J. Quant. Spectrosc. Radiat. Transfer, Vol. 18, pp. 245-248. Pergamon Press 1977. Printed in Great Britain

DOUBLING INITIALIZATION REVISITED

W. J. Wiscombe

National Center for Atmospheric Research, Boulder, CO 80307, U.S.A.†

(*Received* 1 *February* 1977)

Abstract—In previous work, the author studied errors arising solely from initial-layer approximations in doubling. It now appears that misleading conclusions may be drawn from that work, because it failed to consider the *interaction* of angular and initial-layer error. That interaction is such that decreasing initial-layer error (a) often has *no effect* on total error, or (b) sometimes *increases* total error. One concludes from this that, contrary to accepted practice, it is advisable to use an initial layer of fairly *large* optical depth, in order to strike a balance between angular and initial-layer error. The diamond initialization still seems generally superior, but only dramatically so for high orders of angular approximation.

INTRODUCTION

In calculating any radiative quantity such as layer albedo by the doubling method, we shall always arrive at an approximate value, A_{approx}, instead of the exact value, A. The two are related by

$$A_{\text{approx}} = A + \epsilon_{\text{ang}}(M) + \epsilon_{\text{init}}(\Delta\tau), \tag{1}$$

where M is the order of angular approximation and $\Delta\tau$ is the initial-layer optical depth used in the doubling method. Ignoring round-off error, $\epsilon_{\text{ang}} \to 0$ as $M \to \infty$ and $\epsilon_{\text{init}} \to 0$ as $\Delta\tau \to 0$.

Wiscombe[1] extensively analyzed ϵ_{init}, both as it varied with $\Delta\tau$, and with the initial-layer formulas. The objective was to make ϵ_{init} as small as possible, and also to ensure flux conservation. Only three out of the five initializations examined conserved flux and, of these, the expanded diamond (EDI) seemed to offer no particular advantage over its progenitor, the diamond (DI); therefore we restrict our study in this paper to the diamond and infinitesimal generator (IGI) initializations.

The angular error, ϵ_{ang}, was ignored in Ref. (1). The present paper attempts to remedy that omission and in the process we shall arrive at some unexpected conclusions about doubling initialization.

ANGULAR TREATMENT

The angular part of our calculation follows the recently-proposed "δ-M method" [Wiscombe[2]], rather than the renormalization approach outlined in Ref. (1). The δ-M method wraps phase function truncation and renormalization into a single procedure, and replaces the actual phase function by an approximate one having the same first 2M moments. An angular discretization $0 < \mu_1 < \cdots < \mu_M < 1$ is used, where μ is cosine of zenith angle, and μ_i are Gaussian quadrature points for the interval [0, 1]. The supplementary angles $-\mu_i$ are used on [−1, 0], for a total of 2M "streams". All other features of the doubling calculation are as described in Ref. (1).

Note that no truncation was used on the phase functions in Ref. [1], which was justified: (a) by the caveat that angular error was not being considered (truncation is used to reduce angular error); and (b) by the fact that we considered only Henyey–Greenstein phase functions, for which then-known truncation procedures seemed ill-adapted. [With the δ-M method, on the other hand, truncation becomes a formality, equally applicable to *all* phase functions.]

AN EXAMPLE

We move immediately to a particular example of the phenomena we wish to point out. Consider a homogeneous non-absorbing layer having the Mie phase function shown in Fig. 1. This phase function is typical of a water cloud in the visible spectrum; the relevant details are

†The National Center for Atmospheric Research is sponsored by the National Science Foundation.

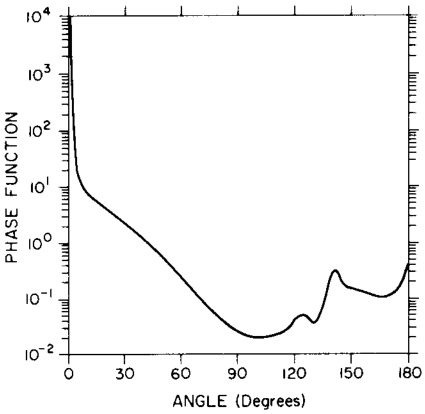

Fig. 1. Mie phase function vs angle for a polydispersion, of effective radius 10μ, of water drops having index of refraction 1.335 at wavelength 0.5μ. The gamma distribution of sizes was integrated between radii 0.1 and 40μ using steps 0.1 in Mie size parameter.

given in the figure caption. There is no upwelling flux at the bottom boundary of the layer. We calculate the albedo A_{approx} of the layer using doubling with the δ-M method for $M = 2, 4, 8$; for layer optical depths $\tau = 0.1, 1, 10$; and for incident beam zenith angle cosines $\mu_0 = 0.1, 0.5, 1.0$.

Four different initializations are used, in order to vary ϵ_{init} over several orders of magnitude. Two are the DI with $\Delta\tau = \mu_1$ and $\Delta\tau = \mu_1/10$, respectively; and two are the IGI with $\Delta\tau = \mu_1$ and $\Delta\tau = \mu_1/100$, respectively. For each formulation (DI or IGI), the second value of $\Delta\tau$ reduces ϵ_{init} a hundredfold compared to the first; this is because ϵ_{init} is linear in $\Delta\tau$ for the IGI, and quadratic in $\Delta\tau$ for the DI.

Total albedo errors,

$$\epsilon_{\text{tot}} = \epsilon_{\text{ang}} + \epsilon_{\text{init}} = A_{\text{approx}} - A, \tag{2}$$

for this calculation are shown in Table 1. This table appears rather formidable, but with it one can make numerous cross-comparisons, which would be rendered much more difficult were the

Table 1. Albedo error using the δ-M method for $M = 2, 4, 8$ and doubling with various initializations, for a homogeneous non-absorbing layer having the Mie phase function of Fig. 1. For each entry of the form $\begin{pmatrix}\Delta_1/\Delta_2\\ \Delta_3/\Delta_4\end{pmatrix}$, Δ_1 and Δ_2 are for the IGI with $\Delta\tau = \mu_1$ and $\Delta\tau = \mu_1/100$, respectively; while Δ_3 and Δ_4 are for the DI with $\Delta\tau = \mu_1$ and $\Delta\tau = \mu_1/10$, respectively. Errors below $1E$-4 for $\mu_0 = 0.1$ and below $1E$-6 for $\mu_0 = 0.5$, 1 are written as zero

		μ_0		
τ	M	0.1	0.5	1.0
0.1	2	-3E-2/-3E-2 *-3E-2/-3E-2*	7E-4/1E-3 *1E-3/1E-3*	-4E-4/-3E-4 *-3E-4/-3E-4*
	4	-6E-3/-3E-3 *-5E-3/-3E-3*	-1E-3/1E-6 *-4E-5/6E-6*	-2E-4/-3E-5 *-3E-5/-3E-5*
	8	4E-4/8E-4 *7E-4/8E-4*	-2E-4/2E-5 *2E-5/2E-5*	-3E-5/1E-5 *1E-5/1E-5*
1	2	-6E-3/-3E-3 *-2E-2/-3E-3*	-5E-3/6E-3 *4E-3/6E-3*	-8E-3/-2E-3 *-2E-3/-2E-3*
	4	4E-3/8E-4 *-1E-3/7E-4*	-3E-3/5E-5 *-8E-5/7E-5*	-2E-3/-1E-4 *-1E-4/-1E-4*
	8	2E-3/7E-4 *5E-4/7E-4*	-8E-4/1E-6 *-4E-6/6E-6*	-4E-4/3E-6 *5E-6/6E-6*
10	2	3E-3/-2E-3 *-5E-3/-2E-3*	-4E-4/-2E-4 *-6E-4/-2E-4*	-2E-4/2E-3 *2E-3/2E-3*
	4	2E-3/0 *-3E-4/0*	7E-6/-7E-5 *-1E-4/-7E-5*	-6E-4/-2E-5 *-2E-5/-1E-5*
	8	9E-4/3E-4 *2E-4/3E-4*	3E-5/0 *-3E-6/-1E-6*	-1E-4/3E-6 *4E-6/5E-6*

numbers to be dissected into separate tables. "Exact" albedos A, good to four decimal places for $\mu_0 = 0.1$ and at least six for $\mu_0 = 0.5$, 1.0, are deduced from an $M = 60$ solution, as described in Ref. (2).

Observe first that, while ϵ_{init} *falls* one hundred fold, ϵ_{tot} actually *rises or stays the same* in 16 out of the 27 cases for the DI, and 6 out of the 27 cases for the IGI. Keeping $\Delta\tau$ fixed at μ_1, but going from the IGI to the more accurate DI, reduces ϵ_{tot} in the majority (19) but, *by no means, all* of the cases.

Consider now just the DI. In reducing $\Delta\tau$ from μ_1 to $\mu_1/10$, ϵ_{tot} changes less than a factor of three (excepting only 3 cases), even though ϵ_{init} falls a factor of 100. It is easy to deduce from this fact, plus eqn (2), that ϵ_{ang} and ϵ_{init} differ in magnitude by no more than a factor of four when $\Delta\tau = \mu_1$. Therefore, it seems pointless and indeed wasteful of computation to take $\Delta\tau$ significantly smaller than μ_1, since doing so only reduces ϵ_{init} and can, at best, effect only a moderate reduction in ϵ_{tot}. Such a reduction of ϵ_{init} can sometimes even raise ϵ_{tot}, in the circumstance that ϵ_{init} and ϵ_{ang} have opposite signs when $\Delta\tau = \mu_1$ and partially cancel one another. Table 1 contains several examples of the latter phenomenon.

Evidence for larger values of M reinforces the observations of the last paragraph. For τ small (≤ 1, say), numerical experiments at $M = 50$ showed that using $\Delta\tau \ll \mu_1$ rather than $\Delta\tau = \mu_1$ changed albedos, etc. in at most the 7th or 8th decimal place (of course μ_1 is quite small at $M = 50$). Second, for large τ (≥ 10, say), using $\Delta\tau \ll \mu_1$ introduces serious round-off error accumulation, causing, among other things, large spurious absorptions. For $\tau = 100$ and $M = 50$, round-off usually affected the 4th or 5th decimal places of albedos and absorptivities for $\Delta\tau = 0.02\mu_1$; it retreated to the 6th or 7th places upon increasing $\Delta\tau$ to μ_1; the computation carried 14 significant digits. In sum, it seems unwise as well as wasteful to take $\Delta\tau < \mu_1$ when using the DI, whatever the value of M.

Is the DI still preferable at all, in light of the above findings? We still believe so, but the case for it is not nearly so strong as an examination of ϵ_{init} alone indicates. We have already seen in about 1/3 of the cases in Table 1 that the IGI may actually lead to *better* accuracy when $\Delta\tau = \mu_1$. Furthermore, Table 1 shows that the IGI with $\Delta\tau = \mu_1/100$ gives errors nearly identical to the DI with $\Delta\tau = \mu_1/10$ (with only 4 exceptions); both initializations have at this point caused $|\epsilon_{\text{init}}| \ll |\epsilon_{\text{ang}}|$ and are therefore equally preferable. But an optimal initialization would use as large a $\Delta\tau$ as possible, in order to minimize computation and round-off. At the same time, this large value of $\Delta\tau$ should cause the two errors to be comparable in magnitude, $|\epsilon_{\text{init}}| \sim |\epsilon_{\text{ang}}|$, by the following reasoning. Suppose, in fact, $|\epsilon_{\text{init}}| = |\epsilon_{\text{ang}}|$. Then what can we gain by reducing either error? If both have the same sign, at most a factor of two reduction in ϵ_{tot} is achieved; if opposite signs, ϵ_{tot} is increased from zero up to, at most, twice the error which was not reduced. Since no clear advantage accrues to reducing either error, and since it requires extra computation to do so, we are best off where we started, that is, with $|\epsilon_{\text{init}}| = |\epsilon_{\text{ang}}|$. We cannot of course hope to achieve equality, but it is still desirable, by the same arguments, to strive for comparability, $|\epsilon_{\text{init}}| \sim |\epsilon_{\text{ang}}|$.

The DI generally meets the criteria laid out in the last paragraph better than the IGI. Indeed, $|\epsilon_{\text{init}}| \gg |\epsilon_{\text{ang}}|$ in general for the IGI with $\Delta\tau = \mu_1$, as indicated by the frequently large factors (especially for $M = 4$, 8) by which ϵ_{tot} plummets when ϵ_{init} is reduced a hundredfold. The superiority of the DI is manifest particularly for larger values of M; at $M = 60$, for the same situation as in Table 1, and with $\Delta\tau = \mu_1$, DI albedos, etc. are fully two decimal places more accurate than IGI ones. Thus the DI seems to be the best all-around initialization, an "initialization for all seasons", as it were.

SUMMARY

The conclusions of this paper have been drawn in the context of a specific example; but while the example is specific, the conclusions are general. We have observed similar phenomena in a wide variety of cases.

The basic conclusion is that total error, eqn (2), need not be reduced when initial-layer error is reduced, especially when working at the low orders of angular approximation (M = 2–16, say) that many investigators would prefer to use. Effort spent to reduce initial-layer error, unless it significantly exceeds angular error in magnitude, is largely wasted. Indeed, initial-layer and angular error sometimes partially cancel one another in eqn (2), leading to an *increase* in total error when either one alone is decreased.

The most desirable state of affairs, in view of all this, is to tie angular and initial-layer error together, in such a way that they remain roughly comparable, $|\epsilon_{\text{init}}| \sim |\epsilon_{\text{ang}}|$. This seems best accomplished, from the numerous examples we have examined, by using the diamond initialization with $\Delta\tau \sim \mu_1$.

REFERENCES

1. W. J. WISCOMBE, *JQSRT* **16**, 637 (1976).
2. W. J. WISCOMBE, The delta-M method: rapid yet accurate radiative flux computations for strongly asymmetric phase functions. *J. Atmos. Sci.* (submitted Jan. 1977).

J. Quant. Spectrosc. Radiat. Transfer, Vol. 16, pp. 637–658. Pergamon Press 1976. Printed in Great Britain

ON INITIALIZATION, ERROR AND FLUX CONSERVATION IN THE DOUBLING METHOD

W. J. Wiscombe

National Center for Atmospheric Research,† Boulder, CO 80303, U.S.A.

(*Received* 28 *July* 1975)

Abstract—Truncation errors and flux conservation errors in the doubling method are examined. The error properties of five different initial-layer approximations are compared as a function of initial-layer size, layer optical depth, single-scattering albedo, and phase function asymmetry parameter. The "diamond" initial-layer approximation is found to be orders of magnitude more accurate than the others for fixed initial-layer size, or of equivalent accuracy starting from a very much larger initial layer. The commonly used single-scattering initialization is shown to lead to serious flux conservation errors. Analytic error estimates, based upon a new derivation of the single-scattering initialization directly from doubling, are shown to be useful when the layer optical depth is on the order of 10 or less. Finally, questions of round-off error, calculation of an "exact" answer using Richardson extrapolation, and computational efficiency are all addressed briefly.

INTRODUCTION

The doubling method has become a mainstay for accurate radiative transfer calculations for plane-parallel, horizontally-homogeneous, absorbing-scattering atmospheres. For a layer of arbitrary optical depth and arbitrary (but constant) single-scattering albedo and phase function, the method provides the internal sources, the reflectivity, and the transmissivity. By decreasing the initial-layer size and increasing the number of discrete angles at which these quantities are calculated, fluxes and intensities may be computed to any desired accuracy (limited only by round-off error). Doubling was proposed in its present form by Van de Hulst[(1)] and independently by Twomey *et al.*[(2)] and has been applied to atmospheric problems by Hansen[(3–6)] and Grant and Hunt[(7)] among others. A historical review and more extensive references are given by Plass *et al.*[(8)]

In spite of the broad acceptance of doubling for benchmark radiative transfer calculations, very little discussion of its flux conservation properties and of its error, as a function of initial-layer approximation and of parameters such as the single-scattering albedo, has appeared. Those who apply the method generally obtain at least plotting accuracy (1% or so) by experimenting with the initial-layer size and angular resolution; only Van de Hulst and Grossman[(9)] have gone beyond this, and their comments are brief and fall far short of a thorough error analysis. As for flux conservation, Grant and Hunt[(10,11)] have provided a nice conceptual framework within which to discuss the problem, but they show very few actual examples. Finally, there are a variety of ways to initialize the doubling process (see Section 3), and it is not clear which, if any, is superior.

One might well ask, of course, if error analysis and initialization are not rather academic concerns. After all, by appropriately reducing the initial-layer size and re-computing results it is possible to experimentally confirm whether one's solution has converged to the desired number of significant digits (if round-off can be neglected). For many simple applications this is undoubtedly sufficient. But there are a growing number of more sophisticated applications, such as in the Grant–Hunt[(10)] method for vertically inhomogeneous atmospheres, where doubling is only a component in a larger scheme. The author (Wiscombe and Freeman;[(12)] Wiscombe[(13)]) has described a radiative transfer calculation for the entire solar and i.r. spectrums which requires roughly 100 spectral intervals, within each of which the Grant–Hunt method may be called repeatedly based on the number of terms in an exponential-sum fit to the transmission. Thousands of doubling results, for widely-varying layer sizes, single-scattering albedos, and phase functions, are required in the course of such a calculation. It is time-consuming and wasteful to have to repeat such massive calculations with smaller initial-layer sizes to verify levels of

†The National Center for Atmospheric Research is sponsored by the National Science Foundation.

significance, and *a priori* error estimates such as are given here relieve much of this burden. Furthermore they make possible potentially large savings of computer time, since lacking such estimates one must generally be over-conservative (no more striking example of this exists that HANSEN's[3] use of initial layers of thickness 2^{-25}, which the present study shows to be orders of magnitude smaller than necessary).

In the following sections, we shall study error and flux conservation in the doubling method for five initial-layer approximations (hereinafter called simply "initializations") as a function of (1) initial-layer size (2) layer optical depth (3) single-scattering albedo and (4) phase function asymmetry parameter. Other new results are: a derivation of the single-scattering approximation directly from doubling; analytic error estimates which are useful when the layer optical depth does not exceed 10; and methods for avoiding round-off error when doubling results of high accuracy are desired. Some computation-saving short cuts are also given in Appendix A.

The question of error due to angular discretization is not specifically addressed here (although we do examine errors for different numbers of angles). Within a particular angular discretization we shall only be concerned with error due to the use of finite initial-layer sizes. (HUNT[14] has made some empirical studies of angular error.)

2. THE DOUBLING METHOD

In this section we give the doubling formulas and phase function used in the calculations and show how an "exact" solution, necessary for computing errors, was obtained.

Our interest shall center on the $m \times m$ reflection and transmission matrices

$$r_n \equiv r(2^n \Delta\tau), \quad t_n \equiv t(2^n \Delta\tau),$$

where $\Delta\tau$ is the initial-layer size, $\tau = 2^N \Delta\tau$ will be the total layer optical depth, and where the goal is to double from r_0, t_0 (derived from some initialization) to r_N, t_N. The quantities $r(\tau)$ and $t(\tau)$ give, when operating on angularly-discretized vectors of incident intensity, the reflected and transmitted intensity vectors respectively. These relationships are summarized in the "interaction principle" (GRANT and HUNT[10]), which when specialized to a layer across which the single-scattering albedo and phase function are constant, is

$$i^+(\tau) = r(\tau)i^-(\tau) + t(\tau)i^+(0) + \sum{}^+ (0, \tau), \tag{1a}$$

$$i^-(0) = r(\tau)i^+(0) + t(\tau)i^-(\tau) + \sum{}^- (0, \tau), \tag{1b}$$

where

$$i^\pm(\hat{\tau}) \equiv \begin{bmatrix} i(\hat{\tau}, \pm\mu_1 \\ \vdots \\ i(\hat{\tau}, \pm\mu_m) \end{bmatrix}, \tag{2}$$

and where $i(\hat{\tau}, \mu)$ is the azimuthally-averaged intensity (we ignore azimuthal dependence here since our interest centers on fluxes). The angular discretization is $0 < \mu_1 < \cdots < \mu_m \leq 1$, where $\mu = \cos\theta$ and θ is the angle measured from the normal to the layer. The internal sources $\Sigma^\pm$ are discussed in a companion paper (WISCOMBE[15]). For those unfamiliar with the interaction principle, Fig. 1 of GRANT and HUNT[10] provides a nice schematic interpretation of eqn (1).

The doubling formulas may be derived directly from collapsing eqns (1) and their analogue for the layer $[\tau, 2\tau]$ into a single interaction principle, or alternatively by specializing the layer addition formulas in GRANT and HUNT,[10] to yield

$$t_{n+1} = t_n(I - r_n r_n)^{-1} t_n, \tag{3a}$$

$$r_{n+1} = r_n + t_n(I - r_n r_n)^{-1} r_n t_n. \tag{3b}$$

Equations (3) were iterated from $n = 0$ to $n = N - 1$ to obtain r_N and t_N. All computations were done in 29-significant-digit arithmetic unless otherwise stated, in order to minimize round-off error.

Appendix A gives a technique which greatly reduces the computational burden of the doubling formulas (3).

For simplicity, and because it simulates the effect of an arbitrary phase function quite well (HANSEN[4]), the Henyey–Greenstein phase function

$$P(\mu)\equiv\frac{1-g^2}{(1+g^2-2g\mu)^{3/2}}$$

was employed in this study. The parameter g is identical to the asymmetry factor $\langle\cos\theta\rangle$ for this phase function. We actually require the azimuthally-averaged form of P, viz.

$$\begin{aligned}\bar{P}(\mu_i,\mu_j)&\equiv\frac{1}{\pi}\int_0^{\pi}P(\mu_i\mu_j+\sqrt{1-\mu_i^2}\sqrt{1-\mu_j^2}\cos\phi)\,\mathrm{d}\phi\\&=\frac{2(1-g^2)}{\pi(a+b)^{3/2}}\frac{E\left(\sqrt{\dfrac{2b}{a+b}}\right)}{1-\dfrac{2b}{a+b}},\end{aligned}\tag{4}$$

where E is the complete elliptic integral of the second kind and

$$a\equiv1+g^2-2g\mu_i\mu_j,\quad b\equiv2g\sqrt{1-\mu_i^2}\sqrt{1-\mu_j^2}.$$

This form for $\bar{P}$ has not been previously noted in the literature to the author's knowledge; it is useful in that fast standard subroutines are available for elliptic integrals. The values of $\bar{P}$ from eqn (4) were re-normalized with Grant's method (see Section 6).

The "exact" reflection and transmission matrices, defined as

$$r_\infty,t_\infty=\lim_{\substack{\Delta\tau\to0\\ \tau\text{ fixed}}}r_N,t_N,\tag{5}$$

are necessary in order to perform an error analysis. Letting $\Delta\tau\to0$ on the computer, however, is not possible in principle because of round-off error growth. And because we set ourselves the goal of obtaining r_∞ and t_∞ correct to at least 8 significant digits, it was generally impossible in practice as well. However, we found a way to sidestep the round-off problem. By using element-by-element Richardson extrapolation (see the excellent review of JOYCE[16]) of r_N's and t_N's computed from successively halved initial-layer sizes,

$$(\Delta\tau)_k=(\Delta\tau)_0/2^k\quad(k=0,1,2,\ldots),$$

it proved possible to compute r_∞ and t_∞ to the desired accuracy, provided τ was not too large. For large τ the extrapolation tended to lose utility because the higher-order terms in the expansion $r_N=r_\infty+a\Delta\tau+b(\Delta\tau)^2+\cdots$ were no longer small compared to the low-order terms which were being successively eliminated (some insight into the reason for this behavior is furnished in Section 7). The value at which τ became "large" depended crucially on the initialization; for the diamond initialization (Section 3b) it was as small as 1–2, while for the less accurate infinitesimal generator initialization (Section 3d) it was at least 100 (the largest value of τ considered). But for large τ we availed ourselves of the phenomenal accuracy of the diamond initialization (Section 7); our desired level of significance was easily obtained using this initialization with a moderately small $\Delta\tau$.

Since r_∞ and t_∞ are crucial to our analysis, considerable effort was expended to make sure they were accurate. The error computations were performed with varying $(\Delta\tau)_0$ and using extrapolation with the diamond, expanded diamond (Section 3c), and infinitesimal generator initializations. In no case were the error plots, presented below, discernibly changed. It is interesting to note, however, that extrapolation "converged" (ceased to decrease the error) more quickly for the more accurate initializations.

3. INITIALIZATIONS

Here we review a general theory from which any number of specific initializations may be derived. Five initializations are singled out for further study.

(a) *General theory*

Let us write the azimuthally-averaged monochromatic radiative transfer equation in the form

$$\mu\frac{\partial i}{\partial\tau}+i=\sigma(\tau,\mu)+\frac{\omega}{2}\int_{-1}^{1}\bar{P}(\mu,\mu')i(\tau,\mu')\,\mathrm{d}\mu', \tag{6}$$

where σ is the source term, ω the single-scattering albedo and $\bar{P}$ the azimuthally-averaged phase function from eqn (4). Replace the integral using a quadrature formula for [0, 1] with points $0<\mu_1<\cdots<\mu_m\le 1$ and corresponding weights c_i, and set $\mu=\mu_i$ to yield

$$\mu_i\frac{\partial i(\tau,\mu_i)}{\partial\tau}+i(\tau,\mu_i)=\sigma(\tau,\mu_i)+\frac{\omega}{2}\sum_{j=1}^{m}c_j\{p_{ij}^{+-}i(\tau,-\mu_j)+p_{ij}^{++}i(\tau,\mu_j)\},$$

where

$$p^{+\pm}=[\bar{P}(\mu_i,\pm\mu_j)].$$

An analogous equation follows from putting $\mu=-\mu_i$ rather than $\mu=\mu_i$. The two equations may immediately be written in matrix-vector form as

$$\pm M\frac{\partial i^{\pm}}{\partial\tau}+i^{\pm}=\sigma^{\pm}(\tau)+\frac{\omega}{2}[p^{++}ci^{\pm}+p^{+-}ci^{\mp}], \tag{7}$$

where $i^{\pm}$ are as defined in eqn (2) and

$$\sigma^{\pm}(\tau)=\begin{bmatrix}\sigma(\tau,\pm\mu_1)\\ \vdots\\ \sigma(\tau,\pm\mu_m)\end{bmatrix},\qquad M=[\mu_i\delta_{ij}],\quad c=[c_i\delta_{ij}].$$

Now integrate eqn (7) across a thin layer $[\tau_0,\tau_1]$ of thickness $\Delta\tau$ to yield

$$\pm M(i_1^{\pm}-i_0^{\pm})+i_{1/2}^{\pm}\Delta\tau=\sigma_{1/2}^{\pm}\Delta\tau+\frac{\omega}{2}\Delta\tau[p^{++}ci_{1/2}^{\pm}+p^{+-}ci_{1/2}^{\mp}]. \tag{8}$$

The quantities $i_0^{\pm}$ and $i_1^{\pm}$ are simply $i^{\pm}$ evaluated at $\tau=\tau_0$ and $\tau=\tau_1$, respectively. The "central intensities" $i_{1/2}^{\pm}$ and "central sources" $\sigma_{1/2}^{\pm}$ are defined as

$$i_{1/2}^{\pm}\equiv\frac{1}{\Delta\tau}\int_{\tau_0}^{\tau_1}i^{\pm}\,\mathrm{d}\tau,\quad \sigma_{1/2}^{\pm}\equiv\frac{1}{\Delta\tau}\int_{\tau_0}^{\tau_1}\sigma^{\pm}\,\mathrm{d}\tau. \tag{9}$$

A variety of different thin-layer approximations can be obtained by assuming that the central intensities are related linearly to the two boundary intensities. The most general such assumption which does not mix up intensities along different quadrature directions μ_i is

$$i_{1/2}^{+}=X^{+}i_0^{+}+(I-X^{+})i_1^{+}, \tag{10a}$$

$$i_{1/2}^{-}=X^{-}i_1^{-}+(I-X^{-})i_0^{-}, \tag{10b}$$

where X^{+} and X^{-} are arbitrary diagonal matrices. Upon inserting these relations into eqn (8), it is possible to manipulate eqn (8) into interaction principle form [eqns (1)] and thereby identify expressions for r, t and $\Sigma^{\pm}$.

(*b*) *Diamond initialization* (*DI*)

A natural choice for the $X^{\pm}$ in eqn (10) is

$$X^{+}=X^{-}=\frac{1}{2}I$$

since this is equivalent to assuming that the intensity is linear in optical depth. Carlson[17] coined the name "diamond scheme" for this assumption in an entirely different radiative transfer framework, and we shall retain that usage.

Putting the diamond scheme into eqn (8) leads to

$$(I+T)i_1^{+} = (I-T)i_0^{+} + R(i_0^{-} + i_1^{-}) + \hat{\sigma}^{+}_{1/2}, \tag{11a}$$

$$(I+T)i_0^{-} = (I-T)i_1^{-} + R(i_0^{+} + i_1^{+}) + \hat{\sigma}^{-}_{1/2}, \tag{11b}$$

where

$$T \equiv \hat{T}\frac{\Delta\tau}{2}, \quad \hat{T} = M^{-1}\left(I - \frac{\omega}{2}p^{++}c\right) \tag{11c}$$

$$R \equiv \hat{R}\frac{\Delta\tau}{2}, \quad \hat{R} = \frac{\omega}{2}M^{-1}p^{+-}c, \tag{11d}$$

$$\hat{\sigma}^{\pm}_{1/2} \equiv M^{-1}\sigma^{\pm}_{1/2}\Delta\tau. \tag{11e}$$

Equation (11a) would be in interaction principle form [eqn (1a)] were it not for the i_0^{-} term. But eqn (11b) can be solved for i_0^{-} in terms of the other intensities, and substituted into eqn (11a) to eliminate i_0^{-}. After further manipulation we may then identify the reflection and transmission matrices and source vectors in the interaction principle as follows:

$$r(\Delta\tau) = 2\Gamma R(I+T)^{-1}, \tag{12a}$$

$$t(\Delta\tau) = 2\Gamma - I, \tag{12b}$$

$$\sum{}^{\pm}(\Delta\tau) = \Gamma\{\hat{\sigma}^{\pm}_{1/2} + R(I+T)^{-1}\hat{\sigma}^{\mp}_{1/2}\}, \tag{12c}$$

where

$$\Gamma \equiv [I + T - R(I+T)^{-1}R]^{-1}.$$

The identities

$$2(I+T)^{-1} = I + (I+T)^{-1}(I-T), \quad 2I - \Gamma^{-1} = I - T + R(I+T)^{-1}R,$$

have been used to simplify the equations for $r(\Delta\tau)$ and $t(\Delta t)$, respectively.

The actual computation of eqns (12) is facilitated by using the techniques in Appendix A.

Equations which are reducible to eqns (12) are buried in an appendix of Grant and Hunt.[18] However, it is not clear from the context if they used these equations to initialize doubling, and it is certain that no one else has.

(c) *Expanded diamond initialization (EDI)*

When preliminary numerical experiments showed the DI to have accuracy $O[(\Delta\tau)^2]$, it seemed logical to avoid the matrix inversions in eqns (12) by expanding in powers of $\Delta\tau$ out to $(\Delta\tau)^2$. The truncated expansions so derived are:

$$r(\Delta\tau) = \hat{R}\Delta\tau - (\hat{R}\hat{T} + \hat{T}\hat{R})\frac{(\Delta\tau)^2}{2}, \tag{13a}$$

$$t(\Delta\tau) = I - \hat{T}\Delta\tau + (\hat{T}^2 + \hat{R}^2)\frac{(\Delta\tau)^2}{2}. \tag{13b}$$

$\hat{R}$ and $\hat{T}$ are defined in eqns (11c, d). The source vector expansions are not given here since they depend on the explicit form of the source function $\sigma(\tau, \mu)$. Equations (13) will be referred to as the "expanded diamond" initialization.

(d) *Infinitesimal generator initialization (IGI)*

The simplest and most basic initialization can be derived either by taking $X^{+} = X^{-} = I$ in eqns (10), whereby eqns (8) reduce immediately to interaction principle form, or by expanding the DI to O($\Delta\tau$) only. The result is

$$r(\Delta\tau) = \hat{R}\Delta\tau, \quad t(\Delta\tau) = I - \hat{T}\Delta\tau, \quad \sum^{\pm}(\Delta\tau) = \hat{\sigma}^{\pm}_{1/2}, \tag{14}$$

where $\hat{R}$, $\hat{T}$, and $\hat{\sigma}$ are defined in eqns (11c–e). This initialization is probably the one used most frequently (e.g. Twomey *et al.*[2]) Only Grant and Hunt[10] have given it a name, the "infinitesimal generator" (from group theory), and to avoid a proliferation of nomenclature we will adopt that usage here.

(e) *Single-scattering initialization (SSI)*

The single-scattering approximation in radiative transfer is well known (Chandrasekhar[19]). It corresponds to solving the *integral* form of the radiative transfer equation to first order in ω. It does not, however, proceed from eqns (8) and (10) for any choice of the weights $X^{\pm}$, for the simple reason that eqns (8) are extracted from the *integrodifferential* form (6) of the transfer equation.

Nevertheless, it is possible to derive the single-scattering approximation from nothing more than the IGI [eqns (14)] and the doubling formulas (3). We rewrite eqns (14) in order to explicitly display their ω-dependence as

$$r_0 = \omega\hat{r}, \quad t_0 = \hat{e} + \omega\hat{t},$$

where

$$\hat{r} \equiv \frac{1}{2}M^{-1}p^{+-}c\Delta\tau, \quad \hat{t} \equiv \frac{1}{2}M^{-1}p^{++}c\Delta\tau, \quad \hat{e} \equiv I - M^{-1}\Delta\tau. \tag{15}$$

Clearly r_0 is O(ω), and therefore from the doubling formula (3b), r_n will be O(ω) for all n. Hence to O(ω) the matrix inverse in both doubling formulas can be replaced by the identity. Performing the first two doublings explicitly and keeping only O(ω) terms,

$$\begin{aligned} r_1 &= r_0 + t_0 r_0 t_0 = \omega(\hat{r} + \hat{e}\hat{r}\hat{e}), \\ t_1 &= t_0^2 = \hat{e}^2 + \omega(\hat{t}\hat{e} + \hat{e}\hat{t}), \\ r_2 &= r_1 + t_1 r_1 t_1 = \omega[\hat{r} + \hat{e}\hat{r}\hat{e} + \hat{e}^2\hat{r}\hat{e}^2 + \hat{e}^3\hat{r}\hat{e}^3], \\ t_2 &= t_1^2 = \hat{e}^4 + \omega(\hat{t}\hat{e}^3 + \hat{e}\hat{t}\hat{e}^2 + \hat{e}^2\hat{t}\hat{e} + \hat{e}^3\hat{t}). \end{aligned}$$

The emerging pattern is clear, and it can be proved by induction that

$$r_n = \omega\sum_{k=0}^{2^n-1} \hat{e}^k\hat{r}\hat{e}^k, \quad t_n = \hat{e}^{2^n} + \sum_{k=0}^{2^n-1} \hat{e}^k\hat{t}\hat{e}^{2^n-k-1}.$$

Because $\hat{e}$ is diagonal [eqn (15)], these reduce to geometric series which can be summed in closed form. For r_n we have

$$(r_n)_{ij} = \omega\hat{r}_{ij}\sum_{k=0}^{2^n-1}(e_ie_j)^k = \omega\hat{r}_{ij}[1-(e_ie_j)^{2^n}]/(1-e_ie_j), \tag{16}$$

where e_i is the ith diagonal element of $\hat{e}$:

$$e_i \equiv 1 - (\Delta\tau/\mu_i).$$

For t_n we have similarly

$$\begin{aligned} (t_n)_{ij} &= e_i^{2^n}\delta_{ij} + \omega\hat{t}_{ij}e_j^{2^n-1}\sum_{k=0}^{2^n-1}(e_i/e_j)^k \\ &= \begin{cases} e_i^{2^n} + 2^n\omega\hat{t}_{ii}e_i^{2^n-1} \\ \omega\hat{t}_{ij}(e_j^{2^n} - e_i^{2^n})/(e_j - e_i) & i \neq j. \end{cases} \end{aligned} \tag{17}$$

Let us now take the limit as $\Delta\tau \to 0$ of these results, holding the layer size $\hat{\tau} = 2^n \Delta\tau$ fixed. Since

$$\lim_{\Delta\tau\to 0} e_i^{2^n} = \lim_{n\to\infty}\left(1 - \frac{\hat{\tau}/\mu_i}{2^n}\right)^{2^n} = e^{-\hat{\tau}/\mu_i},$$

it is clear that the limiting forms of eqns (16) and (17) are

$$(r_\infty)_{ij} = \frac{\omega}{2}\frac{c_j}{\mu_i} p_{ij}^{+-}\,[1 - \mathrm{e}^{-(\hat{\tau}/\mu_i + \hat{\tau}/\mu_j)}]\Big/\left(\frac{1}{\mu_i} + \frac{1}{\mu_j}\right) \tag{18a}$$

$$(t_\infty)_{ij} = \begin{cases} \mathrm{e}^{-\hat{\tau}/\mu_i}\left(1 + \dfrac{\omega}{2}\dfrac{c_i}{\mu_i} p_{ii}^{++}\hat{\tau}\right) & i = j \\[2ex] \dfrac{\omega}{2}\dfrac{c_j}{\mu_i} p_{ij}^{++}(\mathrm{e}^{-\hat{\tau}/\mu_j} - \mathrm{e}^{-\hat{\tau}/\mu_i})\Big/\left(\dfrac{1}{\mu_i} - \dfrac{1}{\mu_j}\right) & i \neq j. \end{cases} \tag{18b}$$

With $\hat{\tau} = \Delta\tau$, eqns (18) become the single-scattering initialization. It has been used most prominently by HANSEN.[3,5]

(f) *Plass initialization* (*PI*)

This initialization appears in the appendix of a paper by PLASS *et al.*,[8] where it is mistakenly associated with Hansen, who in fact uses the SSI. It is a slight variant on the IGI, in which it is recognized in eqn (14) that when $\omega = 0$ the IGI transmission

$$t(\Delta\tau) = I - M^{-1}\Delta\tau = \left[\left(1 - \frac{\Delta\tau}{\mu_i}\right)\delta_{ij}\right],$$

is merely a two-term expansion of the correct transmission

$$E(\Delta\tau) = [\mathrm{e}^{-\Delta\tau/\mu_i}\,\delta_{ij}].$$

Therefore, the transmission is initialized by

$$t(\Delta\tau) = E(\Delta\tau) + \frac{1}{2}\,\omega M^{-1} p^{++} c\,\Delta\tau, \tag{19}$$

which is then exact at $\omega = 0$. The remainder of eqn (14) is used as it stands.

Equation (19) is inconsistent in that an incipient power series in $\Delta\tau$ has been summed explicitly in only one part of eqn (14). The price of this inconsistency will become clear in Sections 5 and 6.

(g) *Restrictions on* $\Delta\tau$

GRANT and HUNT[10,11] thoroughly discuss the importance of having all elements of the reflection and transmission matrices non-negative, which one should expect on physical grounds. They show that the property of non-negativity is preserved by doubling, so that it is sufficient to have all elements of r_0 and t_0 non-negative. These matrices are *always* non-negative for the SSI and PI, for all $\Delta\tau > 0$. But for the other three initializations, non-negativity imposes an upper limit on $\Delta\tau$.

For the IGI, this upper limit is

$$(\Delta\tau)_{\max} \equiv \min_i \left(\frac{\mu_i}{1 - \frac{1}{2}\omega p_{ii}^{++} c_i}\right). \tag{20}$$

In the case of the EDI, non-negativity constrains $\Delta\tau$ to satisfy a complicated set of inequalities. For typical small $\Delta\tau$, however, they are only slightly more liberal than $\Delta\tau \leqslant (\Delta\tau)_{\max}$. Therefore these inequalities are omitted.

For the DI, we can show that an upper limit of $2(\Delta\tau)_{\max}$ is sufficient for non-negativity (the proof is given in Appendix B), but are unable to establish the necessity of this bound. Numerical experiments using the DI over wide ranges of parameters and with $\Delta\tau > 2(\Delta\tau)_{\max}$ have however convinced us that (1) $\Delta\tau \leqslant 2(\Delta\tau)_{\max}$ is a necessary as well as a sufficient condition for r_0, $t_0 \geq 0$ (2) computed fluxes may remain accurate even when $\Delta\tau$ exceeds $2(\Delta\tau)_{\max}$ by up to an order of magnitude and (3) initially negative elements (in r_0 and t_0) disappear after only a very few doubling steps. This remarkable recovery ability is not shared by either the IGI or EDI—if they start out negative, they grow more so. The number of doubling steps needed for the DI to recover non-negativity is a function primarily of the extent to which $\Delta\tau$ exceeds $2(\Delta\tau)_{\max}$. For $\Delta\tau/[2(\Delta\tau)_{\max}] = 2$, for example, only one doubling step is required; for $\Delta\tau/[2(\Delta\tau)_{\max}] = 5$, 3 steps. We conclude that it will be safe to use $\Delta\tau > 2(\Delta\tau)_{\max}$ in the DI if one is doubling to fairly large optical depths, but not otherwise.

(h) *Sources*

We do not examine error in doubling for sources here, but we note that this error is unnecessarily magnified when the central source vectors $\sigma^{\pm}_{1/2}$ [eqn (9)] are taken as

$$\sigma^{\pm}_{1/2} = \sigma^{\pm}(\bar{\tau}), \tag{21}$$

where $\bar{\tau}$ is the midpoint of $[\tau_0, \tau_1]$. Both GRANT and HUNT[10] and the author in earlier work (WISCOMBE and FREEMAN[12]) have used this approximation, which is exact only for some unknown $\bar{\tau}$ in $[\tau_0, \tau_1]$, according to the mean value theorem. But when the τ-dependence of $\sigma^{\pm}$ is known analytically, as for example in the direct beam source, which is exponential in τ, then $\sigma^{\pm}_{1/2}$ can be computed exactly. Even if the analytic dependence on τ is not known, it should be possible to perform a more sophisticated numerical quadrature on eqn (9) than a midpoint rule, which leads to eqn (21).

4. NORMS AND FLUX ERRORS

Examining the differences between exact and approximate reflection and transmission matrices element-by-element would be painful and unilluminating. Fortunately, there is a natural matrix norm for radiative transfer which distills this error into a single number. This number can be used immediately to put a bound on the flux error.

First a vector norm is defined, following GRANT and HUNT,[11] so that when the vector is an intensity its norm is the quadrature formula for the flux,

$$\|i\| \equiv \sum_{j=1}^{m} 2\pi c_j \mu_j |i_j|.$$

The corresponding matrix norm is defined in terms of the vector norm as

$$\|A\| = \max_{\|i\|} \frac{\|Ai\|}{\|i\|}, \tag{22a}$$

$$= \max_{j} \sum_{i=1}^{m} c_i \mu_i |A_{ij}| \frac{1}{c_j \mu_j}, \tag{22b}$$

The equivalence of eqns (22a and b) is shown in ISAACSON and KELLER, Ref. (20), p. 9. If A were a reflection matrix, definition (22a) shows that $\|A\|$ would be the maximum possible albedo of a layer; if A were a transmission matrix, $\|A\|$ would be the maximum transmissivity. Thus this norm has simple physical interpretations.

From definition (22a) it is evident that

$$\|Ai\| \leq \|A\| \, \|i\|.$$

Subtracting the interaction principles [eqn (1)] for the exact ($\Delta\tau \to 0$) and approximate (finite $\Delta\tau$) cases, taking norms, and using this property leads to the following error-bounds for fluxes leaving a layer:

$$\|\Delta i_1^+\| \le \|\Delta r\|\,\|i_1^-\| + \|\Delta t\|\,\|i_0^+\| + \|\Delta\Sigma^+\|,$$
$$\|\Delta i_0^-\| \le \|\Delta r\|\,\|i_0^+\| + \|\Delta t\|\,\|i_1^-\| + \|\Delta\Sigma^-\|, \tag{23}$$

where Δ indicates the difference between exact and approximate values ($\Delta r = r_\infty - r_N$, $\Delta t = t_\infty - t_N$, etc.). Thus, barring anomalously high source errors, the error in emergent flux will not exceed max ($\|\Delta r\|$, $\|\Delta t\|$) times the incident flux, roughly speaking. If, for example, we pick $\Delta\tau$ so that $\|\Delta r\|$ and $\|\Delta t\|$ do not exceed 10^{-4}, then errors in the emergent fluxes will not exceed 0·01% of the incident fluxes.

5. COMPARISON OF INITIALIZATIONS

The errors $\|\Delta r\|$ and $\|\Delta t\|$ for the five initializations are now compared.

Figure 1 shows $\|\Delta r\|$ vs initial-layer size $\Delta\tau$ for layers of optical depth $\tau = 0{\cdot}1$, 1 and 10 and single-scattering albedos $\omega = 0{\cdot}1$, 0·5 and 1. Figure 2 shows $\|\Delta t\|$ vs $\Delta\tau$ for the same parameters. Only the asymmetry factor $g = 0{\cdot}9$ is considered since the plots varied in only minor ways for $0 \le g \le 0{\cdot}9$, and all of the important conclusions to be drawn from them were independent of g. Figures 1 and 2 are for $m = 8$ Gaussian quadrature points μ_i for the interval [0, 1] and all five initializations of Section 3(b–f) are represented on each plot (although for two of the $\tau = 10$ cases in Fig. 1 the DI is so accurate that its curve falls entirely below $\|\Delta r\| = 10^{-8}$ and so does not appear).

Perhaps the most striking result of these plots is the phenomenal accuracy of the DI relative to all the others. Only for $\Delta\tau$ near its upper bound $2(\Delta\tau)_{\max}$, discussed in Section 3(g), and for $\tau = 0{\cdot}1$ do $\|\Delta r\|$ and $\|\Delta t\|$ become even as large as 10^{-2}. For the wide ranges of τ, ω, g, and

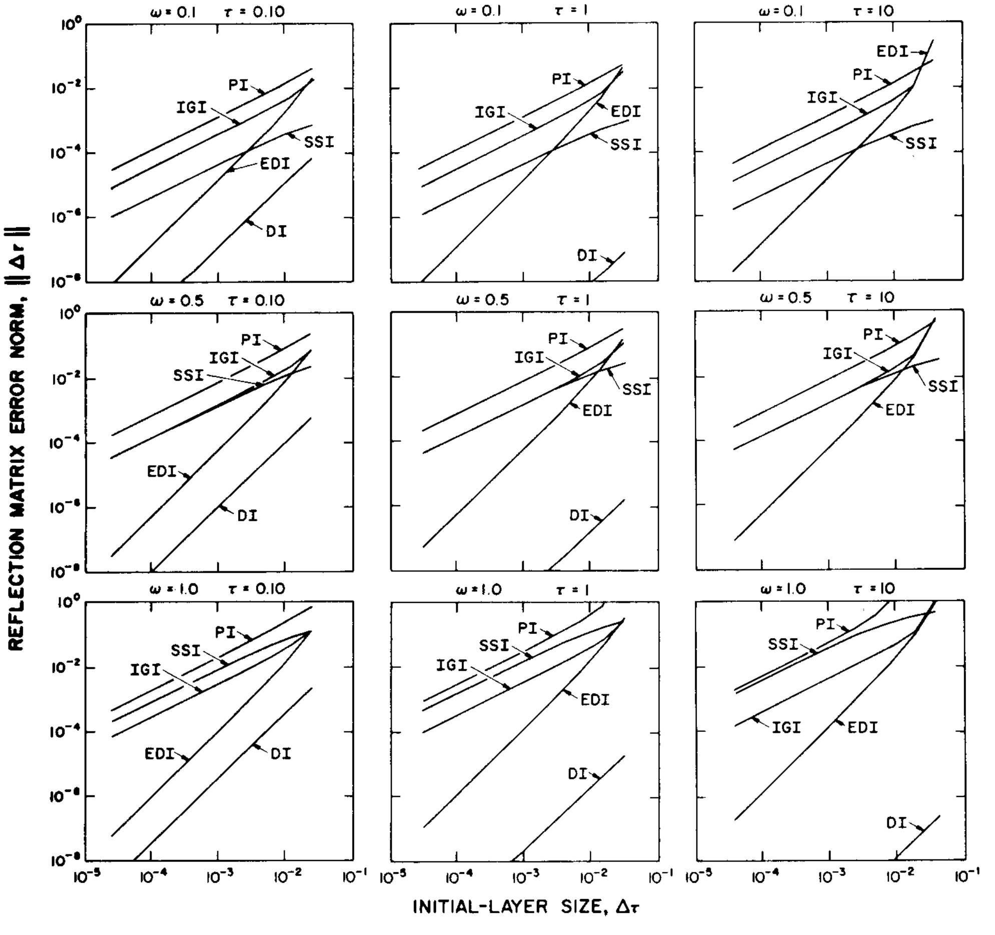

Fig. 1. $\|\Delta r\|$ vs $\Delta\tau$ for $g = 0{\cdot}9$, $m = 8$ Gaussian angles, $\omega = 0{\cdot}1, 0{\cdot}5, 1{\cdot}0$, and $\tau = 0{\cdot}1, 1, 10$.

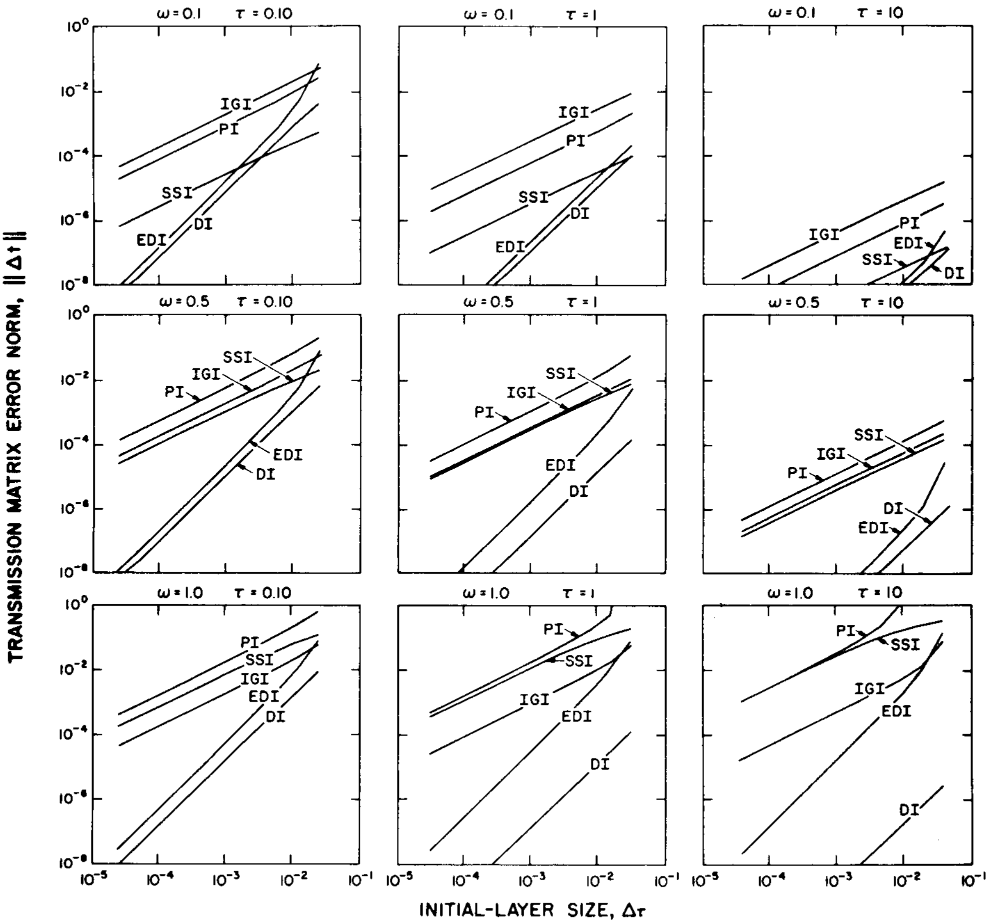

Fig. 2. $\|\Delta t\|$ vs $\Delta\tau$, same cases as Fig. 1.

m which we have studied, we have never observed a case in which $\|\Delta r\|$ and $\|\Delta t\|$ for the DI exceed 10^{-1}, and the worst cases are always when $\tau \approx \mu_1$ (μ_1 is a good approximation to $(\Delta\tau)_{\max}$) and $\Delta\tau = \tau$, so that no doubling is done. In general, using $\Delta\tau \approx \mu_1$ with the DI will almost always produce fluxes accurate to better than 1%, using the arguments of Section 4 relating $\|\Delta r\|$ and $\|\Delta t\|$ to flux error. This 1% limit will actually be approached for optically thin layers, but by the time one has doubled up to $\tau = 1$ a figure of 0·01% maximum flux error is more appropriate, and this figure continues to plummet dramatically as one doubles beyond $\tau = 1$. In fact, another remarkable property of the DI is just this dramatic error reduction with each doubling, as one may observe by looking across the rows of Figs. 1 and 2. It is particularly noteworthy that the curves for the other four initializations in Fig. 1 hardly move at all as τ increases across each row, while at the same time the diamond curve is falling precipitously. All the curves fall as we look across the rows of Fig. 2, due to the decline in transmission as $\tau \to \infty$, but clearly the diamond curve falls faster than the others, the more so the larger ω is. As we proceed down the columns of Fig. 1, corresponding to increasing ω, the diamond curve rises 1–2 orders of magnitude. All the other initializations experience a comparable loss in accuracy as ω increases with the exception of the SSI where the loss in accuracy is rather more drastic.

Note that, in order to match the accuracy of the DI with say $\Delta\tau = 10^{-2}$ ($\approx \mu_1$), the other schemes, with the occasional exceptions of SSI and EDI, require $\Delta\tau$'s orders of magnitude smaller. Since a single doubling involves more computation than the entire DI (if the shortcuts of Appendix A are used), it follows that the DI can guarantee a given level of flux accuracy—always better than 1%—with significantly less computation. If smaller $\Delta\tau$'s than 10^{-2} are desired, in order to obtain improved accuracy, the advantage of the DI over all the others except the EDI escalates rapidly, due to its superior convergence properties (steeper slope of its curves in Figs. 1 and 2).

In order to show that our conclusions are not sensitive to angular discretizations, we performed comparisons analogous to Figs. 1 and 2 for $m = 4$ and $m = 16$ Gaussian quadrature points also. A few of the $\tau = 1$ results for $m = 16$ are shown in Fig. 3 ($\|\Delta r\|$) and Fig. 4 ($\|\Delta t\|$), again for $\omega = 0{\cdot}1$, $0{\cdot}5$ and 1, but this time for $g = 0$ rather than $g = 0{\cdot}9$. Comparison of corresponding plots in Figs. 1 and 3 and in Figs. 2 and 4 leads to the conclusion that the relative ordering of the curves, their separations, and in general their relations to one another are rarely altered by the large changes in m and g. Even the actual values of $\|\Delta r\|$ and $\|\Delta t\|$ do not change by much more than an order of magnitude. These conclusions are borne out by more extensive comparisons. Thus Figs. 1 and 2 furnish a fairly universal set of error estimates in spite of being restricted to a particular m and g.

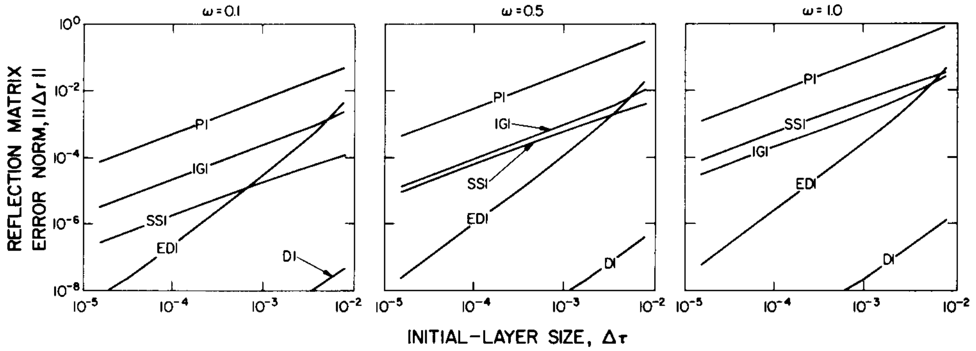

Fig. 3. $\|\Delta r\|$ vs $\Delta\tau$ for $m = 16$ Gaussian angles, $g = 0$, $\tau = 1$ and $\omega = 0{\cdot}1, 0{\cdot}5, 1{\cdot}0$.

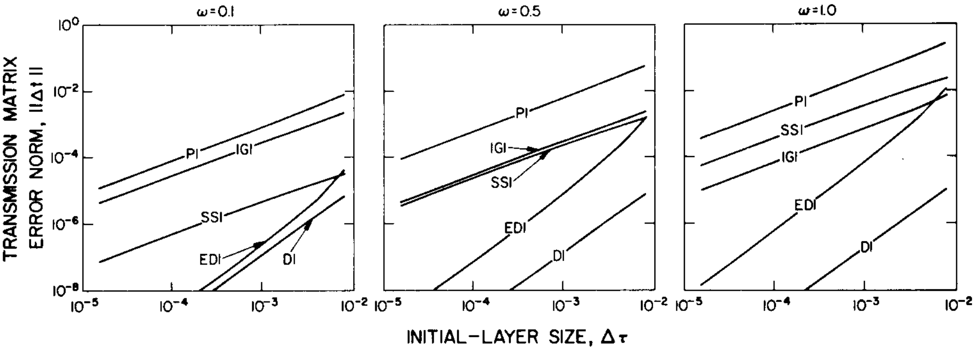

Fig. 4. $\|\Delta t\|$ vs $\Delta\tau$, same cases as Fig. 3.

We now discuss the remaining initializations, roughly in order of decreasing accuracy and desirability. The EDI is clearly competitive with the DI with regard to rapidity of convergence and, for small τ, in size of $\|\Delta r\|$ and especially $\|\Delta t\|$. For larger τ, it tends to be tremendously inferior to the DI for computing reflection, but is still fairly competitive in computing transmission as long as ω is not near unity. Unfortunately, since the EDI [eqns (13)] involves four matrix multiplications, it offers no clear computational advantage over the DI if the DI is computed as in Appendix A.

The IGI is consistently inferior in accuracy to the DI and EDI and to the SSI as well if $\omega \le 0{\cdot}5$. For $\omega \ge 0{\cdot}5$ it is superior in accuracy to the SSI, which seems strange in view of the derivation [Section 3(e)] which implies that the SSI is a more accurate O(ω) approximation than the IGI. However, that derivation was inconsistent in that a series was summed explicitly in the O(ω) terms and not in the O(ω^2) and higher terms. A similar inconsistency was noted in the derivation of the PI [Section 3(f)]. Such inconsistencies are no doubt responsible, not only for inferior accuracy, but for the flux conservation difficulties which we shall encounter in the next section.

The PI is consistently the least accurate of the five. Note that both the PI and SSI are constructed so as to give the correct transmissions (i.e. have zero-error) at $\omega = 0$, but that their accuracies have already deteriorated seriously by $\omega = 0{\cdot}1$, relative to schemes which do not limit properly as $\omega \to 0$. Having the correct behavior as $\omega \to 0$ seems to hurt rather than help the accuracy and flux conservation ability of an initialization.

It is a simple matter to show empirically that the IGI, SSI and PI all have $O(\Delta\tau)$ convergence, while the DI and EDI have $O[_{\Delta}\tau^2]$ convergence. The quantities $\|\Delta r\|$ and $\|\Delta t\|$ from Figs. 1 and 2 have been divided by $(\Delta\tau/\mu_1)$ for the $O(\Delta\tau)$ initializations and by $(\Delta\tau/\mu_1)^2$ for the $O[\Delta\tau^2]$ initializations. Sample results from this procedure are presented in Fig. 5. Since the curves are nearly horizontal, the rate of convergence has obviously been deduced correctly. The deviation from horizontality as $\Delta\tau$ increases reflects the influence of higher powers of $\Delta\tau$ in the Taylor expansion of the error.

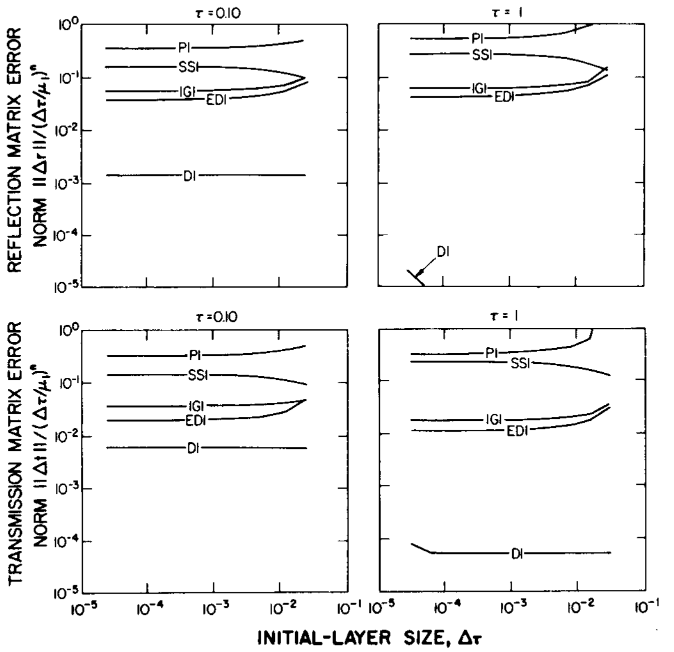

Fig. 5. $\|\Delta r\|$ (top row) and $\|\Delta t\|$ (bottom row) divided by $(\Delta\tau/\mu_1)^n$ vs $\Delta\tau$ for $m = 8$, $g = 0{\cdot}9$, $\omega = 1$ and $\tau = 0{\cdot}1$, 1; $n = 2$ for DI and EDI, $n = 1$ otherwise.

For the purpose of studying round-off error growth in the various initializations, the computations were downgraded from 29 to 14 significant digits (CDC single precision, IBM double precision) and the $\Delta\tau$ range was extended down to 10^{-8}. Sample results for $\|\Delta r\|$ and $\|\Delta t\|$ vs $\Delta\tau$, for $\omega = 1$ and $\tau = 1$, are shown in Fig. 6. The DI succumbs to round-off error somewhat below $\Delta\tau = 10^{-4}$, and the EDI somewhat below $\Delta\tau = 10^{-5}$. (As a general rule, the more accurate the approximation, the more susceptible it will be to round-off error.) The other initializations just begin to succumb between $\Delta\tau = 10^{-7}$ and 10^{-8}; the IGI breaks down somewhat before the others. It is curious that Hansen,[3] using $\Delta\tau = 2^{-25} = 3 \times 10^{-8}$, may have been operating very near the threshold of round-off catastrophe. Of course, the round-off limit on $\Delta\tau$ depends strongly on the machine precision and to a lesser extent on ω, τ and m, so no hard and fast rules can be given here. But it would seem unwise to use $\Delta\tau$'s much smaller than $0{\cdot}1\mu_1$ with the DI, especially since $\Delta\tau = 0{\cdot}1\mu_1$ will give more than sufficient accuracy for most applications.

A final conclusion to be drawn from Figs. 1 to 4 is that so-called "benchmark" calculations using the doubling method, with the popular IGI or SSI, may compute fluxes correct to only one or two significant digits if $\Delta\tau$ is not sufficiently small. Heating rates computed from differencing such fluxes may have no significant digits whatsoever.

6. FLUX CONSERVATION AND PHASE FUNCTION RENORMALIZATION

In order to conserve flux, it is essential (Twomey *et al.*;[2] Grant and Hunt[10]) that the phase function normalization condition be satisfied in its quadratured form:

$$\sum_{i=1}^{m} c_i(p_{ij}^{++} + p_{ij}^{+-}) = 2 \quad (j = 1, \ldots, m). \qquad (24)$$

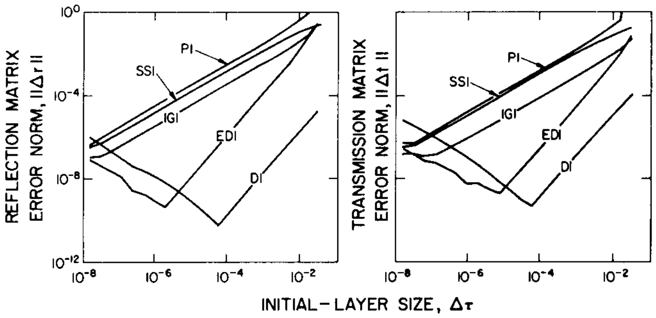

Fig. 6. $\|\Delta r\|$, $\|\Delta t\|$ vs extended range of $\Delta\tau$ for $m = 8$, $g = 0{\cdot}9$, $\omega = 1$ and $\tau = 1$.

This is a system of m equations and therefore, at best, can only determine m unknowns. Hence the problem of determining $(m^2 + m)$ correction factors, one for each element of p^{++} and p^{+-} (the number is reduced from $2m^2$ by symmetry), is underdetermined. This arbitrariness has led to several proposed methods for renormalizing p^{++} and p^{+-}, which are discussed below in order of increasing complexity.

The simplest is due to Grant (private communication), who corrects only the m diagonal elements, identically for p^{++} and p^{+-}. Since the diagonal elements are larger than the off-diagonal ones, this puts the entire correction where it has the least relative impact. To give an idea of the size of this correction, for $g = 0{\cdot}9$ the largest one is 46% for $m = 4$, 21% for $m = 8$ and 3·7% for $m = 16$ (the correction obviously goes to zero as $m \to \infty$ or as $g \to 0$). Thus it may cause a substantial distortion in a highly asymmetric phase function. If the unrenormalized matrices are denoted by $\hat{p}^{++}$ and $\hat{p}^{+-}$, the Grant method is

$$p_{ij}^{+\pm} = (1 + \epsilon_j \delta_{ij}) \hat{p}_{ij}^{+\pm}.$$

From substituting this into eqn (24), one finds

$$\epsilon_j = \frac{2 - b_j}{c_j(\hat{p}_{jj}^{++} + \hat{p}_{jj}^{+-})},$$

where

$$b_j \equiv \sum_{i=1}^{m} c_i(\hat{p}_{ij}^{++} + \hat{p}_{ij}^{+-}).$$

The Grant method was used in the present computations.

A method in which every element is corrected was proposed by Wiscombe *et al.*:[21]

$$p_{ij}^{+\pm} = (1 + \epsilon_i + \epsilon_j) \hat{p}_{ij}^{+\pm}.$$

Equation (24) then yields a set of linear equations for the ϵ's:

$$\sum_{i=1}^{m} B_{ij} \epsilon_i = 2 - b_j \quad (j = 1, \ldots, m),$$

where

$$B_{ij} = b_j \delta_{ij} + c_i(\hat{p}_{ij}^{++} + \hat{p}_{ij}^{+-}).$$

The general effect of this method is similar to Grant's, that is, the largest corrections turn out to be on the diagonals, but it distributes the corrections more uniformly over p^{++} and p^{+-}.

Next up in complexity is Hansen's[6] method, in which only p^{++} is corrected; the corrections are computed iteratively from a formula of unstated origin.

The most complex method is that of Twomey *et al.*,[2] which involves finding the eigenvectors and eigenvalues of

$$A = \begin{pmatrix} c\hat{p}^{++} & c\hat{p}^{+-} \\ c\hat{p}^{+-} & c\hat{p}^{++} \end{pmatrix}.$$

By the Perron–Frobenius theorem (Todd, Ref. (22), p. 290), since A has all positive elements it has a positive eigenvalue λ and a unique corresponding eigenvector e_λ with positive components. All the other eigenvectors must have at least one negative component since they are orthogonal to e_λ. The eigenvector e_λ should be all ones by eqn (24); if it isn't, it is made so, λ is changed to 2, and the whole set of eigenvectors reorthogonalized. A new A-matrix is then constructed from the modified eigenvalues and eigenvectors, from which p^{++} and p^{+-} are then extracted. This method is more elegant than the previous ones, but at the price of considerably more computation. It would be interesting to know which of these renormalization methods, if any, leads to the smallest flux error.

We now address the question of flux conservation, or alternatively spurious absorption, in layers for which $\omega = 1$. Of course spurious absorption does not disappear when $\omega < 1$, but then it is masked by, and usually insignificant compared to, true absorption. Nevertheless, any numerical scheme exhibiting non-negligible spurious absorption is seriously deficient; such is the case for the single-scattering initialization, as we shall see below.

If we add together eqns (1a and b) of the interaction principle, ignore sources, and take the norm of both sides, we obtain

$$F_{\text{out}} \leq \|S\| F_{\text{in}}, \tag{25}$$

where $\|S\| = \|r + t\|$, $F_{\text{out}} = \|i^+(\tau) + i^-(0)\|$, $F_{\text{in}} = \|i^+(0) + i^-(\tau)\|$. F_{out} and F_{in} are the total fluxes out of and into the layer, respectively. From eqn (25), we deduce immediately that $\|S\| \geq 1$ is necessary for flux conservation ($F_{\text{out}} = F_{\text{in}}$), but not sufficient. Thus when $\|S\| < 1$ we know flux cannot be conserved, whereas when $\|S\| \geq 1$ we can make no conclusions regarding flux conservation. Grant and Hunt[11] give this necessary condition in the more restricted form $\|S\| = 1$, but the author has observed situations [all for $\Delta\tau$ exceeding the upper bounds of Section 3(g)] in which $\|S\| \gg 1$ and yet flux is conserved to many significant figures. It is quite simple to show that flux conservation is preserved by layer addition—one need only add together the equations $F_{\text{out}} = F_{\text{in}}$ for each layer. Thus, if the initial layer $\Delta\tau$ for doubling conserves flux, so will the entire layer τ. Since the initial layer cannot conserve flux unless

$$\|S_0\| = \|r_0 + t_0\| \geq 1, \tag{26}$$

then it is clear that having $\|S_0\| \geq 1$ when $\omega = 1$ is a desirable property for an initialization to possess.

For the IGI, $\|S_0\| \geq 1$ may be proven directly from the matrix norm definition (22b) and the renormalization condition (24). For $\omega = 1$ and

$$\Delta\tau \leq (\Delta\tau)_s = \min_i \left[\frac{\mu_i}{1 - \frac{1}{2}c_i(p_{ii}^{++} + p_{ii}^{+-})}\right]$$

we have $\|S_0\| = 1$, and for $\Delta\tau > (\Delta\tau)_s$ we have $\|S_0\| > 1$. Note that $(\Delta\tau)_s$ is somewhat larger than $(\Delta\tau)_{\max}$ [eqn (20)].

The complicated nature of the DI [eqns (12)] precludes the use of the norm definition (22b) to calculate $\|S_0\|$. We have only been able to prove rigorously that, if $\Delta\tau \leq 2(\Delta\tau)_S$, $\|S_0\| < 1$ when $\omega < 1$ and $\|S_0\| \leq 1$ when $\omega = 1$. This proof is given in Appendix C. But computational experience has shown that $\|S_0\| = 1$ when $\omega = 1$ to at least 14 significant digits for the DI, provided $\Delta\tau \leq 2(\Delta\tau)_s$. For $\Delta\tau > 2(\Delta\tau)_s$ we find $\|S_0\| > 1$.

The EDI behaves very similarly, that is, $\omega = 1$ implies $\|S_0\| = 1$ to as many significant digits as the diamond case until $\Delta\tau$ exceeds some limit slightly larger than $(\Delta\tau)_s$; beyond this limit, $\|S_0\| > 1$. It should be emphasized, however, that the condition $\|S_0\| > 1$ does not necessarily destroy flux conservation, even when doubling to large optical depths. This holds true for the DI, the EDI and the IGI. Of course, there may be serious accuracy and non-negativity problems in such cases.

Figure 7 shows $\|S_N\| = \|r_N + t_N\|$ as a function of initial-layer size $\Delta\tau$ for $\tau = 2^N\Delta\tau = 0{\cdot}1$, 1 and 10. Results for both $g = 0$ and $g = 0{\cdot}9$ are included to show that renormalization (absent for $g = 0$) cannot be responsible for the observed behavior of $\|S_N\|$. As $\Delta\tau$ and/or τ increases, we see that $\|S_N\|$ for the SSI becomes increasingly less than unity; thus the SSI is incapable of conserving flux. The PI causes $\|S_N\|$ to increase rapidly above unity as $\Delta\tau$ and/or τ increase, so that it may still conserve flux; but whether it does or not is really irrelevant in view of the poor accuracy of this scheme (see

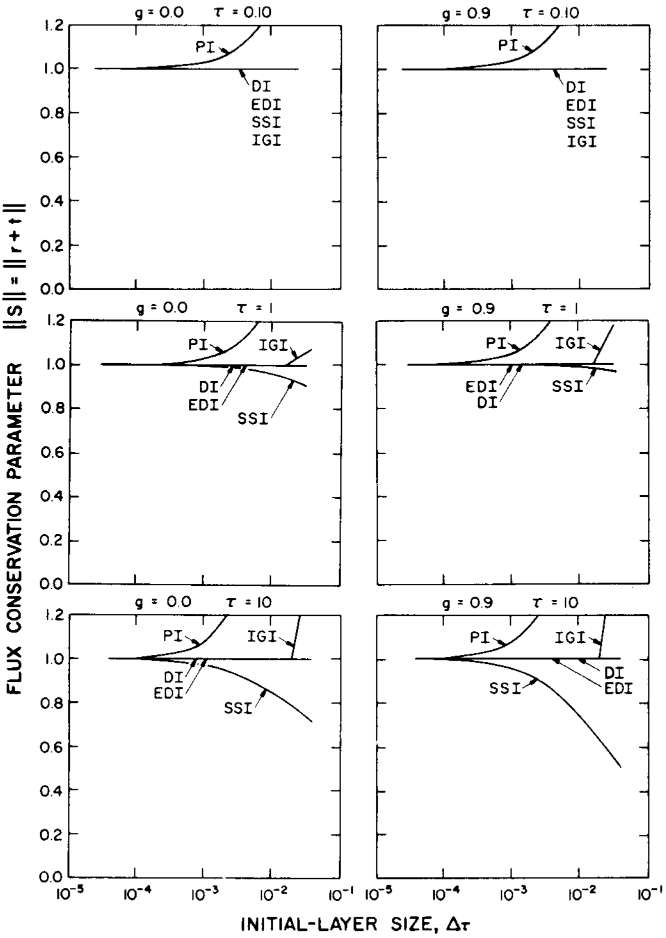

Fig. 7. $\|S_N\|$ vs $\Delta\tau$ for $\omega = 1$, $m = 8$, $g = 0$ and 0·9 and $\tau = 0{\cdot}1, 1, 10$.

Section 5), and we consider it no further. Figure 7 also shows $\|S_N\| = 1$ for the other three initializations (except where the IGI curve angles steeply upward for $\Delta\tau > (\Delta\tau)_s \sim 0{\cdot}02$).

In order to directly study the spurious absorption due to the various initializations, we computed the absorptivity

$$a \equiv (F_{\text{in}} - F_{\text{out}})/F_{\text{in}}$$

when $\omega = 1$ for layers of various optical depths τ and assuming various distributions of incident intensity. Figure 8 shows a vs τ for $i^-(\tau) = 0$ (no flux into the bottom of the layer) and for $i^+(0)$ either along the most nearly-grazing or along the most nearly-normal quadrature angle. Other specifications of incident intensity led to absorptivities between the normal and grazing cases. Figure 8 is for $g = 0{\cdot}9$ and $m = 8$, but qualitatively, and quantitatively to within an order of magnitude, a is the same for $0 \leq g \leq 0{\cdot}9$ and $4 \leq m \leq 16$. Results are shown for $\Delta\tau = \mu_1/10$ (0·002) and for $\Delta\tau = \mu_1(0{\cdot}02)$. The absorptivity increases about linearly with $\Delta\tau$ and with τ; therefore it doubles with each doubling step. It is negligibly small and almost the same for the DI, the EDI and the IGI. It is intolerably large for the SSI, easily reaching 1–10% and more, unless $\Delta\tau$ is taken very much smaller than the values of Fig. 8. Perhaps this explains why Hansen[3] took $\Delta\tau \sim 10^{-8}$, a value much smaller than accuracy alone would dictate.

From eqn (25) and the definition of absorptivity a, it follows that

$$a \gtrsim 1 - \|S\|.$$

In general we find that $a \gg 1 - \|S\|$ initially and for the first few doubling steps (provided $\|S\| < 1$), but after that $a \simeq 1 - \|S\|$, so that the absorptivity can be estimated from the layer property $\|S\|$ without postulating a specific incident intensity.

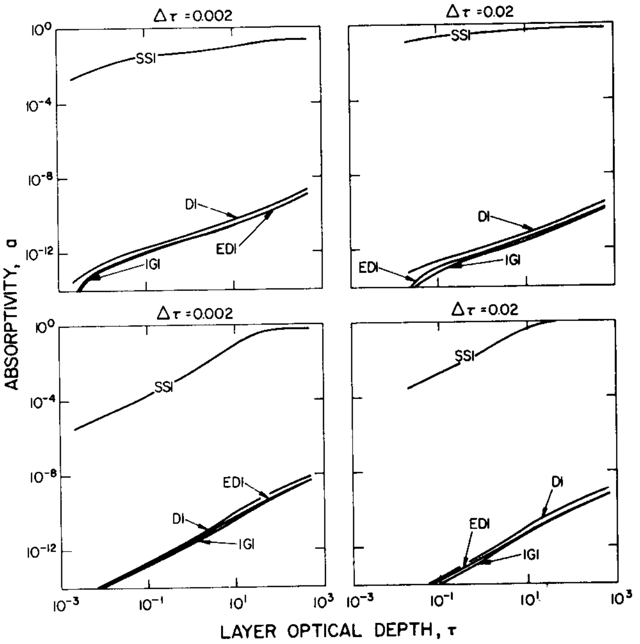

Fig. 8. Spurious absorptivity a vs τ for $m = 8$, $g = 0{\cdot}9$, $\Delta\tau = \mu_1/10$ and μ_1, and for incident intensity near-grazing (top row) and near-normal (bottom row).

7. ANALYTIC ERROR ESTIMATES

Restricting our attention to the DI, we shall derive analytic estimates of $\|\Delta r\|$ and $\|\Delta t\|$ which are useful for $\tau \leq 10$. An analogous treatment is equally useful for the IGI but is omitted here.

We begin by showing, in Fig. 9, examples of how the errors $\|\Delta r\|$ and $\|\Delta t\|$, divided by $(\Delta\tau/\mu_1)^2$ to normalize out the $\Delta\tau$-variation, vary with τ for $\omega = 0{\cdot}1$, $0{\cdot}5$ and $1{\cdot}0$, for $g = 0{\cdot}9$, and for $m = 4$ and 8 Gaussian angles. We note two things which are typical of such curves. First, the error peaks out and is quite large for optically thin layers but decreases dramatically as we continue doubling to thicker layers. The peak moves toward smaller τ as m increases. Second, the spread

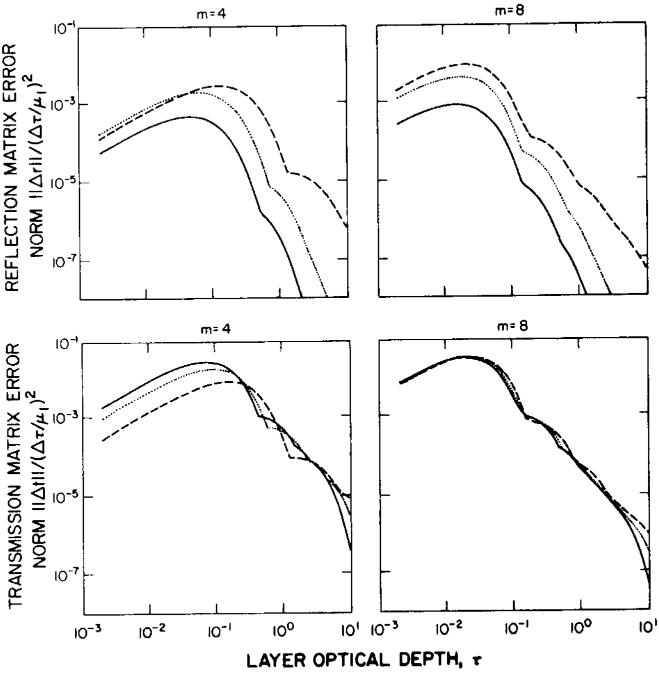

Fig. 9. $\|\Delta r\|$ (top row) and $\|\Delta t\|$ (bottom row) divided by $(\Delta\tau/\mu_1)^2$ vs τ for DI and for $g = 0{\cdot}9$, $m = 4, 8$. Curve labels: ——, $\omega = 0{\cdot}1$; ···, $\omega = 0{\cdot}5$; ---, $\omega = 1{\cdot}0$ (same labels apply to Figs. 10 and 11).

between the $\omega = 0{\cdot}1$ and $\omega = 1{\cdot}0$ curves is reasonably small compared to the range of variation of the error, especially for $\|\Delta t\|$, which suggests that analytic error estimates for $\omega \to 0$ will be useful for all ω.

The exact $\omega = 0$ transmission error can be derived as follows. Putting $r_n = 0$ and hence $\Gamma_n = I$ in the doubling formula (3a) and initializing t_0 from eqn (12b) with $\omega = 0$, we obtain

$$t_N = \left[\left(\frac{1-\dfrac{\Delta\tau}{2\mu_i}}{1+\dfrac{\Delta\tau}{2\mu_i}}\right)^{2^N}\delta_{ij}\right].$$

The exact transmission is

$$t_\infty = \lim_{\substack{\Delta\tau\to 0\\(\tau\ \text{fixed})}} t_N = [\mathrm{e}^{-\tau/\mu_i}\,\delta_{ij}].$$

Therefore the $\omega = 0$ error is

$$E_T \equiv \|t_\infty - t_N\| = \frac{1}{12}\,\tau(\Delta\tau)^2 \max_i \left[\frac{\mathrm{e}^{-\tau/\mu_i}}{\mu_i^3}\right], \tag{27}$$

where we have expanded and neglected terms of order $(\Delta\tau)^4$ and higher, in consonance with our empirical finding in Section 5 that $\|\Delta t\|$ is nearly proportional to $(\Delta\tau)^2$. The value of μ at which the maximum in eqn (27) is assumed depends on τ, and results in E_T having a τ-dependence of the following form

$$E_T \propto \begin{cases} \tau\,\mathrm{e}^{-\tau/\mu_1} & \tau \le 3\mu_1 \\ \tau^{-2} & 3\mu_1 < \tau < 3\mu_m. \\ \tau\,\mathrm{e}^{-\tau/\mu_m} & \tau \ge 3\mu_m \end{cases}$$

The peak is assumed in the first sub-range, at $\tau = \mu_1$, followed by a monotonic decrease through the next two sub-ranges, as we indeed observe in Fig. 9. The τ^{-2} dependence in the mid-range is only approximate since the μ_i are a discrete set, so it is preferable to use eqn (27) directly in the mid-range.

In the process of expansion in powers of $\Delta\tau$ which leads to eqn (27), one finds that higher-order terms may *not* be neglected if

$$\tau(\Delta\tau)^2 \gg 1. \tag{28}$$

This situation was not triggered in any cases considered here since our maximum $\Delta\tau$ was on the order of $2\mu_1$, so that even for only 4 Gaussian angles condition (28) would be at least $\tau \gg 50$. But in a similar analysis for the IGI, eqn (28) is replaced by $\tau\ \Delta\tau \gg 1$, which may invalidate the expansion for fairly small τ's.

In Fig. 10 we show $\|\Delta t\|$ divided by E_T vs τ for the same values of ω as in Fig. 9, for $g = 0$ and $0{\cdot}9$, and for $m = 4$ and 8. When the curves are close to unity, then of course E_T is a good approximation to $\|\Delta t\|$. The approximation is seen to be excellent for $\omega = 0{\cdot}1$, easily within a factor of 2 everywhere, as we might have expected. The approximation grows progressively worse for $\omega = 0{\cdot}5$ and $\omega = 1{\cdot}0$ but is still easily within an order of magnitude almost everywhere. It deteriorates rapidly beyond about $\tau = 5$ and is considerably worse for highly asymmetric scattering ($g = 0{\cdot}9$) than for isotropic scattering ($g = 0$). Increasing the number of angles (m) seems to improve the accuracy of E_T as an error estimate in all cases. The sawtooth appearance of some of the curves is due to the maximum in eqn (27) jumping between values of μ_i, starting at μ_1 for small τ and winding up at μ_m for large τ.

Since $r_n = 0$ for $\omega = 0$, we must perforce go to a more complex $\mathrm{O}(\omega)$ theory to get an estimate for $\|\Delta r\|$. Fortunately the analysis of Section 3(e) can be repeated beginning with the DI [expanded to $\mathrm{O}(\omega)$] rather than the IGI. We dispense with the details and merely give the result corresponding to eqn (16):

$$(r_N)_{ij} = \frac{\omega}{2}\frac{c_j}{\mu_i}\,p_{ij}^{+-}[1-(e_ie_j)^{2^N}]\Big/\left(\frac{1}{\mu_i}+\frac{1}{\mu_j}\right), \tag{29}$$

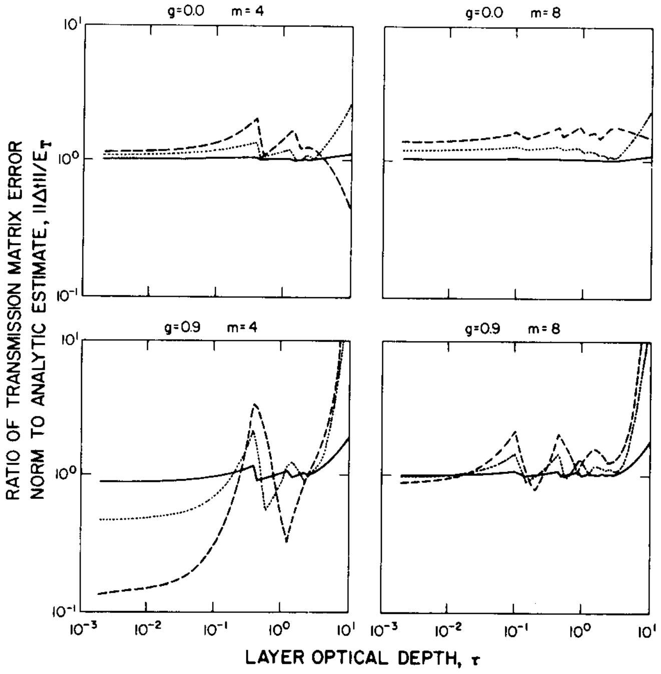

Fig. 10. $\|\Delta t\|$ divided by E_T vs τ for DI and for $g = 0, 0{\cdot}9$ and $m = 4, 8$.

where

$$e_i = \left(1 - \frac{\Delta\tau}{2\mu_i}\right) \bigg/ \left(1 + \frac{\Delta\tau}{2\mu_i}\right).$$

We take the difference between eqn (29) and the single-scattering result of eqn (18a) [to which eqn (29) limits as $\Delta\tau \to 0$ with $2^N \Delta\tau$ fixed] as our error estimate:

$$E_R = \|r_\infty - r_N\|,$$

where

$$(r_\infty - r_N)_{ij} = \frac{\omega}{24} \frac{c_j \mu_j}{\mu_i + \mu_j} \left(\frac{1}{\mu_i^3} + \frac{1}{\mu_j^3}\right) p_{ij}^{+-} \tau \, e^{-[(1/\mu_i)+(1/\mu_j)]\tau} (\Delta\tau)^2,$$

which has been expanded to order $(\Delta\tau)^2$ as we did for E_T (the same caveat on the size of $\tau(\Delta\tau)^2$ applies here). Note that since E_R is linear in ω it is primarily measuring the error due to singly-scattered radiation. To the extent that the reflected radiation is multiply-scattered, E_R will lose accuracy.

It might be remarked that an $O(\omega)$ correction may be added to E_T [eqn (27)] following the approach for E_R. This will improve the estimate somewhat for thinner layers, but since for thicker layers most of the transmitted radiation is multiply scattered, no improvement will result there.

In Fig. 11 we show $\|\Delta r\|$, divided by E_R, vs τ for the same parameter set as in Fig. 10. The utility of E_R as an analytic error estimate is seen to be somewhat less than that of E_T. There is a particularly marked deterioration as ω increases in the $g = 0{\cdot}9$ case, and in general the E_R estimate seems to be substantially better when $g = 0$ than when $g = 0{\cdot}9$. This can be understood in terms of the much larger multiply-scattered component of the reflected radiation when $g = 0{\cdot}9$, which E_R does not account for. As a general rule, E_R can only be relied on for an order of magnitude error estimate for $\tau \leq 0{\cdot}1$; however, depending on the particular combination of ω and g, it may be reliable up to τ's of 5–7.

We remark finally that we have purposely considered only $\tau \leq 10$ in Figs. 9–11 because unless $\omega \approx 1$ there is little significant change in r and t with further doubling. Even for $\omega = 0{\cdot}99$, $\|r\|$ and $\|t\|$ have essentially reached their asymptotic values by $\tau = 100$. As $n \to \infty$ the doubling formulas

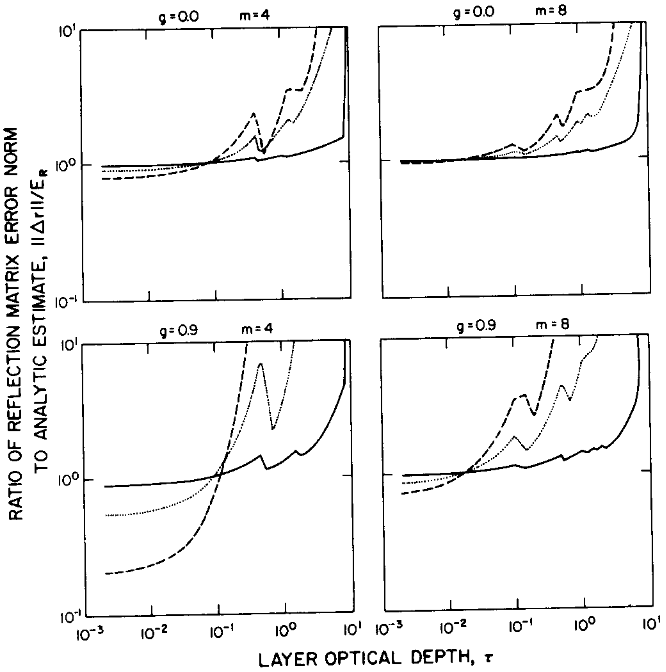

Fig. 11. $\|\Delta r\|$ divided by E_R vs τ, same cases as Fig. 10.

(3) of course reduce to $t_{n+1} \sim 0$, $r_{n+1} \sim r_n$, and the point is that this happens rather rapidly when $\omega < 1$.

8. SUMMARY AND CONCLUSIONS

The main goal of this paper has been to provide estimates of doubling error for reflection and transmission, for a variety of initializations, which can be directly used in estimating flux error [eqn (23]. Figures 1–6 provide an abundance of information in this regard of which we give only the highlights below. The other major goal has been to compare the flux conservation properties of the various initializations.

The doubling initializations studied differ widely in accuracy, and two little-known ones, the diamond and expanded diamond, are found to exhibit vastly superior accuracy and convergence (as $\Delta\tau \to 0$) compared to the commonly-used infinitesimal generator and single-scattering initializations. With the diamond initialization in particular, it is possible to achieve a pre-set accuracy with an initial-layer size many orders of magnitude larger than what the other initializations would require. The single-scattering initialization, surprisingly, is found to be *less* accurate than the simpler infinitesimal generator when $\omega \geq 0{\cdot}5$. The diamond and expanded diamond initializations succumb to round-off error for initial layers $\Delta\tau$ somewhat smaller than 10^{-3}, while the other initializations are stable against round-off down to $\Delta\tau \approx 10^{-7}$ (for 14-significant digit computations).

Richardson extrapolation is exploited, apparently for the first time in this connection, to obtain reflection and transmission matrices accurate to at least 8 significant digits. This permits one to skirt the formidable round-off error problems associated with a straightforward passage to the limit of small initial-layer size. The diamond, expanded diamond, and infinitesimal generator initializations are all shown to have excellent flux conservation properties. The single-scattering initialization, on the other hand, is found to be very poor in this regard.

By judicious use of a new derivation of the single-scattering initialization, analytic error estimates are derived which prove useful for restricted ranges of layer optical depth, single-scattering albedo, and asymmetry parameter.

As a final note, the author hopes that the present work may help to stimulate a somewhat greater awareness of the errors which lurk in supposedly "exact" calculations of radiative fluxes.

Acknowledgements—The author would like to thank Prof. I. P. Grant for stimulating his interest in this subject and Dr. S. Schneider for encouraging him to pursue that interest.

REFERENCES

1. H. C. VAN DE HULST, A new look at multiple scattering, *Tech. Rept.* Goddard Institute for Space Studies, NASA, New York, 81 pp (1963).
2. S. TWOMEY, H. JACOBOWITZ and H. B. HOWELL, *J. atmos. Sci.* **23**, 289 (1966).
3. J. E. HANSEN, *Ap. J.* **155**, 565 (1969).
4. J. E. HANSEN, *J. atmos. Sci.* **26**, 478 (1969).
5. J. E. HANSEN, *J.* ~~*atmos. Sci.* **28**, 120 (1971).~~
6. J. E. HANSEN, *J. atmos. Sci.* **28**, 1400 (1971).
7. I. P. GRANT and G. E. HUNT, *Icarus* **9**, 526 (1968).
8. G. N. PLASS, G. KATTAWAR and F. CATCHINGS, *Appl. Opt.* **12**, 314 (1973).
9. H. C. VAN DE HULST and K. GROSSMAN, In *The Atmospheres of Venus and Mars* (Edited by BRANDT and MCELROY). Gordon & Breach, New York (1968).
10. I. P. GRANT and G. E. HUNT, *Proc. R. Soc. London* **A313**, 183 (1969).
11. I. P. GRANT and G. E. HUNT, *Proc. R. Soc. London* **A313**, 199 (1969).
12. W. J. WISCOMBE and B. FREEMAN, *A Detailed Radiation Model for Climate Studies*, Pre-print volume, Conf. Atmos. Rad. (Ft. Collins, Colo.), Am. Met. Soc., Boston, Mass. (1972).
13. W. J. WISCOMBE, In *Climate of the Arctic* (Edited by G. WELLER and S. BOWLING). 24th Alaskan Science Conference. Geophysical Institute, Univ. of Alaska Press, Fairbanks (1973).
14. G. E. HUNT, *JQSRT* **11**, 309 (1971).
15. W. J. WISCOMBE, *JQSRT* **16** , 477 (1976).
16. D. C. JOYCE, *SIAM Rev.* **13**, 435 (1971).
17. B. CARLSON, In *Methods in Computational Physics*, Vol. 1. Academic Press, New York (1963).
18. I. P. GRANT and G. E. HUNT, *Mon. not. R. astr. Soc.* **141**, 27 (1968).
19. S. CHANDRASEKHAR, *Radiative Transfer.* Dover, New York (1960).
20. E. ISAACSON and H. KELLER, *Analysis of Numerical Methods.* Wiley, New York (1966).
21. W. J. WISCOMBE, B. FREEMAN and W. ENGLAND, The effects of meso-scale and small-scale interactions on global climate, *Rept. SSS-R*-72-1255, Systems, Science and Software, La Jolla, Calif. (1972).
22. J. TODD, [Ed], *Survey of Numerical Analysis.* McGraw-Hill, New York (1962).

APPENDIX A

Speeding up doubling and diamond initialization calculations

The basic idea behind the restructuring of the doubling and diamond initialization computations to be discussed below is that, in principle, it is roughly six times faster to solve a system of linear equations than to invert a matrix and then multiply it by another matrix (ISAACSON and KELLER Ref. (20), p. 34). Further computational savings may also be possible if the LU-decomposition from the linear system solution can be reused, as in the diamond case.

Doubling. Referring to eqns (3) in the paper, and dropping subscripts, it is clear that we must compute the matrix

$$A \equiv t(I - rr)^{-1}.$$

Rather than inverting $(I - rr)$ and multiplying the result by t, one should write the problem as a linear system,

$$(I - rr)^T A^T = t^T \tag{A1}$$

and solve for the columns of A^T (rows of A) using standard Gaussian elimination with LU decomposition. However, because of the way matrices are stored column-by-column in a computer, it may be desirable to modify eqn (A1) by making use of the relationships

$$r^T = DrD^{-1}, \quad t^T = DtD^{-1}, \tag{A2}$$

where

$$D = (\mu_i c_i \delta_{ij})$$

(μ_i and c_i are defined in Section 3). GRANT and HUNT[11] mention that eqns (A2) are valid for the IGI (Section 3d). However, it is not difficult to prove that they are also valid for the DI and the EDI [Sections 3(b) and (c)], and that these relationships are preserved under doubling. Therefore eqns (A2) are quite general. Using eqns (A2) in eqn (A1), one may derive

$$(I - rr)B = t, \tag{A3}$$

where

$$B \equiv D^{-1}A^T D.$$

Solving this last relationship for A,

$$A_{ij} = \frac{\mu_j c_j}{\mu_i c_i} B_{ji} \quad (1 \leq i, j \leq m). \tag{A4}$$

Timing studies have indicated that obtaining A from eqns (A3) and (A4) has the advantage over solving eqn (A1) directly.

It should also be noted that once A is available, no further matrix inverses are required in order to double for internal sources as well (WISCOMBE[15]).

Diamond initialization. Referring to eqns (12), it is seen that two matrix inverses are involved in the DI. The inverse of

$(I+T)$ always occurs in the combination

$$C = R(I+T)^{-1},$$

so that we should solve the linear system

$$(I+T)^T C^T = R^T$$

for C. However, as in the doubling case discussed previously, the easily-proved relationships

$$T^T = DTD^{-1}, \quad R^T = DRD^{-1},$$

allow us to solve instead the system

$$(I+T)B = R,$$

and recover C from

$$C_{ij} = \frac{\mu_j c_j}{\mu_i c_i} B_{ji} \quad (1 \le i, j \le m).$$

The other matrix inversion is bypassed in the following manner. Define

$$G = \tfrac{1}{2}[(I+T) - CR]$$

and obtain r, t and $\Sigma^{\pm}$ by solving the linear systems

$$Gr = C, \quad G(t+I) = I, \quad G\Sigma^{\pm} = \tfrac{1}{2}[\hat{\sigma}^{\pm}_{1/2} + C\hat{\sigma}^{\mp}_{1/2}].$$

The LU-decomposition of G should be saved during solution of the first system and reused in the subsequent solutions.

APPENDIX B

Sufficient upper bound on $\Delta\tau$ for diamond initialization

We prove that

$$\Delta\tau \le 2(\Delta\tau)_{\max},$$

implies that $r(\Delta\tau)$ and $t(\Delta\tau)$ for the DI [eqns (12a and b)] are non-negative matrices. $\Delta\tau \le 2(\Delta\tau)_{\max}$ implies $(I-T) \ge 0$, where T is defined in eqn (11c). We have $R > 0$ from its definition [eqn (11d)]. Since R and $(I-T)$ are the IGI for r and t for a layer of thickness $\frac{1}{2}\Delta\tau$, the norm definition (22a) implies that

$$\|R\| < 1, \quad \|I-T\| < 1. \tag{B1}$$

Defining

$$R_0 \equiv R(I+T)^{-1}, \tag{B2}$$

allows us to write the DI as

$$r(\Delta\tau) = 2(I+T)^{-1}(I-R_0^2)^{-1}R_0, \tag{B3}$$

$$t(\Delta\tau) = 2(I+T)^{-1}(I-R_0^2)^{-1} - I. \tag{B4}$$

In order to expand the inverse matrices in eqns (B2)–(B4), we require the theorem (ISAACSON and KELLER Ref. (20), p. 15) that, if $\|A\| < 1$, then

$$(I-A)^{-1} = I + A + A^2 + \cdots.$$

First prove $\|R_0^2\| < 1$. We have

$$\|R_0^2\| \le \|R_0\|^2 \le \left(\frac{1}{2}\|R\| \left\|\left(I - \frac{1}{2}(I-T)\right)^{-1}\right\|\right)^2$$

$$\le \left(\frac{\|R\|}{2 - \|I-T\|}\right)^2 < 1,$$

where eqn (B1) and two norm inequalities from Isaacson and Keller are used. Because $\|R_0^2\| < 1$ and $\|I-T\| < 1$, eqns (B3) and (B4) may be expanded as follows:

$$r(\Delta\tau) = [I + \tfrac{1}{2}(I-T) + \cdots][I + R_0^2 + \cdots]R_0,$$

$$t(\Delta\tau) = [I + \tfrac{1}{2}(I-T) + \cdots][I + R_0^2 + \cdots] - I.$$

But we may expand R_0 [eqn (B2)] as follows:

$$R_0 = \tfrac{1}{2}R[I - \tfrac{1}{2}(I-T)]^{-1} = \tfrac{1}{2}R[I + \tfrac{1}{2}(I-T) + \tfrac{1}{4}(I-T)^2 + \cdots]$$

so that $R_0 > 0$ because $R > 0$ and $(I-T) \geq 0$. Both r and t are clearly sums of series of non-negative matrices since $R_0 > 0$ and $(I-T) \geq 0$, so they themselves are non-negative. This completes the proof.

APPENDIX C

Flux conservation parameter $\|S_0\|$ *for diamond initialization*

Note that the quantities $(I-T)$ and R in eqns (11c and d) are the infinitesimal generator expressions for transmission and reflection matrices, for a layer of optical thickness $\frac{1}{2}\Delta\tau$. Adding the DI equations for reflection and transmission (12a and b) together and expressing the result in terms of $S_{\mathrm{IGI}} \equiv I - T + R$,

$$\begin{aligned}\|S_0\|_{\mathrm{DI}} &= \|(2I - S_{\mathrm{IGI}})^{-1}S_{\mathrm{IGI}}\| \\ &\leq \|S_{\mathrm{IGI}}\|\,\|(2I - S_{\mathrm{IGI}})^{-1}\| \\ &\leq \frac{\frac{1}{2}\|S_{\mathrm{IGI}}\|}{1 - \frac{1}{2}\|S_{\mathrm{IGI}}\|}\end{aligned}$$

This follows from two standard norm inequalities (ISAACSON and KELLER[20]). But, if $\frac{1}{2}\Delta\tau \leq (\Delta\tau)_s$, then $\|S_{\mathrm{IGI}}\| < 1$ when $\omega < 1$ and $\|S_{\mathrm{IGI}}\| = 1$ when $\omega = 1$. Therefore, $\|S_0\|_{\mathrm{DI}} < 1$ when $\omega < 1$ and $\|S_0\|_{\mathrm{DI}} \leq 1$ when $\omega = 1$.