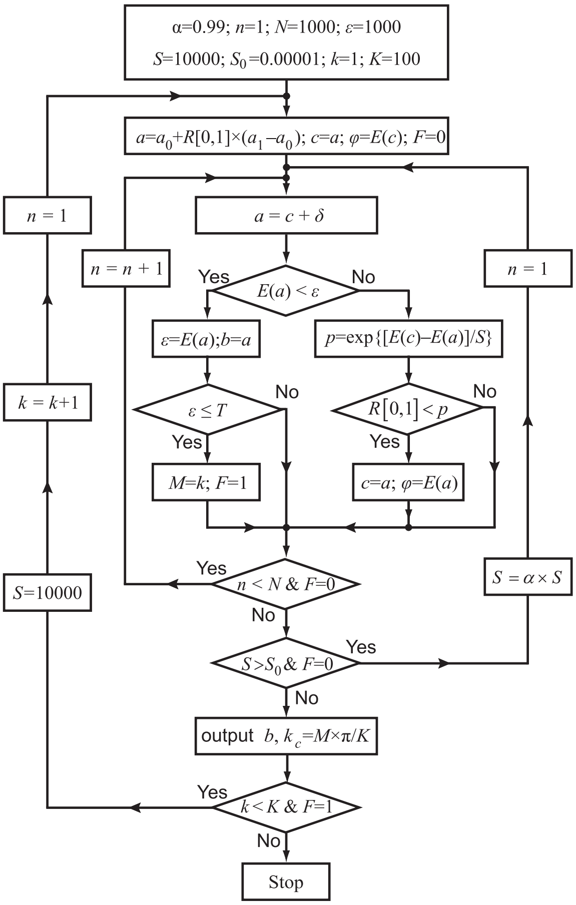

Fig. 1. Flow chart of the optimized scheme using the simulated annealing algorithm. E is the objective functions shown in (23) or (24). b denotes the vector of $b_{-N/2}$ to $b_{N/2}$, δ denotes a small perturbation around b. a_0 and a_1 denote the bottom and upper limits of b, respectively. ε is the absolute error. T is the error limitation. S is the temperature, α is the cooling rate and S_0 is the minimal value of S. $n = 1,2,\ldots,N$ and N is the repeat number under the current temperature S. $k = 1,2,\ldots,K$ and K is the total uniform sampling number of the wave number $k_x \in [0, \pi)$. $k_c = M \times \pi/K$ is the maximum accurate wave number found by the flow chart. M is the index of the maximum accurate wave number k_c. F is a flag: true for 1 and false for 0. $R[0, 1]$ is a function to generate a random number between 0 and 1.

For each searching procedure at the kth wave number, the temperature S would be very high (e.g., 10 000) at the beginning to guarantee that the best solution can be obtained. The temperature would gradually decrease by a factor of $\alpha = 0.99$. For any estimated coefficients c (i.e., $b_{-N/2} \sim b_{N/2}$), some small perturbations δ are added to test whether there are any better

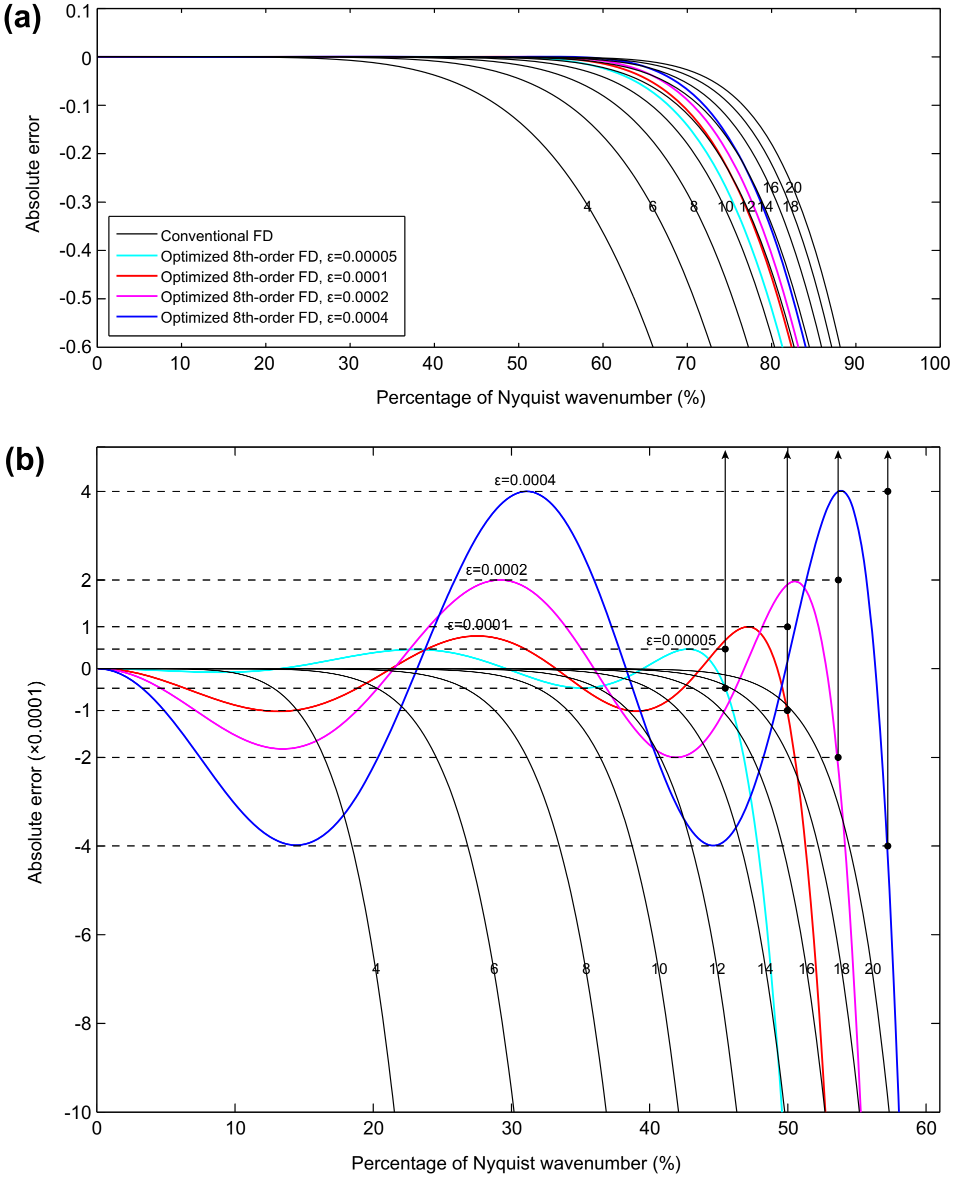

Fig. 2. Influence of the error thresholds on the optimized FD operators. (a) A global view of absolute error within $[-0.6, 0.1]$; (b) a local view within $[-0.001, 0.0004]$. The FD operators are for the second derivative along the spatial direction. Curves denote the absolute errors between the wave number of the FD operators and the analytical wave number. Thin solid curves denote the conventional FD operators with numbers indicating the order. The bold curves denote the optimized 8th-order FD operators using different error limitation ε.

Table 1
Optimized coefficients for high-order FD operators of first derivative.[a]

	4th-Order	6th-Order	8th-Order	10th-Order	12th-Order
b_1	0.67880327	0.77793115	0.84149635	0.88414717	0.91067892
b_2	−0.08962729	−0.17388691	−0.24532989	−0.30233648	−0.34187892
b_3		0.02338713	0.06081891	0.10275057	0.13833962
b_4			−0.00839807	−0.02681517	−0.04880710
b_5				0.00398089	0.01302148
b_6					−0.00199047

[a] $b_0 = 0$, and $b_{-n} = -b_n$, $n = 1, 2, \ldots, N/2$, where N is the order of the optimized FD operator.

coefficients around c. If there are some, the perturbed coefficients $a = c + \delta$ will be set to be the initial values for the next searching procedure at $n + 1$; if there is no, they will still be taken as the potential candidate of the initial values with a random possibility by $R[0,1] < \exp\{[E(c) - E(a)]/S\}$, where E is the maximum of the absolute errors between the analytical wave numbers and the approximated wave numbers for 0 to k (see Eqs. (23) and (24)). If some group of coefficients satisfies

Table 2
Optimized coefficients for high-order FD operators of second derivative.[a]

	4th-Order	6th-Order	8th-Order	10th-Order	12th-Order
b_0	−2.55616844	−2.8215452	−2.97581692	−3.06801592	−3.12513824
b_1	1.37140059	1.57663238	1.70664680	1.78858721	1.84108651
b_2	−0.09331637	−0.18347238	−0.25959423	−0.31660756	−0.35706478
b_3		0.01761260	0.04618682	0.07612137	0.10185626
b_4			−0.00533093	−0.01626042	−0.02924772
b_5				0.00216736	0.00696837
b_6					−0.00102952

[a] $b_{-n} = b_n$, $n = 1,2,\ldots,N/2$, where N is the order of the optimized FD operator.

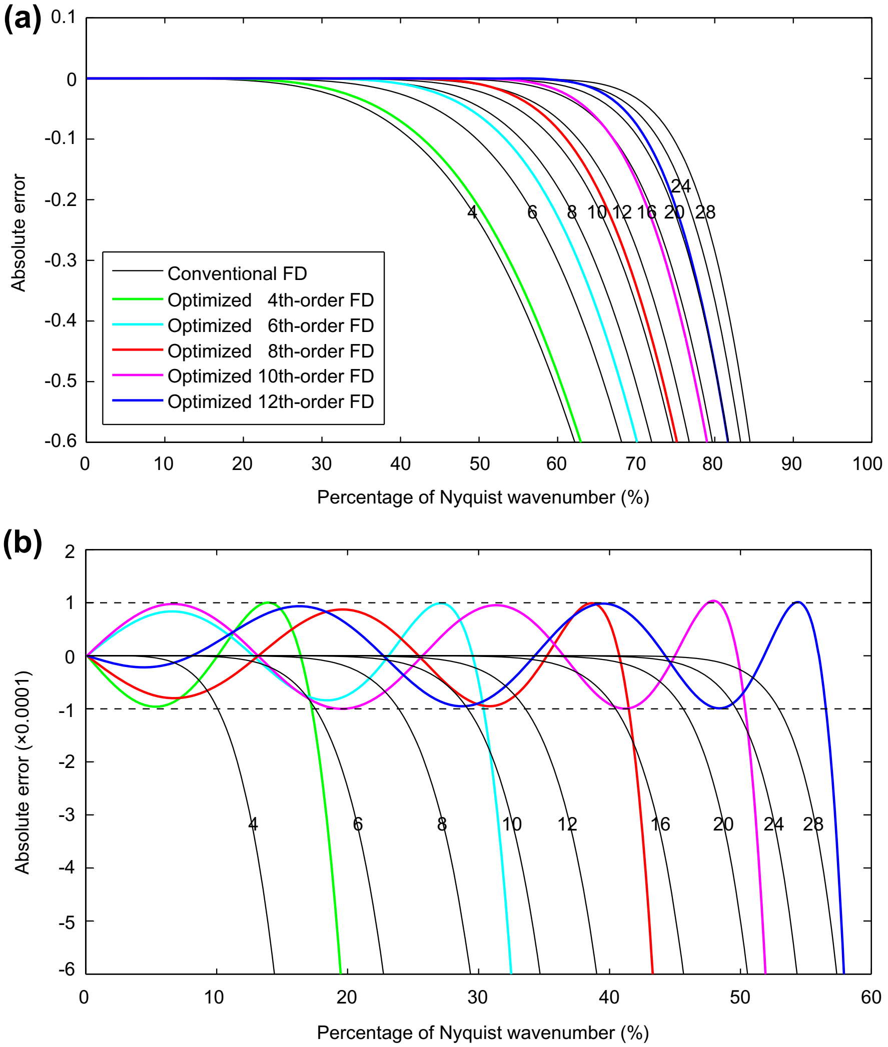

Fig. 3. Accuracy comparison between the conventional and optimized FD operators of the first derivative. (a) A global view of absolute error within $[-0.6, 0.1]$; (b) a local view within $[-0.0006, 0.0002]$. Curves denote the absolute errors between the wave number of the FD operators and the analytical wave number. Thin solid curves denote the conventional FD operators with numbers indicating the order. The bold curves denote the optimized FD operators using the error limitation $\varepsilon = 0.0001$.

$E(a) < T$, then it will be remembered as the potential solutions b for 0 to k. Next, we would further test whether there is some solution for 0 to $k+1$. If there is no solution for 0 to $k+1$, the coefficients b for 0 to k would be the final solution under the given error threshold T, and the maximum accurate wave number k_c is $k \times \pi/K$.

The total number of optimized coefficients is N+1, i.e., $b_{-N/2} \sim b_{N/2}$, which is difficult to determine with the simulated annealing algorithm when N is large. To reduce the optimization effort, Zhang and Yao [26] set up three criteria according to the theories of sinc interpolation [4] and finite impulse response [21] for the second order derivative. We extend Zhang and Yao's criteria to more general cases: (1) the coefficients should be real numbers $b_n \in R$, and the operator should be symmetric for even order derivatives (i.e., $b_{-n} = b_n$) and be anti-symmetric for odd order derivatives (i.e., $b_{-n} = -b_n$); (2) the total energy of the optimized FD operator should be zero for both even and odd order derivatives, that is $\sum_{n=-N/2}^{N/2} b_n = 0$; (3) the coefficients should have an amplitude of damped oscillation away from the center position ($n = 0$), that is $|b_n| > |b_{n+1}|$ and $b_n b_{n+1} < 0$ for $n \geqslant 0$; (4) to cover a much wider wave number range, all coefficients should be as large as possible (including b_0 and $b_{N/2o}$).

Rules 1 and 2 reduce the actual number of coefficients to only $N/2$, since $b_0 = -2\sum_{n=1}^{N/2} b_n$ for even order derivatives and $b_0 = 2\sum_{n=1}^{N/2} b_n = 0$ for odd order derivatives. Thus we can obtain the whole operator by purely determining the independent coefficients $b_1 \sim b_{N/2}$(or $b_{-N/2} \sim b_{-1}$). Rules 3 and 4 greatly decrease the scope of the search and make the simulated annealing algorithm affordable for high-order FD operators. In fact, the original coefficients of the conventional FD operators also obey these criteria.

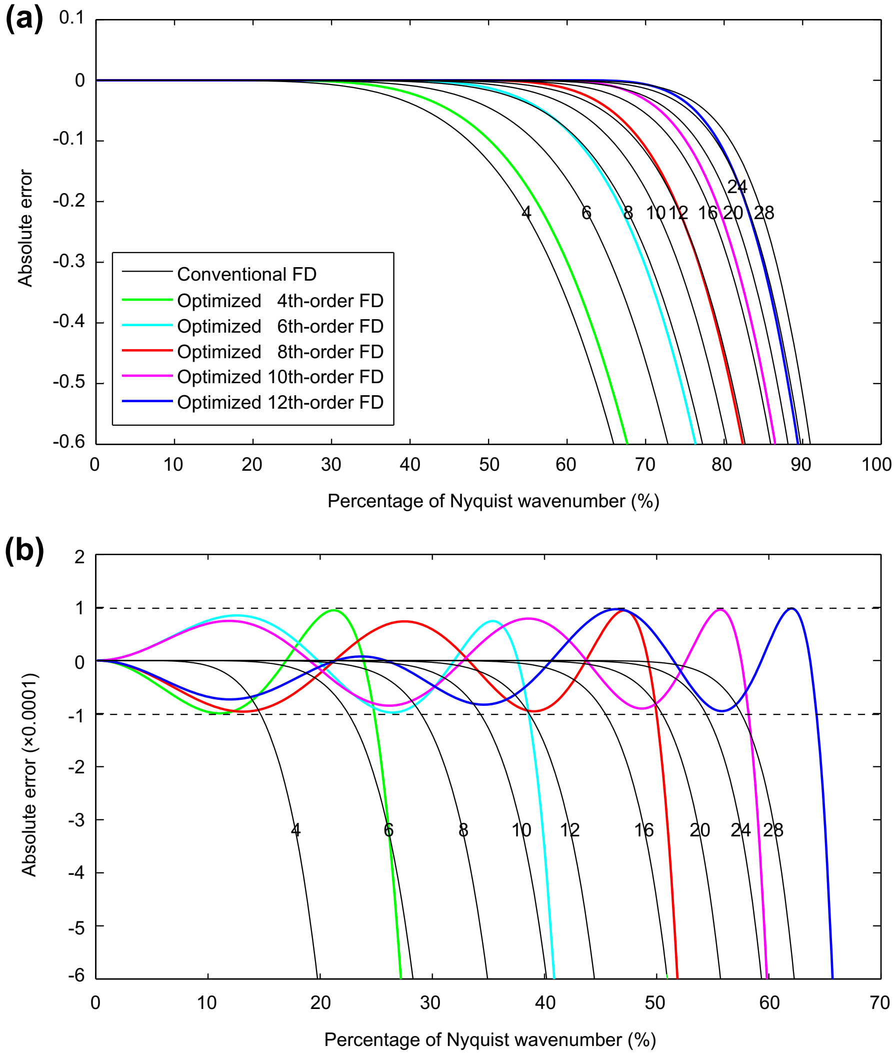

Fig. 4. Accuracy comparison between the conventional and optimized FD operators of the second derivative. See Fig. 3 for detail captions.

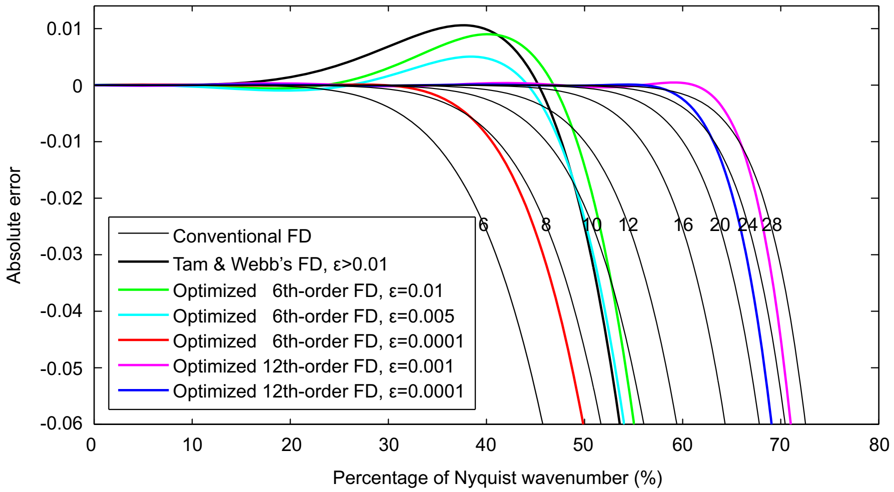

Fig. 5. Accuracy comparison between the Kam and Webb's operator and our optimized FD operators of the first derivative. Thin solid curves denote the conventional FD operators with numbers indicating the order. The bold curves denote the optimized FD operators using different error limitations. The four bold curves on the left side are for the 6th-order FD operators and the two bold curves on the right side are for the 12th-order FD operators, respectively.

Table 3
Optimized coefficients used in Figs. 5–7.[a]

	Tam and Webb's using 0.01	Our using 0.01	Our using 0.005	Our using 0.0005
b_1	0.79926643	0.80961299	0.80359697	0.92014414
b_2	−0.18941314	−0.20184244	−0.19704245	−0.35645100
b_3	0.02651995	0.03151878	0.03021447	0.15232195
b_4				−0.05826689
b_5				0.01743499
b_6				−0.00314725

[a] $b_0 = 0$, and $b_{-n} = -b_n$, $n = 1,2,\ldots,6$. Our optimized coefficients using 0.0001 are listed in Tables 1 and 2.

5. Proper selection of error threshold

The error threshold ε plays an important role in the optimized scheme of the FD operator [25,26]. For a small error limitation (e.g., 0.00001) it can guarantee the accuracy of the resulting operators, but would make it difficult to gain an apparent improvement. For a big error limitation (e.g., 0.0003–0.03 as suggested by Holberg [10] and by many other works) it can easily cover a much wider wave number range; unfortunately, the practical performances shown in numerical experiments may greatly deviate from the theoretical analyses, especially for large travel times or at large distances. Therefore it is necessary to select a proper error limitation for the objective functions to guarantee that the accuracy improvements are apparent and solid.

Using the phase velocity is more convenient than using the group velocity when designing the objective function [6,13,16]. Basically, the absolute error of the operator in the wave number domain is similar to the phase velocity. Whereas Holberg [10] points out that the objective function based on the phase velocity should have a much smaller error limitation than that based on the group velocity, which is about one order of magnitude smaller. However, our experiments show that the practical error limitation is not necessarily as small as suggested by Holberg [10] to obtain the same accuracy. Nevertheless, we still suggest using a tight error limitation since in practice the requirement on the accuracy is always increasing.

6. Absolute-error analyses

We evaluate the accuracy performance of the optimized FD operators by examining its absolute errors in the wave number domain. First, we show the accuracy influences caused by different error limitations. In Fig. 2 we take the 8th-order FD operator of the second derivative as an example. Obviously, the absolute-error curves of the conventional FD operators increase gradually with increasing wave numbers; whereas the absolute-error curves of the optimized FD operators vibrate rapidly within the given error limitation. The optimized FD operators have a much wider accurate-wave number coverage at the cost of much more errors that are evenly distributed within the "accurate-wave number" coverage. Fortunately these error limitations are small enough for many practical applications.

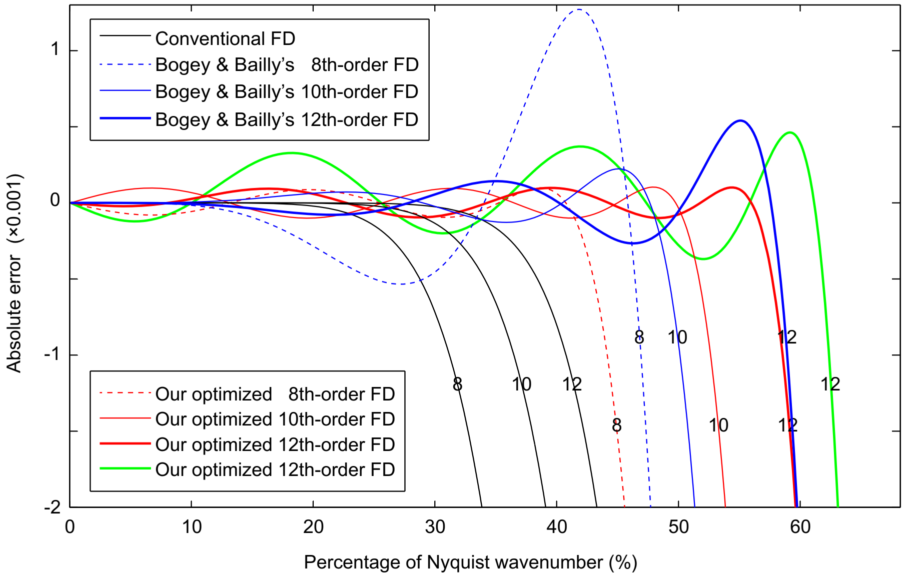

Fig. 6. Accuracy comparison between the Bogey and Bailly's operators and our optimized FD operators of the first derivative. Black thin solid curves denote the conventional FD operators with numbers indicating the order. The dashed curves denote the 8th-order optimized FD operators, the colorful thin curves denote the 10th-order optimized FD operators, and the bold solid curves denote the 12th-order optimized FD operators. The red curves are generated using the error limitation of 0.0001, and the green curve is generated using 0.0005. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

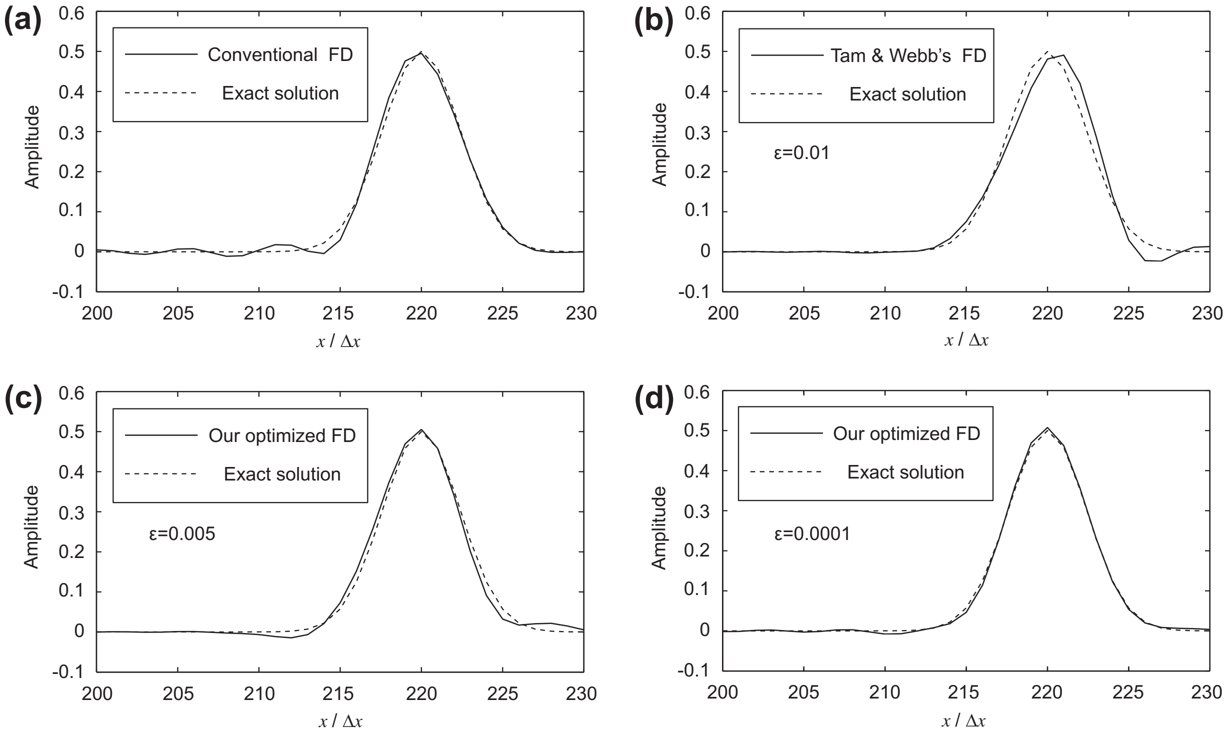

Fig. 7. Comparison between numerical results obtained by 6th-order FD operators. Dashed curves are obtained by the analytical solution, and solid curves are obtained by (a) conventional 6th-order FD operator; (b) Kam and Webb's operator (corresponding to $\varepsilon > 0.01$); (c) optimized 6th-order FD operator using $\varepsilon = 0.0005$; (d) optimized 6th-order FD operator using $\varepsilon = 0.0001$.

The error limitations listed in Fig. 2 (i.e., 0.00005–0.0004) are all far lower than those listed in the literature (e.g., 0.0003–0.03). Obviously, a bigger error limitation would lead to greater accuracy improvements but much larger peak errors; meanwhile, we see that only a doubled error limitation would earn similar accuracy improvements, as indicated by the black dots and the vertical arrows. Therefore, we have to make a balance between the accuracy improvements and the peak errors. We prefer a tight error limitation to avoid rapid error accumulation, especially for large-scale and long-term problems. We select 0.0001 as our error limitation for later experiments. Tables 1 and 2 list the optimized coefficients under this error limitation for the first and second derivatives, respectively.

Figs. 3 and 4 show the 4th- to 12th-order optimized FD operators for the first and second derivatives, respectively. Obviously, the accuracy of the optimized FD operator has a wider accurate wave number coverage than does the conventional same-order FD operator. In addition, the higher-order optimized FD operators have much wider accurate wave number coverage than do the lower-order optimized FD operators. For example, the accuracy of the optimized 4th-order FD operator is only slightly higher than that of the conventional 4th-order FD operator. In contrast, the accuracy of the optimized 8th-order FD operator is much higher than that of the conventional 8th-order FD operator, and even reaches that of the conventional 12th-order FD operator. Furthermore, the accuracy of the optimized 12th-order FD operator is much higher than that of the conventional 12th-order FD operator and even reaches that of the conventional 24th-order FD operator. Therefore, we suggest using the higher-order optimized FD operators in practical applications since they have much higher accuracy compared with the lower-order optimized FD operators.

7. Experiments on 1D advection equation

To illustrate the optimized scheme proposed in this paper, we compare our optimized FD operators with Kam and Webb's optimized FD operator (1993). We consider the 1D advection equation

$$\frac{\partial u}{\partial t}+\frac{\partial u}{\partial x}=0, \tag{25}$$

with an initial disturbance

$$u(x,t=0)=\frac{1}{2}\exp\left[-\ln 2\frac{(x-20)^2}{\sigma}\right], \tag{26}$$

where $0 \leqslant x \leqslant 400$ and the grid interval $\Delta = 1$. Fig. 5 shows the curves of three different 6th-order optimized FD operators using error limitations of 0.01, 0.005 and 0.0001. The optimized coefficients used in Fig. 5 are listed in Table 3. Obviously, our

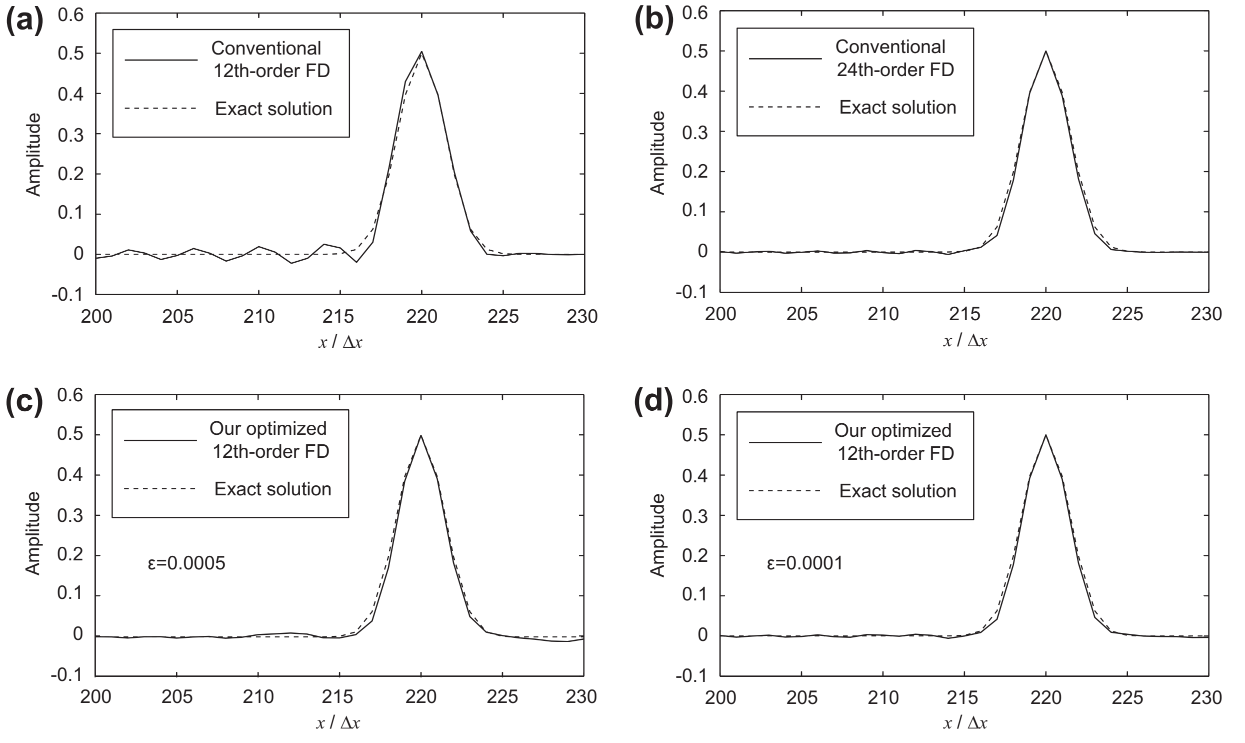

Fig. 8. Comparison between numerical results obtained by 12th-order FD operators. Dashed curves are obtained by the analytical solution, and solid curves are obtained by (a) conventional 12th-order FD operator; (b) conventional 24th-order FD operator; (c) optimized 12th-order FD operator using $\varepsilon = 0.001$; (d) optimized 12th-order FD operator using $\varepsilon = 0.0001$.

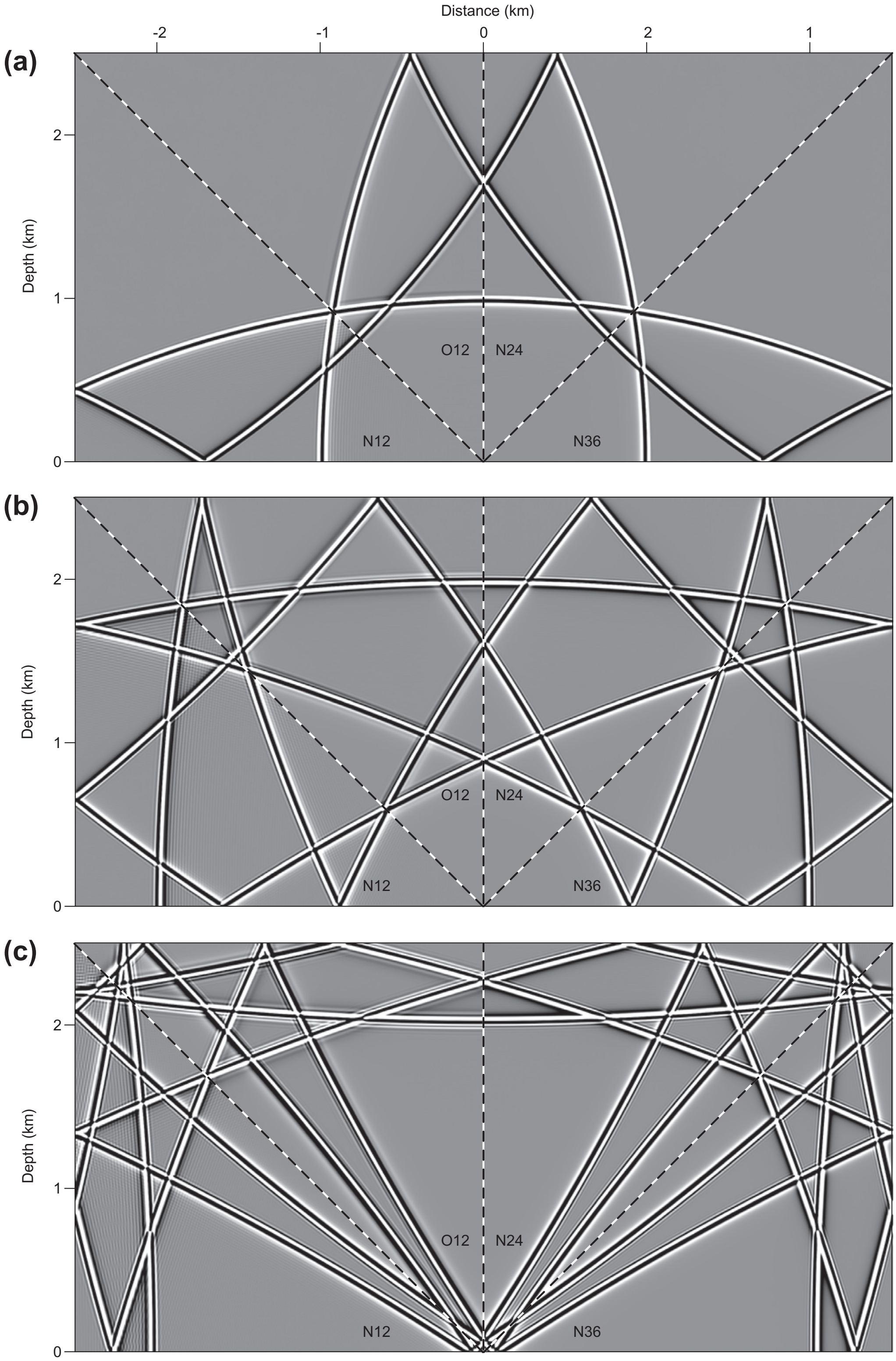

Fig. 9. Snapshot comparison among different methods at three travel times: (a) 3 s; (b) 6 s; (c) 9 s. Each subfigure has four equivalent parts and they are separated by dashed lines. Snapshots are generated by the conventional 12th-order, 24th-order and 36th-order FD methods and the optimized 12th-order FD method, respectively. They are sequentially indicated by N12, N24, N36 and O12. The conventional 36th-order FD method are shown as references.

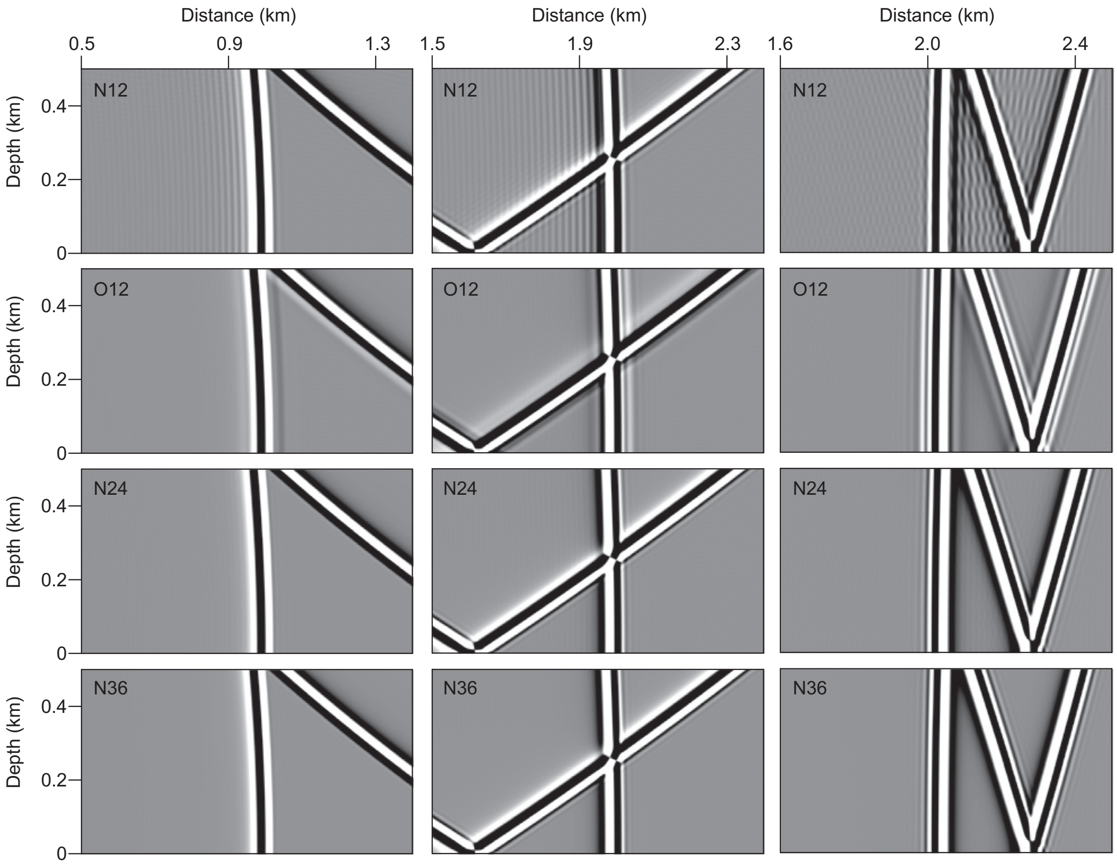

Fig. 10. Comparison of local details in Fig. 9. The depth range is $0 \leqslant z \leqslant 500$ m for all sub Figures.. Left column: $500 \leqslant x \leqslant 1400$ m at 3 s; middle column: $1500 \leqslant x \leqslant 2400$ m at 6 s; right column: $1600 \leqslant x \leqslant 2500$ m at 9 s.

three operators have smaller peak errors than that of Tam and Webb [23]. In addition, one of our operators shows wider wave number coverage as well as smaller peak errors than do Tam and Webb's operator. This indicates that our maximum-norm objective functions as well as the simulated annealing algorithm are better than the 2-norm objective functions and the least squares. On the other hand, the other two operators obtained by our scheme seem to have a narrower wave number coverage than Tam and Webb's operator; surprisingly, our numerical experiments show that this is actually not the case.

We also compare our optimized coefficients with some existing optimized coefficients for high-order FD operators [3]. Bogey and Bailly call the 8th-, 10th- and 12th-order operators here as 9-, 11- and 13-point stencils. Fig. 6 show the difference between our results and their results. For each order listed, our optimized operator generally has quite similar wave number range but much smaller maximum error. For example, our operators have an error limitation of only 0.0001 (see the red curves), but theirs have error limitations of 0.0011, 0.0002 and 0.0006 (see the blue curves), respectively. If we use much looser error limitation, such as 0.0005 (see the green bold curve), the wave number coverage will be much wider compared with the blue bold curve. We do not show the waveform comparison since the difference in waveform is not so significant in small scale model or short duration of the record. However, note that a big error in wave number domain always has a big risk in the presence of either long-period or over-size problems.

As shown in Fig. 7 (corresponding to $\sigma = 8$), the optimized FD operator using 0.005 is apparently superior to Tam and Webb's operator that uses 0.01; in addition, the optimized FD operator using the error limitation of 0.0001 is better than that using 0.005. Fig. 8 (corresponding to $\sigma = 3$) shows the case of the 12th-order optimized FD operators, which also indicates that a small error limitation is better than a large error limitation. Therefore, we should use a small error limitation from now on rather than purely pursuing much wider wave number coverage by arbitrarily relaxing the error limitation.

8. Experiments on 2D wave equation

In this section, we consider the 2D scalar wave equation

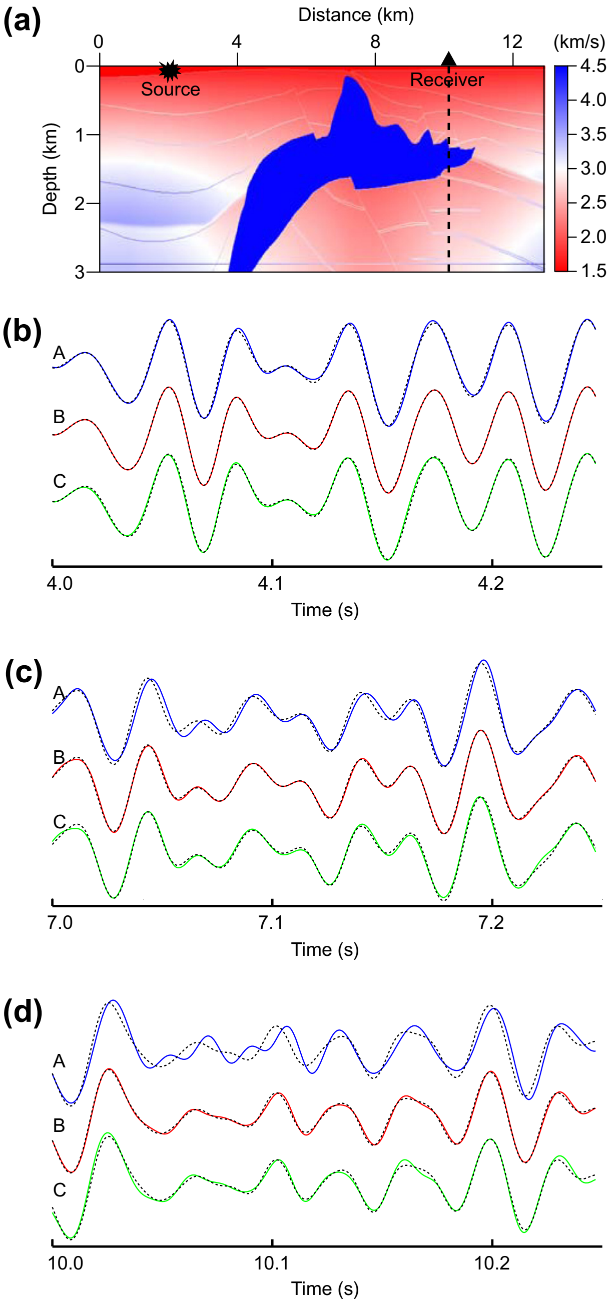

Fig. 11. Waveforms comparison among different methods for SEG/EAGE salt model. (a) Modified SEG/EAGE salt model; (b) 4.0–4.25 s (beginning from the first arrival); (c) 7.0–7.25 s; (d) 10.0–10.25 s (including multiple reflected waves from the salt dome). Waveforms are generated by the conventional 12th-order, 24th-order FD methods and the optimized 12th-order FD method, respectively. The conventional 36th-order FD method are shown as references (see dashed curves).

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial z^2}=\frac{1}{v^2(x,z)}\frac{\partial^2 u}{\partial t^2}+\delta(x_s,z_s)S(x,z;t), \tag{27}$$

where t is the time, $v(x,z)$ is the 2D velocity function, and $u\equiv u(x,z;t)$ is the wave field. We take the Ricker wavelet

$$S(x,z;t)=(1-2\pi^2\omega^2t^2)\exp(-\pi^2\omega^2t^2) \tag{28}$$

as the initial waveform, where ω is the dominant frequency.

First, we illustrate the above absolute-error analyses by impulse responses in a 2D homogeneous medium, where $-2500\leqslant x\leqslant 2500$ m and $-2500\leqslant z\leqslant 2500$ m with a uniform grid spacing of 5 m. The source location is located at

$x_s = 0$ m and $z_s = 0$ m. The velocity is $v = 2000$ m/s and the dominant frequency $\omega = 50$ Hz. Note that the dominant frequency and the velocity used here almost reach the upper limit that is fairly difficult to handle in practice, and the scale of the model is also typical in practical applications in geophysical exploration (e.g., [25]). Figs. 9(a)–(c) show the wavefield snapshots at 3 s, 6 s and 9 s, respectively. Fig. 10 shows the local details of Fig. 9. Obviously, the optimized 12th-order FD methods obtain much better results compared with the conventional 12th-order FD method. In addition, the results obtained by the optimized 12th-order FD method are quite similar to those obtained by the conventional 24th-order FD method. Figs. 9 and 10 indicate that the improvement after using our optimization scheme is significant, even for the large-scale model at a long travel time. Note that these numerical experiments show perfect agreement with the theoretical analyses in the previous section.

To verify the capabilities of our optimized FD operators, we simulate the wave field propagations on a modified SEG/EAGE salt model [1]. Fig. 11(a) shows the velocity model and Fig. 11(b)–(d) show the waveforms at three time windows. For convenience of comparison, we take the waveforms generated by the conventional 36th-order FD method as references, as shown by the dashed curves. We see that the waveforms generated by the conventional 12th-order FD method evidently deviate from the reference waveforms due to numerical dispersion. In contrast, the waveforms obtained by the optimized 12th-order FD method are almost the same as those obtained by the conventional 24th-order FD method. In addition, the waveforms obtained by the optimized 12th-order FD method only show slight differences from the reference waveforms at 10 s (see Fig. 11(d)). Fig. 11 indicates that the optimized FD method is superior to the conventional FD method for the same order. Again, these conclusions are consistent with the theoretical and numerical analyses in the previous sections.

9. Discussions

Only the absolute error is used in our objective functions. In fact, we can also try relative error [3], or add a proper weight function to important wave numbers, or try any other possible forms of the objective function, to obtain further improvement. Fortunately, the extension is straightforward to the proposed scheme. As a general approach for optimizing FD operators, our optimized scheme can be applied to other orders of spatial derivative, for example the third and fourth derivatives [17,20]. In general, the optimized scheme is applicable to various equations that do not contain cross derivatives. This paper only concentrates on the FD discretization of spatial derivatives. We can also take the FD discretization of the temporal derivative into account [3,4,12] or extend our method to higher order time discretization [18]. Of course, this scheme is also available for compact FD operators [17,27] to achieve additional accuracy improvements.

10. Conclusions

We present a new optimization scheme to reduce the numerical dispersions of high-order explicit FD methods. The objective functions are constructed with the maximum norm rather than the traditional 2-norm; in addition, we solve the objective functions using the simulated annealing algorithm rather than the traditional least squares. The maximum norm provides us with the largest number of possible solutions, which greatly enhances the possibility of finding the optimized coefficients for the simulated annealing algorithm over a vast solution set.

We show that the error limitation is essential for solid accuracy improvements. A small error limitation is superior to a large error limitation, although we may draw the opposite conclusion according to the theoretical analyses. This indicates that we should use a small error threshold (e.g., 0.0001) to guarantee accuracy for large-scale modeling with long travel times, rather than purely pursuing the accurate wave number coverage by arbitrarily relaxing the error limitations.

For both the first and second spatial derivatives, our optimized 8th-order FD method has the same accuracy as the conventional 12th-order FD method, and our optimized 12th-order FD method has the same accuracy as the conventional 24th-order FD method. This means we can greatly save on both memory demand and computational cost when using our optimized high-order FD methods.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant No. 41074092 and 41130418) and the National Major Project of China (Grant No. 2011ZX05008-006).

References

[1] F. Aminzadeh, N. Burkhard, J. Long, T. Kunz, P. Duclos, Three dimensional SEG/EAEG models—An update, The Leading Edge 15 (1996) 131–134.

[2] G. Ashcroft, X. Zhang, Optimized prefactored compact schemes, J. Comput. Phys. 190 (2003) 459–477.

[3] C. Bogey, C. Bailly, A family of low dispersive and low dissipative explicit schemes for flow noise and noise computations, J. Comput. Phys. 194 (2004) 194–214.

[4] C. Chu, P.L. Stoffa, Determination of finite-difference weights using scaled binomial windows, Geophysics 77 (2012) W17–W26.

[5] R.C. Eberhart, J. Kennedy, A new optimizer using particle swarm theory. Proceedings of the Sixth International Symposium on Micromachine and Human Science, Nagoya, Japan. 39–43, 1995.

[6] J.T. Etgen, A tutorial on optimizing time domain finite-difference schemes: "Beyond Holberg", Stanford Exploration Project Report 129 (2007) 33–43.

[7] J.T. Etgen, M.J. O'Brien, Computational methods for large-scale 3D acoustic finite-difference modeling, A tutorial, Geophysics 72 (2007) SM223–SM230.

[8] B. Fornberg, Calculation of weights in finite difference formulas, SIAM Review 40 (1998) 685–691.

[9] Z. Haras, S. Ta'asan, Finite difference schemes for long-time integration, J. Comput. Phys. 114 (1994) 265–279.

[10] O. Holberg, Computational aspects of the choice of operator and sampling interval for numerical differentiation in large-scale simulation of wave phenomena, Geophys. Prospect. 35 (1987) 629–655.

[11] J.H. Holland, Genetic Algorithms, Sci. Am. 267 (1992) 66–72.

[12] F.Q. Hu, M.Y. Hussaini, J.L. Manthey, Low dissipation and low-dispersion Runge-Kutta schemes for computational acoustics, J. Comput. Phys. 124 (1996) 177–191.

[13] J.W. Kim, D.J. Lee, Optimized compact finite difference schemes with maximum resolution, AIAA J. 34 (1996) 887–893.

[14] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Optimization by simulated annealing, Science 220 (1983) 671–680.

[15] D.D. Kosloff, E. Baysal, Forward modeling by a Fourier method, Geophysics 47 (1982) 1402–1412.

[16] C. Lee, Y. Seo, A New Compact Spectral Scheme for Turbulence Simulations, J. Comput. Phys. 183 (2002) 438–469.

[17] S. Lele, Compact finite difference schemes with spectral-like resolution, J. Comput. Phys. 103 (1992) 16–42.

[18] X. Li, W. Wang, M. Lu, M. Zhang, Y. Li, Structure-preserving modelling of elastic waves: a symplectic discrete singular convolution differentiator method, Geophys. J. Int. 188 (2012) 1382–1392.

[19] J. Gao, Z. Yang, X. Li, An optimized spectral difference scheme for CAA problems, J. Comput. Phys. 231 (2012) 4848–4866.

[20] Y. Liu, M.K. Sen, A practical implicit finite-difference method: Examples from seismic modeling, J. Geophys. Eng. 6 (2009) 231–249.

[21] A.V. Oppenheim, R.W. Schafer, J.R. Buck, Discrete-time signal processing, 2nd edition., Prentice-Hall, New York, 1999.

[22] M.K. Sen, P.L. Stoffa, Global Optimization methods in geophysical inversion, 2nd edition., Cambridge University Press, Cambridge, 2013.

[23] C.K.W. Tam, J.C. Webb, Dispersion-relation-preserving schemes for computational aeroacoustic, J. Comput. Phys. 107 (1993) 262–281.

[24] A. Tarantola, Inverse problem theory and methods for model parameter estimation, SIAM, Philadelphia, 2005.

[25] J.H. Zhang, Z.X. Yao, Globally optimized finite-difference extrapolator for strongly VTI media, Geophysics 77 (2012) S125–S135.

[26] J.H. Zhang, Z.X. Yao, Optimized finite-difference operator for broad-band seismic wave modeling, Geophysics 77 (2013) A13–A18.

[27] H. Zhou, G. Zhang. Prefactored optimized compact finite-difference schemes for second spatial derivatives, Geophysics 76 (2011) WB87–WB95.