JOURNAL OF COMPUTATIONAL PHYSICS 100, 253–269 (1992)

Stability Analysis of a Non-linear Diffusion-Type Kinetic Equation

HENRY GRANEK

School of Physics, University of Melbourne, Parkville, Victoria, Australia and CSIRO, Division of Atmospheric Research, Aspendale, Victoria, Australia

AND

BRUCE H. J. MCKELLAR

School of Physics, University of Melbourne, Parkville, Victoria, Australia

Received April 30, 1990; revised January 27, 1991

A diffusion-type partial differential equation with non-linear coefficients is analysed for stability in the von Neumann sense, and some numerical examples are given. The equation is a kinetic equation representing an instantaneous injection of energetic photons into a thermalised cosmological background radiation (CBR) and the subsequent time evolution of the electromagnetic spectrum. Compton, double Compton, and bremsstrahlung are the only interactions considered at the relatively photon energies low. The final conservative, implicit, finite difference scheme is a refinement of a similar model developed by Lightman, which is shown to be not stable for some cases considered. A semi-Lagrangian modification is used to account for the expansion of the universe. The full physical derivation of the kinetic equation and the associated parameters are given elsewhere.

CONTENTS

1. INTRODUCTION

Modelling of the Early Universe has enjoyed a lot of interest, particularly over last quarter of the century, when digital computers made it possible for us to numerically test the various scenarios. The period of the Early Universe following the primordial nucleosynthesis is of particular interest, because, it is thought during that period that the fluctuations in the initially homogeneous universe led to the formation of galaxies and stars. Some of the more conventional theories of galaxy formation require the presence of relatively massive neutral particles, that ultimately must decay, with a possible emission of electromagnetic radiation. The radiation emitted in such decays would lead to significant distortions in cosmological blackbody background radiation. If such distortions are not to be present today, the emitted radiation must have been thermalised in the period between the time when this radiation was emitted and the recombination of hydrogen, when electromagnetic radiation finally decoupled from matter.

In this paper we will not dwell in detail on the physics of the problem outlined above, as that is not the purpose of this article, and it is discussed in fine detail elsewhere [6–7] (see also references therein; Ref. [7] is henceforth referred to as GM). This paper is an exposé of the details of the numerical implicit finite difference model developed by the authors, in order to derive the constraints on the distortions of electromagnetic background radiation spectrum in the Early Universe [6–7]. The model assumes that the spectrum is *not* strongly distorted as a result of various processes occurring at those early times. This last assumption may not be correct, if some recent results [8, 22] are confirmed. This possibility, however, does not affect the conclusions from the numerical analyses discussed below, and that is the purpose of the present paper.

In the subsequent sections the finite difference approximations to the kinetic equation representing the interaction of radiation with electrons and protons are analysed in detail. Finite difference schemes are reviewed and applied to the physical model as presented in GM. Deficiencies of the various schemes are pointed out and alternative approaches offered. Stability in the von Neumann sense is investigated in order to predict the

0021-9991/92 $5.00

behaviour of the various numerical solutions. The model suffers from only a very mild case of truncation error, which is well within the limit of the error in computation, but with some effort it may possibly also be removed, while a potential instability in Lightman's [14] scheme due to the nonlinear term has been rectified successfully. The model is tested on a stable solution to confirm its consistency.

2. DESCRIPTION OF THE PROBLEM

The final form of the kinetic equation approximately representing the interaction of matter with electromagnetic radiation, also known as the *Kompaneetz* [12] equation with sources and sinks, is given by Eq. (2.0.1) (Eq. (4.16) of Ref. [6] or Eq. (3.1.10) of GM):

$$\left(\frac{\partial f}{\partial y}\right)_q - q\frac{d\psi}{dy}\left(\frac{\partial f}{\partial q}\right)_{T,t} = q^{-2}\frac{\partial}{\partial q}q^4\left(\left(\frac{\partial f}{\partial q}\right)_{T,t} + f(1+f)\right) + \frac{1}{\sigma_T n_e}\frac{m_e}{T}\{C_B[f] + C_{DC}[f]\}. \quad (2.0.1)$$

Here f is the phase space density function of the photon background (or their spectrum), y is referred to as the *optical depth* and is the equivalent of the time component, while q is a dimensionless momentum magnitude and is the equivalent of the space component in the diffusion-like equation (2.0.1). In fact, the first term on the right-hand side of (2.0.1) represents the photon diffusion in momentum space [19], while the remaining two terms represent a source/sink for the photons. The remaining parameters are T, the temperature of the photon gas at a given value of y or time t, n_e is the rest mass of the electron (we assume the nuclear units so that $c = \hbar = k = 1$, where c is the speed of light in vacuum, $\hbar$ is the Planck's constant divided by 2π and k is the Boltzmann's constant).

Equation (2.0.1) requires a number of subsidiary conditions and equations defining some of the other parameters. These are derived in GM and are only reproduced here for completeness. These are

$$T(y) = T_0\phi/\lambda \quad (2.0.2)$$

$$\psi = \ln\phi \quad (2.0.3)$$

$$\phi^4(y) = J_3(y_0)/J_3(y) \quad (2.0.4)$$

$$J_3(y) = \int_0^\infty q^3 f(q, y)\, dq \quad (2.0.5)$$

$$\frac{d\lambda}{dy} = \left(\frac{3Hm_e}{\sigma_T n_e T} - 4\frac{d\psi}{dy}\right) \times \left\{\frac{(4/3)(\rho_r)_0/\lambda^3 + (\rho_p)_0/\lambda^2}{4(\rho_r)_0/\lambda^4 + 2(\rho_p)_0/\lambda^3}\right\} \quad (2.0.6)$$

$$H^2 = \frac{8\pi G_N}{3}[(\rho_r)_0/\lambda^3 + (\rho_p)_0/\lambda^2], \quad (2.0.7)$$

where the subscript 0 means that the quantity is taken at some initial time $t = t_0$ when $\lambda = \phi = 1$ and $y = 0$, and for completeness, the expression for time is given by

$$t(y) = t_0 + \frac{m_e}{\sigma_T (n_e)_0 T_0}\int_0^y \lambda^4(y')/\phi(y')\, dy' \quad (2.0.8)$$

The problem therefore, is to find a time dependent solution of (2.0.1) with additional equations (2.0.2)–(2.0.7) describing a set of additional parameters. Note that if C_B and C_{DC} vanish in (2.0.1), then only Eqs. (2.0.3)–(2.0.5) are required and λ need not be evaluated, unless the expression for time t is required.

Equation (2.0.1) is a second-order partial differential equation in the *spatial* coordinate q and first order in the *temporal* coordinate y. In addition the problem has nonlinear coefficients. Mathematically, this type of differential equation falls in the general category of *parabolic partial differential equations*, of which the heat and diffusion equation are special cases. It turns out that the Kompaneetz equation part of Eq. (2.0.1); i.e., excluding the last two terms, under certain assumed circumstances has simple, analytical, diffusion equation type solutions [10, 12, 19].

The approximations applied by Kompaneetz [12], Illarionov and Sunyaev [10], and Silk and Stebbins [19] all assume $f(q, y) \ll 1$, so that the coefficients of the Kompaneetz equation are linear in f. Under these circumstances, Kompaneetz showed that the photon spectrum approaches the equilibrium at least as fast as e^{-2y}. Illarionov and Sunyaev [10] and Silk and Stebbins [19] considered small deviations from the Planck spectrum, and for $q \ll 1$ their results agree with those of Kompaneetz (the *Wien limit*), while for $q \gg 1$ it behaves like a growing disturbance which moves to the lower frequencies.

When the problem is solved numerically, these approximations are not necessary, but the results obtained using these approximations can be compared with the numerical results in the regions where the approximation conditions are satisfied. Therefore the aim is to numerically investigate the behaviour of the Kompaneetz equation with modifications allowing for the expanding universe scenario and, also, the behaviour of this equation when either or both of the entropy generating terms such as bremsstrahlung and double Compton terms are included.

The PDE (2.0.1) is of a fairly general format, but a

number of features may be identified, and these are helpful in formulating a satisfactory difference scheme for solving the problem in a fairly general case.

First, the left-hand side of (2.0.1) is basically a form of a wave equation in the $(y, \ln q)$ space. Second, the first right-hand term of (2.0.1) (if we ignore the $(1+f)$ factor, may be transformed into a more conventional diffusion format in $\ln q$ space [19]. Third, the last term of (2.0.1) is a source/sink term.

The difference scheme that solves (2.0.1), without the source/sink term, must conserve both $\phi^3 J_2$ and $\phi^4 J_3$ (defined by Eqs. (2.0.4), (2.0.5), and (3.4.7)) proportional to the photon number and the photon energy of the universe. This suggests that a conservative scheme should be used for this part of the equation, with possibly some special care taken with the non-linear factor $f(1+f)$. The wave component may be handled using a Lagrangian approach by replacing the left-hand side of (2.0.1) with a full drivative with respect to some variable $y' = y - c_s^{-1} \ln q$.

In the following section we describe difference schemes which may be used for the solution of (2.0.1), construct the conservative difference scheme, and discuss its implementation.

3. NUMERICAL TECHNIQUES

Methods of solving differential equations of various types have been studied extensively in the past, even before these methods could be applied in practical cases as a result of improved access to large scale computing. The techniques of analysis of various types of differential equations may be found in a range of texts [9, 11, 15–17]. Despite extensive work done in this area, it soon becomes clear that there are no universal methods of solving partial differential equations, and some insight into the problem and properties of the function(s) investigated is necessary, when the coefficients are not simple, and in particular, if there are non-linear terms involved.

The problem requiring the numerical solution of the kinetic equation (2.0.1) is an initial value—boundary condition problem, and the numerical solution forms a time series, so the problem will be analysed in terms of finite difference schemes which appear to be the most appropriate in this case. This is only one of several major classes of approaches to solving partial differential equations numerically; other techniques include finite element analysis and spectral techniques.

Spectral analysis techniques effectively involve obtaining the solution of the equivalent transform problem, e.g., obtaining the solution of the Fourier transform of the original differential equation. Advances in algorithms used in calculations of Fourier transforms, i.e., the fast Fourier transform techniques, or the FFTs [18], make these techniques particularly useful, but other transforms may be more suited to solving any given problem. Spectral techniques, under certain circumstances, are more efficient and more accurate than the finite difference techniques, but this need not be true in general [5]. Accuracy and efficiency of the procedures is usually dependent on the application of the appropriate set of characteristic functions.

For our problem it is necessary to know the phase space density function $f(q, y)$ at various intermediate stages in order to solve the ordinary differential equation obtained from the equation for entropy

$$S = \frac{R^3}{T} [P + \rho]_{\text{int}} \tag{3.0.1}$$

(see Ref. [21, p. 533]), where S is the entropy to within an additive constant, R is a distance scale, P is the gas pressure, and ρ is the energy density. Consequently, under these circumstances, spectral techniques are neither practical nor efficient. It therefore becomes clear that the most practical approach to the problem is using some finite difference scheme (provided it works), and this will be discussed in the next section. In fact, obtaining the correct finite difference scheme is one of the most important elements in the process of numerically solving a partial differential equation, and this point will be demonstrated here in the case of the Kompaneetz equation.

3.1. *Finite Difference Schemes*

The formulation of a finite difference scheme essentially consists of applying the Taylor expansion series in some form, to approximate the derivatives in the partial differential equation, e.g., by expanding $f(q + \Delta q, y)$ and $f(q, y + \Delta y)$.

Let us first define a notation convention where the superscript refers to the temporal variable y and the subscript refers to the spatial variable q; then $F(q_0 + m\,\Delta q, y_0 + n\,\Delta y) = F_m^n$, y_0 and q_0 are some reference values of y and q, respectively. Since we are dealing with an initial value problem, all points with superscript 0 are known at any given iteration, so the aim is to develop and solve an expression for the value of the function at the next time step.

Let us now consider a general parabolic partial differential equation with variable coefficients

$$\frac{\partial f}{\partial y} = a \frac{\partial^2 f}{\partial q^2} + b \frac{\partial f}{\partial q} + cf, \tag{3.1.1}$$

where a, b, and c are functions of y and q. We now wish to write down an approximation to (3.1.1). The most general estimate for the derivative $\partial f/\partial y$, somewhere in the range $(y, y + \Delta y)$, would be some linear contribution of

derivatives calculated at time level 0 and 1. Thus the difference equation may easily be shown to have the form

$$\alpha\left(\frac{a_0^1}{(\Delta q)^2}+\frac{b_0^1}{2\,\Delta q}\right)f_1^1+\left(\frac{\alpha c_0^1-2\alpha a_0^1}{(\Delta q)^2}+\frac{1}{\Delta y}\right)f_0^1$$
$$+\alpha\left(\frac{a_0^1}{(\Delta q)^2}-\frac{b_0^1}{2\,\Delta q}\right)f_{-1}^1$$
$$=(\alpha-1)\left(\frac{a_0^0}{(\Delta q)^2}+\frac{b_0^0}{2\,\Delta q}\right)f_1^0$$
$$+(\alpha-1)\left(\frac{a_0^0}{(\Delta q)^2}-\frac{b_0^0}{2\,\Delta q}\right)f_{-1}^0$$
$$+\left((\alpha-1)\,c_0^0-\frac{2(\alpha-1)\,a_0^0}{(\Delta q)^2}-\frac{1}{\Delta y}\right)f_0^0$$
$$+O(\Delta y)+O(\Delta q^2). \tag{3.1.2}$$

This is of the form

$$d_1 f_1^1-d_0 f_0^1+d_{-1}f_{-1}^1\approx e^0, \tag{3.1.3}$$

where the d's here are just the coefficients of the function at an appropriate grid point at time $t+\Delta t$ (i.e., the superscript 1 is implied). It is now required to solve Eq. (3.1.3) for f^1. The $\alpha=0$ case is an *explicit* form of the difference equation commonly referred to as *forward in time and centered in space* (FTCS), whereas setting $\alpha=1$ results in an *implicit* scheme. In both of these schemes, the accuracy is of the order Δy in the temporal variable and $(\Delta q)^2$ in the spatial variable. In case of $\alpha=\frac{1}{2}$, the scheme (3.1.2) is accurate to $(\Delta y)^2$. This may be shown by expanding the Taylor series about the point $(q+\frac{1}{2}\Delta q,\ y+\frac{1}{2}\Delta y)$ and could be referred to as centered in time and centered in space, but it is more commonly known as the Crank–Nicholson scheme and is also *implicit*. The results obtained using the implicit and the Crank–Nicholson type scheme in the case of a version of the Kompaneetz equation will be presented later, following the stability analysis of the scheme (3.1.2). We mention here in passing that a further modification of Eq. (3.1.2) is possible in order to achieve the accuracy of order $(\Delta q)^4$, but as will soon become obvious, this is not of primary concern in view of the stability considerations, so it will not be discussed here in detail.[1] This covers the set of general schemes involving two time levels [16, pp. 189–191], but there still exists another class of such schemes involving more than two time levels. These will not be discussed here also since, as it turns out, they complicate the system unnecessarily. Richtmyer and Morton discuss and present algorithms for a wide range of such schemes in the case of a simple heat or diffusion equation, Eq. (3.1.1) with $b=c=0$. In the general case a number of schemes or variations may be developed for the purpose of accommodating the variability of these coefficients.

A very general and systematic method of developing difference schemes to the accuracy of an arbitrary order in the spatial part, may be obtained by the method of Padé approximants (see, e.g., Ref. [4, p. 257]). In this technique, the spatial part of Eq. (3.1.1) is replaced by the lowest order difference scheme approximation (3.1.2), ignoring the time level for now. This results in a matrix expression of the form

$$\frac{df}{dy}=-\hat{C}f, \tag{3.1.4}$$

where f is the vector of the spatial function values and $\hat{C}$ is a tri-diagonal matrix approximating the spatial derivative. For small time intervals Δy, the solution to this is

$$f(y+\Delta y)=e^{-\hat{C}\,\Delta y}f(y). \tag{3.1.5}$$

The idea now is to approximate the exponential function by the Padé approximant

$$r_{m,n}(u)=[q_n(u)]^{-1}\,p_m(u),$$

where p_m and q_n are polynomials in u of degree m and n, respectively, in the matrix $u=\hat{C}\,\Delta y$, so that (3.1.5) becomes

$$f(y+\Delta y)=q^{-1}(\hat{C}\,\Delta y)\,p(\hat{C}\,\Delta y)\,f(y). \tag{3.1.6}$$

In the case $m=n=1$ the polynomials are $p(u)=1-u/2$ and $q(u)=1+u/2$, which gives the Crank–Nicholson type scheme.

3.2. *Stability Considerations*

Stability analysis of finite difference equations is often a difficult procedure due to the fact that in order to find the analytical solution of the difference equation it is necessary to solve a partial differential equation of higher order than the original one.

There are a number of techniques of estimating the stability of the difference equation approximation. One of them consists of investigating the evolution of some error quantity $\varepsilon^n=f^n-f^{*n}$, where f^n is the calculated value, while f^n is the exact solution of the difference equation time series. The problem is stable if $\varepsilon^{n+1}/\varepsilon^n\to 0$ as $n\to\infty$, or, alternatively, $|\varepsilon^{n+1}/\varepsilon^n|<1$. In addition, if $\varepsilon^{n+1}/\varepsilon^n\geqslant 0$ for all n, then no *overshoot* occurs in the solution (i.e., the solution will *not* oscillate).

One of the most popular stability analysis techniques is that due to *von Neumann*, originally developed around 1944, but formally first published by Crank and Nicholson

[1] See Ref. [16] for the procedure of obtaining higher order corrections to various types of partial differential equations.

[3] and later by von Neumann and Richtmyer [20]. The aim of this technique is to evaluate the amplification factor G (for some wavelength associated with the grid spacing), where

$$V^{n+1} = GV^n \tag{3.2.1}$$

and V^n is the magnitude of some spectral component at time level n. Now each solution component ζ_m^n may be written in the form

$$\zeta_m^n = V^n e^{imk\,\Delta q} \tag{3.2.2}$$

where k is the smallest wavenumber associated with the grid and the extent has been normalised to $q_{\max} = 1$ [17]. If we now let

$$\theta = k\,\Delta q \tag{3.2.3}$$

then (3.2.2) becomes

$$\zeta_m^n = V^n e^{im\theta} \tag{3.2.4}$$

If the number of grid points in the spatial part of the grid is M, then

$$\theta_{\min} = \frac{\pi}{M-1} \tag{3.2.5a}$$

$$\theta_{\max} = \pi. \tag{3.2.5b}$$

Substitution from (3.2.4) into the difference equation results in the amplification factor G (Eq. (3.2.1)). This may then be analysed for stability, which requires that $|G| \leqslant 1$ for all possible values of θ. We shall apply this technique, in the remainder of this paper, to analyse the stability of the various numerical approaches.

3.3. *Metric Transformation*

In the Kompaneetz equation the coefficient $a = q^2$ indicates a possible singularity in f at the origin. Furthermore, at the origin $q^2 f = 0$ forms the boundary condition from the particle flux considerations [12] and the behaviour of the function near the origin is of major interest because of the level of permissible distortion in the phase space density function f. This means that f should be considered on a logarithmic scale, at least at the lower end of the spectrum. At that end, however, the boundary condition is rather awkward to implement on a logarithmic scale [14], so in order to avoid such difficulties, a metric transformation is applied to the spatial domain

$$q = \beta(e^u - 1), \tag{3.3.1}$$

where β is some small arbitrary number, approximately $q_{\min}/\Delta u$, as discussed further below and u is the new spatial variable. In addition, substituting for f

$$f = F/q^2, \tag{3.3.2}$$

we obtain

$$\frac{\partial F}{\partial y} = \left(\frac{q}{q+\beta}\right)^2 \frac{\partial^2 F}{\partial u^2} + \left(\frac{q}{q+\beta}\right)\left(q - \frac{q}{q+\beta} + 2F/q + \frac{d\psi}{dy}\right)\frac{\partial F}{\partial u} + 2\left(q - 1 - \frac{d\psi}{dy}\right)F \tag{3.3.3}$$

which becomes the new equation to be analysed, and it is this equation, one hopes, that the difference equation will represent.

3.4. *Stability of the Approximation*

We now compare Eq. (3.3.3) with (3.1.1) and obtain

$$a = \left(\frac{q}{q+\beta}\right)^2; \tag{3.4.1a}$$

note that $a \to 0$ as $q \to 0$ and $a \to 1$ as $q \to \infty$,[2]

$$b = \left(\frac{q}{q+\beta}\right)\left(q - \frac{q}{q+\beta} + 2F/q + \frac{d\psi}{dy}\right) \tag{3.4.1b}$$

and

$$c = 2\left(q - 1 - \frac{d\psi}{dy}\right). \tag{3.4.1c}$$

Assuming that $d\psi/dy$ is a slowly varying function of y, we substitute Eqs. (3.4.1) into the amplification factor equation for G (3.2.1) and obtain

$$G = \frac{1 - (1-\alpha)(A - iB)}{1 + \alpha(A - iB)}, \tag{3.4.2}$$

where

$$A = 2\,\Delta y\left(2\gamma^2 \sin^2\frac{1}{2}\theta - 1 + q - \frac{d\psi}{dy}\right) \tag{3.4.3a}$$

$$B = \Delta y \gamma \sin\theta\left(q - \frac{q}{q+\beta} + \frac{d\psi}{dy} + 2\frac{F}{q}\right) \tag{3.4.3b}$$

$$\gamma = \frac{q}{(q+\beta)\,\Delta u}. \tag{3.4.3c}$$

[2] The potential problem near $q = 0$ may not be too serious if only a few points occur in the region where $q \ll \beta$.

The quantities A, B, and γ are all real, so that in the case $\alpha = \frac{1}{2}$ the condition $|G| \leqslant 1$ is satisfied by the requirement $A \geqslant 0$. This case is important because when this condition is satisfied, the difference equation approximation error is $o(\Delta y)^2$ rather than $O(\Delta y)$. This scheme is likely to have some problems when $d\psi/dy > -1$ and q is small, since $\gamma^2 \sin^2 \frac{1}{2}\theta$ cannot be guaranteed to be large enough so that $A \geqslant 0$ at all times. This in fact may be observed in the estimated value of $d\psi/dy$ at each step, which tends to overshoot somewhat when it is significantly greater than zero, although it still converges in the end despite this problem, when its value drops sufficiently low.

Let us now examine the coefficients of the difference equation (3.1.3) in the case of Kompaneetz equation. Here

$$d_{-1} = \alpha \left\{ \left(\frac{q}{q+\beta} \right)^2 \frac{1}{(\Delta u)^2} - \frac{1}{2\,\Delta u} \left(\frac{q}{q+\beta} \right) \times \left(q - \frac{q}{q+\beta} + 2F/q + \frac{d\psi}{dy} \right) \right\} \tag{3.4.4a}$$

$$d_1 = \alpha \left\{ \left(\frac{q}{q+\beta} \right)^2 \frac{1}{(\Delta u)^2} + \frac{1}{2\,\Delta u} \left(\frac{q}{q+\beta} \right) \times \left(q - \frac{q}{q+\beta} + 2F/q + \frac{d\psi}{dy} \right) \right\} \tag{3.4.4b}$$

$$d_0 = \frac{1}{\Delta y} + \frac{2\alpha}{(\Delta u)^2} \left(\frac{q}{q+\beta} \right)^2 - 2\alpha \left(q - 1 - \frac{d\psi}{dy} \right) \tag{3.4.4c}$$

form a *tri-diagonal* system of the form

$$df = e, \tag{3.4.5}$$

where d is an $M \times M$ matrix with only super-diagonal, diagonal, and sub-diagonal elements non-zero, while f and e are vectors with M components each. The advantage of such a scheme is that the solution time only grows linearly with M as opposed to approximately M^2 in the case of inverting an $M \times M$ matrix. The round off errors (see Ref. [17, Appendix A]) for such a scheme will be kept low, provided

$$d_{-1} + d_1 < d_0, \tag{3.4.6}$$

where d_{-1}, d_0, and $d_1 > 0$. It is quite obvious from Eqs. (3.4.4) that the condition (3.4.6) is *not* automatically satisfied. In fact, large $d\psi/dy$ again could lead to problems, but this time just the fact that it is large in amplitude could force the coefficients d_1 and d_{-1} to change sign and also violate the condition (3.4.6). To avoid this problem it is necessary to keep Δu small. In practice it turns out that the condition (3.4.6) is sufficient but need *not* be satisfied everywhere for the system to be stable. In fact it turns out that d_i need not be all positive, provided d_0 is largest in magnitude and the condition (3.4.6) is satisfied. It will be shown later that the system is stable if we remove the factor $1/\Delta y$ from d_0 and condition (3.4.6) is satisfied with d_0 dominating (i.e., it is larger in magnitude than either of) the other coefficients. In other words, one of d_1 and d_{-1} may become negative, provided it is not large in magnitude. To ensure diagonal domination in (3.4.4), it is also necessary to keep Δy sufficiently small to compensate for a large negative $d\psi/dy$. In fact it is clear that the round-off errors will *not* be contained for small q, since there the coefficient of $(\Delta u)^{-2}$ becomes very small, resulting in condition (3.4.6) being satisfied only provided that Δy is very small. It should be noted that q in Eqs. (3.4.4a)–(3.4.4b) may cause problems as well, since it has large values at the upper end of its domain, so that it may dominate (3.4.4). In fact this does happen when the initial spectrum is strongly distorted in that region. One way of getting around this problem is to make Δu very small indeed, but this has the disadvantage of leading to an unacceptable large value of M, the size of the tri-diagonal matrix. Another alternative is to use a somewhat different metric transformation in place of Eq. (3.3.1). An appropriate metric would be one that does not compress the region of large q as strongly as the log function, so that the grid interval in terms of q is smaller in the region in question. A third alternative is to use the idea proposed by Lightman [14], where the domain of q is divided into two or more regions with different values of Δu. In this case arbitrary decisions must be made to decide on the resolution and number of intervals to be used. Clearly the former procedure is more systematic, provided a suitable transformation that does not result in undue complications may be found, but either will work if applied correctly.

The difference equation "representing" the Kompaneetz equation is solved numerically *without* the additional modifications mentioned above, for several values of M (or Δu) and for variable and constant (small) Δy attempting to keep the problems mentioned above to a minimum. The result of these calculations, for the case of $\alpha = \frac{1}{2}$, is shown in Fig. 1. The plot of the dimensionless chemical potential (ξ_{ass}) is quite enlightening here. Physically, the phase space density function tends to the equilibrium and, because the number of particles is conserved, this quantity, which is calculated from the expected equilibrium spectrum, should remain fixed during the calculations. This clearly is not so in Fig. 1, and so it demonstrates the departures of the difference equation solutions from the solutions of the differential equation (see Ref. [16, Section 1.6). It may be immediately verified that this departure or *drift* away from the expected value is of the order $(\Delta u)^2$ by comparing the slopes of the $M = 500$ and $M = 2000$ and $\alpha = 1, \frac{1}{2}$ curves [6]. It is also clear that the error is $O(\Delta y)$, since the effect due to varying Δy is very small when comparing with the fixed Δy case. Clearly this is the fundamental problem here and

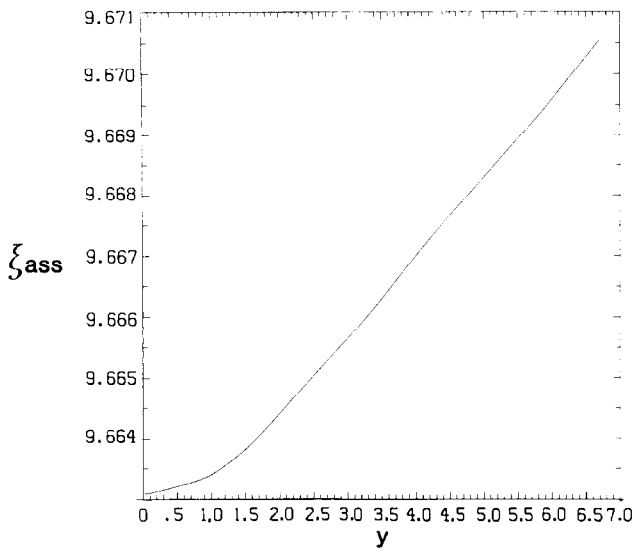

FIG. 1. Sample results for the non-conservative model of the Kompaneetz equation in the context of expanding universe equations (3.1.2) and (3.3.2) of GM, described in this paper for a Crank–Nicholson type scheme ($\alpha = \frac{1}{2}$), keeping the temporal variable increments fixed. The physical model parameters for all the figures, unless otherwise specified, are $m_H = 1.0$ MeV, $m_L = 0$ eV, the decay temperature $T_D = 1$ KeV, branching ratio for electromagnetic decays $B = 0$, $\eta = 10^{-10}$, the number of light neutrino species $N_l = 2$, and the number of heavy neutrino species $N_H = 1$. The assymptotic value of the chemical potential of radiation ϕ_{ass} is given as a function of optical depth y evaluated by solving the equations

$$J_3(t)\,\phi^4 = \frac{g_\gamma}{2\pi^2}\,\phi_{ass}^4 \int_0^\infty (e^{q-\xi_{ass}} - 1)^{-1}\, q^3\, dq,$$

$$J_2(t)\,\phi^3 = \frac{g_\gamma}{2\pi^2}\,\phi_{ass}^3 \int_0^\infty (e^{q-\xi_{ass}} - 1)^{-1}\, q^2\, dq,$$

at each iteration step.

consequently this approach is unsatisfactory. It is therefore necessary to find an alternative approach by identifying the root of the problem in the original difference equation, if reasonable results are to be obtained.

A brief analysis of the situation shows that this *drift* is occurring while $d\psi/dy$ is nearly zero. Furthermore, it is noted that the photon number density–energy density relationships indicate that either energy density or photon number density or *both* are not conserved during the calculations. Mathematically, if we define moments of the phase space density function f to be

$$J_l = \int_0^\infty f q^l\, dq, \tag{3.4.7}$$

then $\phi^4 J_3$ is proportional to the total energy of the universe and is time independent (see Eq. (3.1.3) of GM). This may be derived from of the modified Kompaneetz equation $Df = C_c[f]$. Clearly something has gone wrong while formulating the difference equation, and a new approach needs to be found.

It may also be appropriate to mention here that schemes of a similar type, but higher order than that used above, tend to be less stable than the above scheme and consequently were rejected following a brief evaluation of several cases. In any case, even if a stable scheme was found, it is unlikely that the drift in particle numbers would be reduced sufficiently to make the scheme consistent.

3.5. *A Conservative Scheme Approach*

As has been noted, a brute strength approach does not guarantee success, so in order to achieve the expected conservative behaviour, it is essential to build the difference equation with this consideration in mind. The general technique of obtaining conservative differencing schemes in multidimensional systems has been described by Arakawa and Mintz [1], including the simple argument to follow, that applies in the one-dimensional case discussed here. It should be noted, however, that it does not automatically guarantee an accurate solution (see, e.g., Ref. [1, Example 3–3, p. 368]), hence consideration of the former approach may be justified.

First, it is necessary to define the conserved quantity. In fact there are two conservation requirements. The first one is due to conservation of energy and requires that the third moment of Eq. (2.0.1) vanish (i.e., the total energy of the system is invariant) and that the second conserved quantity relates to the number of photons. This second quantity is only conserved in entropy conserving single Compton collisions requiring that the second moment of the Kompaneetz term of Eq. (2.0.1) vanish.

To establish a conservative difference scheme it has to be set up in some control region as a difference of a function at the boundaries [1]

$$\frac{d\chi_n}{dy} = A_n - A_{n-1} \tag{3.5.1}$$

so that

$$\frac{dX}{dy} = \sum_n \frac{d\chi_n}{dy} = A_N - A_1 \tag{3.5.2}$$

and therefore dX/dy depends purely on the boundary conditions of function A. If these vanish then the quantity X is conserved numerically. The question now becomes: what are these quantities χ_n and A_n? In the context of flow of some quantity μ we may look at χ_n as the amount of μ within the region bounded by points $u - \Delta u/2$ and $u + \Delta u/2$, while A_{n-1} and A_n may be considered as the flux of μ in and out of the region at the boundaries [17]. Ideally we would

like the quantity μ to be the energy and χ_n to be the energy density of particles with momentum in the range corresponding to the interval defined above; unfortunately Eq. (2.0.1) is not easily transformed into the appropriate form. The other alternative is to try to conserve the photon number, where it should be conserved, hoping that it will bring the error down to a reasonable level, and later try to refine the scheme to improve the accuracy even further.

A brief examination of the Kompaneetz equation indicates that it is already in the form required by (3.5.1) because of the method of derivation. Therefore the quantities required, ignoring the entropy generating terms initially, may be defined as

$$\chi_n = \int_{u_n - \Delta u/2}^{u_n + \Delta u/2} Df q^2 \left(\frac{dq}{du}\right) du = \int_{u_n - \Delta u/2}^{u_n + \Delta u/2} \frac{d}{dq} q^4 \left\{\frac{df}{dq} + f(1+f)\right\}\left(\frac{dq}{du}\right) du; \quad (3.5.3)$$

this gives the function A as

$$A(q) = q^4 \left\{\frac{df}{dq} + f(1+f)\right\}. \quad (3.5.4)$$

The integral over the region $[q_n - \frac{1}{2}\Delta q, q_n + \frac{1}{2}\Delta q]$ of the

$$\int_{q_n - 1/2}^{q_n + 1/2} \left(\frac{\partial f}{\partial y} - \frac{d\psi}{dy} q \frac{\partial f}{\partial q}\right) q^2 \, dq = \left[\frac{\partial}{\partial y}(q_n^2 f_n) - \frac{d\psi}{dy} q_n^3 \left(\frac{\partial f}{\partial q}\right)_n\right]\left(\frac{dq}{du}\right)_n \Delta u \quad (3.5.5)$$

and the integral of the entropy generating terms by

$$\int_{q_n - 1/2}^{q_n + 1/2} K(q)[1 - (e^q - 1)f] \, q^2 \, dq = K(q_n)[1 - (e^{q_n} - 1)f_n]\left(\frac{dq}{du}\right)_n q_n^2 \, \Delta u, \quad (3.5.6)$$

where the derivative $(\partial f/\partial q)_n$ is approximated by

$$\left(\frac{\partial f}{\partial q}\right)_n = \frac{f_{n+1/2} - f_{n-1/2}}{\Delta u}\left(\frac{du}{dq}\right)_n. \quad (3.5.7)$$

If we now also set $f = F/q^2$, so that the new function F has a boundary condition zero at the origin, then the conservative part becomes

$$\int_{u_n - \Delta u/2}^{u_n + \Delta u/2} \frac{d}{dq} q^4 \left\{\frac{df}{dq} + f(1+f)\right\}\left(\frac{dq}{du}\right) du = E_{n+1/2} - E_{n-1/2}, \quad (3.5.8a)$$

with

$$E_{n+1/2} = q_{n+1/2}^4 \left(\frac{d(F/q^2)}{dq}\right)_{n+1/2} + F_{n+1/2}(q_{n+1/2}^2 + F_{n+1/2}), \quad (3.5.8b)$$

$$E_{n-1/2} = q_{n-1/2}^4 \left(\frac{d(F/q^2)}{dq}\right)_{n-1/2} + F_{n-1/2}(q_{n-1/2}^2 + F_{n-1/2}), \quad (3.5.8c)$$

$$\left(\frac{\partial(F/q^2)}{\partial q}\right)_{n+1/2} = \frac{F_{n+1}/q_{n+1}^2 - F_n/q_n^2}{\Delta u}\left(\frac{dq}{du}\right)_{n+1/2}, \quad (3.5.8d)$$

$$\left(\frac{\partial(F/q^2)}{\partial q}\right)_{n-1/2} = \frac{F_n/q_n^2 - F_{n-1}/q_{n-1}^2}{\Delta u}\left(\frac{dq}{du}\right)_{n-1/2}, \quad (3.5.8e)$$

$$F_{n+1/2} = (F_{n+1} + F_n)/2, \quad (3.5.8f)$$

$$F_{n-1/2} = (F_n + F_{n-1})/2. \quad (3.5.8g)$$

Assembling everything together, the expression obtained is

$$\left[\frac{\partial F_n}{\partial y} - \frac{d\psi}{dy} q_n^3 \left(\frac{F_{n+1/2}/q_{n+1/2}^2 - F_{n-1/2}/q_{n-1/2}^2}{\Delta u (dq/du)_n}\right)\right]\left(\frac{dq}{du}\right)_n \Delta u$$

$$\left[\left(\frac{d(F/q^2)}{dq} \right)_{n+1/2} \right]$$

$$- \left[q_{n-1/2}^4 \left(\frac{d(F/q^2)}{dq}\right)_{n-1/2} + F_{n-1/2}(q_{n-1/2}^2 + F_{n-1/2}) \right]$$

$$+ K(q_n)[q_n^2 - (e^{q_n} - 1)F_n]\left(\frac{dq}{du}\right)_n \Delta u \quad (3.5.9)$$

and $(dq/du)_i$ is obtained by differentiating the metric transformation, e.g., Eq. (3.3.1). Dividing (3.5.9) by $\Delta q_n = (dq/du)_n \Delta u$ and collecting coefficients of the F_i, this becomes[3]

$$\frac{\partial F_n}{\partial y} = a_n F_{n-1} + b_n F_n + c_n F_{n+1} + d_n \quad (3.5.10a)$$

and

$$a_n = \left(\frac{du}{dq}\right)_n \left[-\frac{d\psi}{dy} \frac{q_n^3}{2q_{n-1/2}^2 \Delta u} + \frac{q_{n-1/2}^4}{q_{n-1}^2 (\Delta u)^2}\left(\frac{du}{dq}\right)_{n-1/2} - \frac{q_{n-1/2}^2 + F_{n-1/2}}{2\Delta u} \right] \quad (3.5.10b)$$

[3] This scheme is very similar to that of Lightman (Ref. [14]), except that he considered the system interacting with electrons at fixed temperature so it could be assumed that $d\psi/dy \equiv 0$. This expression was, however, derived independently before he communicated his scheme to us.

$$b_n = \left(\frac{du}{dq}\right)_n \left[\left(\frac{d\psi}{dy}\right) q_n^3 \frac{1/q_{n+1/2}^2 - 1/q_{n-1/2}^2}{2\,\Delta u} - \frac{q_{n+1/2}^4 (du/dq)_{n+1/2} + q_{n-1/2}^4 (du/dq)_{n-1/2}}{q_n^2 (\Delta u)^2} + \frac{(q_{n+1/2}^2 + F_{n+1/2}) - (q_{n-1/2}^2 + F_{n-1/2})}{2\,\Delta u}\right] - K(q_n)(e^{q_n} - 1) \tag{3.5.10c}$$

$$c_n = \left(\frac{du}{dq}\right)_n \left[\frac{d\psi}{dy} \frac{q_n^3}{2q_{n+1/2}^2\,\Delta u} + \frac{q_{n+1/2}^4}{q_{n+1}^2 (\Delta u)^2} \left(\frac{du}{dq}\right)_{n+1/2} + \frac{q_{n+1/2}^2 + F_{n+1/2}}{2\,\Delta u}\right] \tag{3.5.10d}$$

$$d_n = q_n^2 K(q_n). \tag{3.5.10e}$$

Note that in this scheme the value of $d\psi/dy$ is used explicitly. Ideally this should be obtained by solving the system (3.5.10) simultaneously with the expression for conservation of energy

$$\sum_n \left(\frac{du}{dq}\right)_n \Delta u \left\{\frac{\partial}{\partial y}(q_n F_n) - \frac{1}{2}\left(\frac{d\psi}{dy}\right) q_n^4 [F_{n+1}/q_{n+1/2}^2 + F_n(1/q_{n+1/2}^2 - 1/q_{n-1/2}^2) - F_{n-1}/q_{n-1/2}^2]\right\} \equiv 0. \tag{3.5.11}$$

Unfortunately the insertion of Eq. (3.5.11) into the system (3.5.10) destroys its tri-diagonal property and, as a result, would lead to a very large increase in amount of computation required to solve it. It also turns out that Eq. (3.5.11) cannot be solved by simultaneously reducing it, while the tri-diagonal system is being solved, because numerically accurate solution requires pivoting about the dominant term of each column, and this requirement is then not satisfied. This means that the calculations must be based on a value of $d\psi/dy$, estimated using some other procedure.

Equation (3.5.10) may now be set up in the form of the Crank–Nicholson type scheme, with the coefficients a, b, c, and d common to both time levels, so that the expression may be written from (3.5.10a) as

$$\frac{F_n^{i+1} - F_n^i}{\Delta y} = \alpha(a_n F_{n-1}^{i+1} + b_n F_n^{i+1} + c_n F_{n+1}^{i+1}) + (1-\alpha) \times (a_n F_{n-1}^i + b_n F_n^i + c_n F_{n+1}^i) + d_n. \tag{3.5.12}$$

The stability of this scheme is quaranteed, provided $|G| \leqslant 1$, where, using the von Neumann analysis technique, as before,

$$G = \frac{1/\Delta y + d + (\alpha - 1)[2(a+c)\sin^2 \tfrac{1}{2}\theta - (a+b+c) + i(a-c)\sin\theta]}{1/\Delta y + \alpha[2(a+c)\sin^2 \tfrac{1}{2}\theta - (a+b+c) + i(a-c)\sin\theta]}, \tag{3.5.13}$$

where a, b, c, and d are the coefficients a_n, b_n, c_n, and d_n of Eq. (3.5.12). When $d = 0$, the expression (3.5.13) satisfies the von Neumann stability criterion, mentioned earlier, for $\alpha \geqslant \frac{1}{2}$, provided the condition $-b + |a+c| \geqslant 0$; i.e., $a+b+c \leqslant 0$ and $a-b+c \geqslant 0$ are satisfied. Now

$$a+b+c = \left(\frac{du}{dq}\right)_n \left\{q^3 \frac{d\psi}{dy} \frac{1/q_{n+1/2}^2 - 1/q_{n-1/2}^2}{\Delta u} + \frac{(q_{n+1/2}^2 + F_{n+1/2}) - (q_{n-1/2}^2 - F_{n-1/2})}{\Delta u} + \left(q_{n+1/2}^4 \left(\frac{du}{dq}\right)_{n+1/2} \left(\frac{1}{q_{n+1}^2} - \frac{1}{q^2}\right) + q_{n-1/2}^4 \left(\frac{du}{dq}\right)_{n-1/2} \left(\frac{1}{q_{n-1}^2} - \frac{1}{q^2}\right)\right) \Big/ (\Delta u)^2\right\} \tag{3.5.14a}$$

and

$$a-b+c = \left(\frac{du}{dq}\right)_n \left(q_{n+1/2}^4 \left(\frac{du}{dq}\right)_{n+1/2} \left(\frac{1}{q_{n+1}^2} + \frac{1}{q^2}\right) + q_{n-1/2}^4 \left(\frac{du}{dq}\right)_{n-1/2} \left(\frac{1}{q_{n-1}^2} + \frac{1}{q^2}\right)\right) \Big/ (\Delta u)^2 \tag{3.5.14b}$$

and, since q is a monotonically increasing function of u, then $du/dq > 0$, so that clearly the stability condition is *not* automatically satisfied, and the problem becomes worse the more negative $d\psi/dy$ becomes.

The second difficulty with the scheme (3.5.9) is that it includes terms of the type q_n^{-1}, which need to be treated specially on the lower end boundary, since $q_1 = 0$ there. This problem may be avoided by varying the scheme somewhat, so that instead of differentiating $(\partial f/\partial q) = \partial(F/q^2)/\partial q$ as a whole, we expand it, inside the Kompaneetz term only, to give

$$\left(\frac{\partial f}{\partial q}\right) = \frac{1}{q^2}\left(\frac{\partial F}{\partial q}\right) - 2F/q^3 \tag{3.5.15}$$

and at the same time we may also integrate the second term of (3.5.5) by parts to give

$$\int_{q_n - 1/2}^{q_n + 1/2} \left(\frac{\partial f}{\partial q}\right) q^3\, dq = (qF)_{n+1/2} - (qF)_{n-1/2} - 3\int_{u_n - \Delta u/2}^{u_n + \Delta u/2} F \frac{dq}{du}\, du \tag{3.5.16}$$

and, consequently, the scheme (3.5.10) may be rewritten as

$$a_n=\left(\frac{du}{dq}\right)_n\left[-\frac{d\psi}{dy}\frac{q_{n-1/2}}{2\,\Delta u}+\frac{q_{n-1/2}^2}{(\Delta u)^2}\left(\frac{du}{dq}\right)_{n-1/2}-\frac{q_{n-1/2}^2-2q_{n-1/2}+F_{n-1/2}}{2\,\Delta u}\right] \tag{3.5.17a}$$

$$b_n=\left(\frac{d\psi}{dy}\right)\left[\left(\frac{du}{dq}\right)_n\frac{q_{n+1/2}-q_{n-1/2}}{2\,\Delta u}-3\right]-\left(\frac{du}{dq}\right)_n\left[\frac{q_{n+1/2}^2(du/dq)_{n+1/2}+q_{n-1/2}^2(du/dq)_{n-1/2}}{(\Delta u)^2}+\frac{\begin{pmatrix}(q_{n+1/2}^2-2q_{n+1/2}+F_{n+1/2})\\-(q_{n-1/2}^2-2q_{n-1/2}+F_{n-1/2})\end{pmatrix}}{2\,\Delta u}\right]-K(q_n)(e^{q_n}-1) \tag{3.5.17b}$$

$$c_n=\left(\frac{du}{dq}\right)_n\left[\frac{d\psi}{dy}\frac{q_{n+1/2}}{2\,\Delta u}+\frac{q_{n+1/2}^2}{(\Delta u)^2}\left(\frac{du}{dq}\right)_{n+1/2}+\frac{q_{n+1/2}^2-2q_{n+1/2}+F_{n+1/2}}{2\,\Delta u}\right] \tag{3.5.17c}$$

$$d_n=q_n^2K(q^2), \tag{3.5.17d}$$

where Eq. (3.5.12) still applies and in the coefficients one must use an explicit form for $F_{n\pm1/2}$:

$$F_{n\pm1/2}=(F_n^i+F_{n\pm1}^i)/2. \tag{3.5.18}$$

This again may be shown to have stability problems similar to those of the scheme (3.5.10), using an equation equivalent to (3.5.14),

$$a+b+c=\frac{d\psi}{dy}\left[\left(\frac{du}{dq}\right)_n\frac{q_{n+1/2}-q_{n-1/2}}{\Delta u}-3\right]+\left(\frac{du}{dq}\right)_n\frac{\begin{matrix}(q_{n+1/2}^2-2q_{n+1/2}+F_{n+1/2})\\-(q_{n-1/2}^2-2q_{n-1/2}+F_{n-1/2})\end{matrix}}{\Delta u} \tag{3.5.19a}$$

and

$$a-b+c=3\frac{d\psi}{dy}\left(\frac{du}{dq}\right)_n+\frac{2}{(\Delta u)^2}\left(\frac{du}{dq}\right)_n\times\left\{q_{n+1/2}^2\left(\frac{du}{dq}\right)_{n+1/2}+q_{n-1/2}^2\left(\frac{du}{dq}\right)_{n-1/2}\right\}. \tag{3.5.19b}$$

Note here that if we ignore the non-linear terms in (3.5.19) then the second term will be negative for $q_n<1$ and positive for $q_n>1$; however, if the function F takes values close to unity in the region $q_n<1$ then the difference of the non-linear terms $F_{n+1/2}-F_{n-1/2}=(F_{n+1}-F_{n-1})/2$ will, in general, be positive in that region and may dominate it, leading to instabilities, as will be demonstrated later. Also note that from Eq. (3.5.16)

$$\int_0^\infty\left(\frac{\partial f}{\partial q}\right)q^3\,dq=-3\sum_n\int_{u_n-\Delta u/2}^{u_n+\Delta u/2}F\frac{dq}{du}\,du=-3\sum_n F_n\left(\frac{dq}{du}\right)\Delta u+O(\Delta u)^2=-3N(1+O(\Delta u)), \tag{3.5.20}$$

where N is defined here to be

$$N=\sum_n F_n\left(\frac{dq}{du}\right)\Delta u$$

and therefore we obtain a numerical relationship equivalent to Eq. (3.1.1) of GM, but this time for photon number density. If $K(q)\equiv0$ is set, then the relation becomes

$$\sum_n\left\{\frac{F_n^{i+1}-F_n^i}{\Delta y}+3\left(\frac{d\psi}{dy}\right)\left(\frac{dq}{du}\right)_n\times[(1-\alpha)F_n^{i+1}+\alpha F_n^i]\right\}\Delta u\approx0. \tag{3.5.21}$$

This type of approach certainly removes the problem of the non-conservative scheme of Section 3.1, but it suffers from the low order instability due to the $d\psi/dy$ term (see Fig. 2). The oscillations can become much more severe

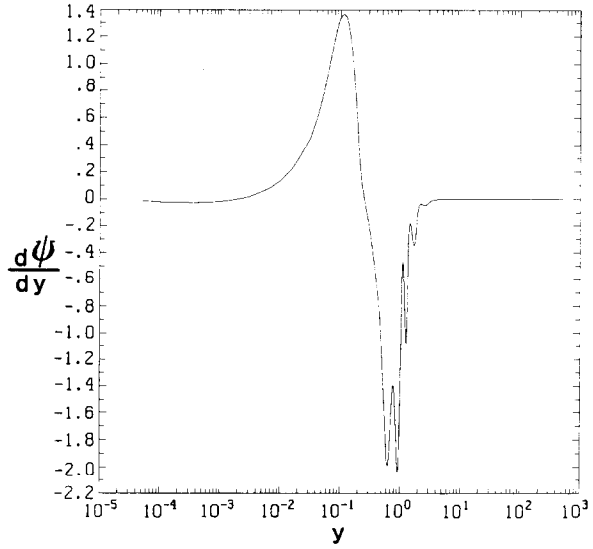

FIG. 2. Evolution curves for $d\psi/dy$ as a function of the optical depth parameter y. Conservative scheme, non-Lagrangian treatment.

under appropriate circumstances. Note also the rather large excursions towards the positive value of the estimated $d\psi/dy$.

To summarise here, it is clear that some improvement has been made in the model, but obviously it is still far from a satisfactory outcome. It is therefore necessary to consider further refinements before reliable results are obtained.

3.6. *Semi-Lagrangian Approach*

The oscillations in the estimated values of $d\psi/dy$ indicate a problem related to that term in the model.

The conservative scheme, developed in the previous section, considers differential segments of particle numbers or the "fluid" being transported in the dimensionless momentum space. This type of approach is referred to as *Eulerian*. In problems involving wave-type forms it may be advantageous to follow the motion of a "single particle of fluid" instead. This is known as the *Lagrangian* viewpoint [17]. In *Lagrangian* schemes a set of "particles" is followed during their evolution. This has the effect of distorting the grid and making it difficult for numerical integration. To avoid such problems a variation on the theme may be introduced by tracking where the "particles" came from at each time step at a fixed grid point, which means that the "particles" followed are *changing* from step to step. This technique is known as the *semi-Lagrangian* approach (see, e.g., Ref. [2]). In case of Eq. (2.0.1) it may, however, be best to use a combination of both approaches, each where it is best suited. To see this, let us write the left side of Eq. (2.0.1) in the form of a full derivative,

$$\hat{D} = \frac{\partial f}{\partial y} - q\frac{d\psi}{dy}\frac{\partial F}{\partial q} = \frac{d}{dy} f(c_s^{-1} \ln q - y), \qquad (3.6.1)$$

where

$$c_s = d\psi/dy,$$

and it is assumed that c_s is either *independent* of y or varies *very slowly* with y. For our purpose, the expression (3.6.1) needs to be modified to become an equation for $F = q^2 f$, so that

$$\frac{\partial F}{\partial y} - q\frac{d\psi}{dy}\frac{\partial F}{\partial q} + 2\frac{d\psi}{dy}F = \frac{dF'}{dy} + 2\frac{d\psi}{dy}F', \qquad (3.6.2)$$

where

$$F' = F(\ln q - c_s y). \qquad (3.6.3)$$

The energy integral of the function F' can then be calculated to be

$$\int_0^\infty qF(\ln q - c_s y)\, dq = e^{2c_s y}\int_0^\infty qF(\ln q)\, dq \qquad (3.6.4)$$

and each term on the right-hand side of (3.6.2) contributes a term like the right-hand side of (3.6.4). Therefore, if the energy is to be conserved in the semi-Lagrangian stage of the integration, the function required is

$$F^*(\ln q) = \kappa F(\ln q - c_s y), \qquad (3.6.5a)$$

where

$$\kappa = \int_0^\infty qF(\ln q,\ y)\, dq \left[\int_0^\infty qF(\ln q - c_s y,\ y)\, dq\right]^{-1}. \qquad (3.6.5b)$$

Note that Eq. (3.6.5) allows us to ignore the second term on the right side of (3.6.2), because it has been absorbed during the integration step of Eq. (3.6.5).

Physically this can be interpreted very simply. As the photons are thermalised and the heating factor ϕ increases, the space q needs to be *rescaled* to keep the spectrum positioned correctly relative to its mean value in q space and at the same time the energy integral must be conserved. The former requirement is satisfied using the transformation

$$\ln q' = \ln q - c_s y \qquad (3.6.6)$$

while the latter is achieved with the application of Eq. (3.6.5). We may now write a modified difference equation approximation of Eq. (2.0.1) by setting $d\psi/dy \equiv 0$ in Eq. (3.5.9), except for one term and replacing $(F_n^{i+1} - F_n^i)/\Delta y$ with $(F_n^{i+1} - F_n^{*i})/\Delta y$, where

$$\begin{aligned} F_n^{*i} &= e^{-2(d\psi/dy)\,\Delta y} F(q_n', y_i) \\ &= F(q_n e^{-(d\psi/dy)\,\Delta y}, y_i) \end{aligned} \qquad (3.6.7)$$

and, consequently, the overall scheme is like that of Lightman [14], except that all the references to the function F at time level i are replaced by F^{*i}. Equations (3.5.17) may now be rewritten as

$$a_n = \left(\frac{du}{dq}\right)_n \left[\frac{q_{n-1/2}^2}{(\Delta u)^2}\left(\frac{du}{dq}\right)_{n-1/2} - \frac{q_{n-1/2}^2 - 2q_{n-1/2} + F_{n-1/2}^*}{2\,\Delta u}\right] \qquad (3.6.8a)$$

$$b_n = \left(\frac{du}{dq}\right)_n \left[-\frac{q_{n+1/2}^2 (du/dq)_{n+1/2} + q_{n-1/2}^2 (du/dq)_{n-1/2}}{(\Delta u)^2} + \frac{(q_{n+1/2}^2 - 2q_{n+1/2} + F_{n+1/2}^*) - (q_{n-1/2}^2 - 2q_{n-1/2} + F_{n-1/2}^*)}{2\,\Delta u}\right] - K(q_n)(e^{q_n} - 1) \qquad (3.6.8b)$$

$$c_n = \left(\frac{du}{dq}\right)_n \left[\frac{q^2_{n+1/2}}{(\Delta u)^2}\left(\frac{du}{dq}\right)_{n+1/2} + \frac{q^2_{n+1/2} - 2q_{n+1/2} + F^*_{n+1/2}}{2\,\Delta u}\right], \tag{3.6.8c}$$

while d_n remains as before. The sum of these coefficients now becomes

$$a + b + c = \left(\frac{du}{dq}\right)_n \frac{(q^2_{n+1/2} - 2q_{n+1/2} + F^*_{n+1/2}) - (q^2_{n-1/2} - 2q_{n-1/2} + F^*_{n-1/2})}{\Delta u} \tag{3.6.9a}$$

and

$$a - b + c = \frac{2}{(\Delta u)^2}\left(\frac{du}{dq}\right)_n \left\{q^2_{n+1/2}\left(\frac{du}{dq}\right)_{n+1/2} + q^2_{n-1/2}\left(\frac{du}{dq}\right)_{n-1/2}\right\} \tag{3.6.9b}$$

and, consequently, the scheme will be stable for all $d\psi/dy$, provided the numerator of Eq. (3.6.9a) is negative. In general, F is an increasing function of u for $u < 1$ and a decreasing function of u for $u > 1$, so that the scheme may become unstable when F is close to 1 due to the non-linear term. This problem does *not* go away if the non-linear term is expanded using Taylor series as in Lightman [14],

$$(F^{i+1}_n)^2 = F^{*i}_n F^{i+1}_n - (F^{*i}_n)^2, \tag{3.6.10}$$

and the typical unstable behaviour is shown in Fig. 3. Lightman's calculations did not suffer from this instability due to the fact that he started with very few photons in the low energy end of the spectrum, while the instability becomes important only when the equilibrium distribution is approached from "above," i.e., when during thermalisation too many photons appear in the low q region. To overcome this problem the non-linear term must be treated in a somewhat different fashion. That is, the coefficients a_n and c_n must be set up, if possible, in such a way that when they are added together, the non-linear term disappears. One way of achieving this is by using the approximation

$$\left(\frac{dF^2}{dq}\right)_n \approx (F^*_{n+1} + F^*_{n-1})\frac{F_{n+1} - F_{n-1}}{2\,\Delta q} \approx F^*_n \frac{F_{n+1} - F_{n-1}}{\Delta u}\left(\frac{du}{dq}\right)_n. \tag{3.6.11}$$

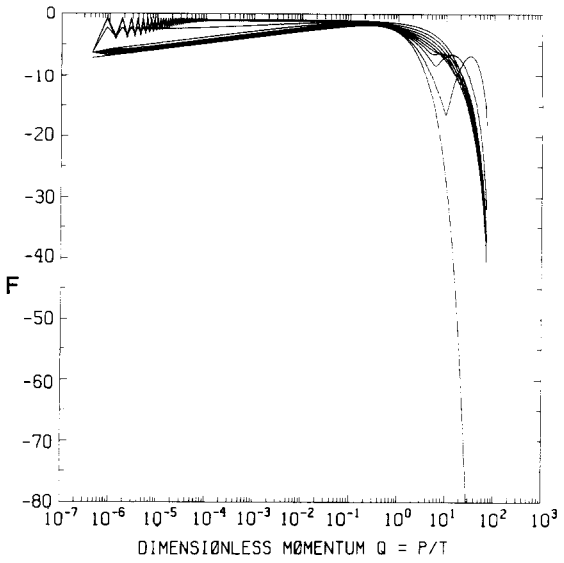

FIG. 3. Time evolution curves of the photon spectrum for the semi-Lagrangian conservative scheme but with the non-linear term treated using the scheme of Lightman [14], for the case of decay temperature T_D of 1 KeV, massive neutrino mass $m_H = 1$ MeV and electromagnetic decay branching ratio $B = 10^{-3}$. Only Kompaneetz and double Compton terms included. Note the instability in the evolving spectrum.

This then gives the final form of the difference scheme with coefficients of (3.5.12) being

$$a_n = \left(\frac{du}{dq}\right)_n \left[\frac{q^2_{n-1/2}}{(\Delta u)^2}\left(\frac{du}{dq}\right)_{n-1/2} - \frac{q^2_{n-1/2} - 2q_{n-1/2} + 2F^*_n}{2\,\Delta u}\right] \tag{3.6.12a}$$

$$b_n = \left(\frac{du}{dq}\right)_n \left[-\frac{q^2_{n+1/2}(du/dq)_{n+1/2} + q^2_{n-1/2}(du/dq)_{n-1/2}}{(\Delta u)^2} + \frac{(q^2_{n+1/2} - 2q_{n+1/2}) - (q^2_{n-1/2} - 2q_{n-1/2})}{2\,\Delta u}\right] - K(q_n)(e^{q_n} - 1) \tag{3.6.12b}$$

$$c_n = \left(\frac{du}{dq}\right)_n \left[\frac{q^2_{n+1/2}}{(\Delta u)^2}\left(\frac{du}{dq}\right)_{n+1/2} + \frac{q^2_{n+1/2} - 2q_{n+1/2} + 2F^*_n}{2\,\Delta u}\right] \tag{3.6.12c}$$

and, for the stability equation,

$$a + b + c = \left(\frac{du}{dq}\right)_n \frac{(q^2_{n+1/2} - 2q_{n+1/2}) - (q^2_{n-1/2} - 2q_{n-1/2})}{\Delta u} \tag{3.6.13a}$$

and

$$a-b+c=\frac{2}{(\Delta u)^2}\left(\frac{du}{dq}\right)_n\left\{q_{n+1/2}^2\left(\frac{du}{dq}\right)_{n+1/2}+q_{n-1/2}^2\left(\frac{du}{dq}\right)_{n-1/2}\right\}. \quad (3.6.13b)$$

In practice, despite the fact that (3.6.13) is positive in some regions, the system is still stable. This is because the amplification factor G is allowed to exceed unity by a term of $O(\Delta y)$ and the system will still remain stable (see, e.g., Ref. [16]) and this is in fact the case here for any local grid region in the spatial coordinate space.

Let us now briefly return to the amplification factor equation (3.5.13). Until now it was assumed that $d=0$ (i.e., only the single Compton or Kompaneetz term was considered). If we now allow $d>0$ then it is obvious from Eq. (3.5.13) that the amplification factor increases and the system becomes less stable. This means that under such circumstances the $\alpha=1$ model is much more likely to remain stable than the $\alpha=\frac{1}{2}$ model and, in fact, the latter does suffer some difficulties resulting in very slow integration in the full model. Evolution curves for various parameters are given in Fig. 4 for the case corresponding to all collision types included and for the $\alpha=1$ model.

At this stage we would like to make several observations. In Fig. 4, y is effectively a measure of time, such that the single Compton effect always becomes important for $y\geqslant 1$, while the source/sink terms, because they are much weaker, do not become important until much later, say $y\approx 5$ in Fig. 4. This means that, up to $y\approx 1$, not a lot of changes are taking place as far as the diagnostic parameters t, T, H, λ, ϕ, $\phi_{\rm ass}$, etc. are concerned (see Figs. 4a–b, d–f); i.e., the behaviour of the system is the same as for a case of a non-expanding, fixed-temperature cavity, without any source/sink terms. The only effect taking place for $y\leqslant 1$ is a rapid change in the spectrum F $(=q^2f)$ see Fig. 4i. The main difference between the spectra for small y, in the case with source/sink terms and those without, is that there is always some small q ($\sim 10^{-6}-10^{-5}$ in Fig. 4i) where the spectrum quickly attains the Planck's spectrum values due to the increasing strength of the source/sink terms with decreasing value of q). For $y>1$, the Wien hump (the region around $q=1$, where f has the Boltzmann's distribution, $f=\exp(-q-\xi)$) is established, and because the source/sink terms for large q are negligible, the hump grows slowly as the particles drift in energy space from the small values at $q\sim 10^{-5}$ to around $q\sim 1$. An interesting and noticeable feature of the spectrum evolution for $y>1$ (in the case examined here, Fig. 4j) is that the Wien hump has practically fixed boundaries; i.e., it is spread over the approximate region $10^{-1}<y<10^2$. Also, because of the behaviour for small y, the spectrum can never become a true Bose–Einstein distribution, even if the values of ξ and $\xi_{\rm ass}$ are practically the same; i.e., the only equilibrium distribution possible, is the Planck spectrum and all other spectra result due to premature decoupling of the system. In the general case, such spectrum consists of three parts: (i) the Planckian segment ($q\leqslant 10^{-4}$ in Fig. 4j); (ii) the Wien hump ($q\geqslant 0.1$ in Fig. 4j); and (iii) the intermediate region that connects the regions (i) and (ii) [10].

3.7. *Stability, Consistency, and Convergence*

Before the results of the model are accepted it is necessary to ensure that it converges onto the required solution. This is done with the aid of the *Lax's equivalence theorem* which states

> Given a properly posed initial-value problem and a finite-difference approximation to it that satisfies the consistency condition, stability is the necessary and sufficient condition for convergence [16].

This means that since the stability of the scheme (3.6.12) has already been established; it is only necessary to ensure that the scheme is consistent, i.e., that the change in the value of the function F goes to zero as $\Delta y\to 0$. Numerically this may be tested by repeating each iteration during the solving procedure, which should reduce the error in the calculated solution, and by running the model with initial conditions consisting of the steady state, in which case only truncation errors should be noticeable.

Both of these tests were done and the results were consistent, producing a Planckian spectrum in $B=0$ case, while the case with iterations repeated two and three times at each time step produced results that were almost completely identical to the single iteration case. The reason for this is that $d\psi/dy$ is relatively small and, so if the step size is very small as well, the effect due to this variable's error becomes negligible and it does not noticeably affect the results. Analytically, the above results mean that if

$$F^{n+1}=\hat{C}(\Delta y)\,F^n \quad (3.7.1a)$$

and

$$\frac{F^{n+1}-F^n}{\Delta y}=\hat{A}F^n, \quad (3.7.1b)$$

then

$$\left|\left(\frac{\hat{C}(\Delta y)-\hat{I}}{\Delta y}-\hat{A}\right)F\right|\to 0 \quad \text{as} \quad \Delta y\to 0, \quad (3.7.1c)$$

where $\hat{I}$ is an identity operator.

Before the discussion of the model is completed, several points should be made here. First, it is quite obvious from Eqs. (3.6.12)–(3.6.13) that this scheme, although stable in the von Neumann analysis sense, violates the truncation error requirements (see Ref. [17, Appendix A], also mentioned briefly in Section 3.4) for sufficiently large values of q. This means that the scheme may suffer from round-off

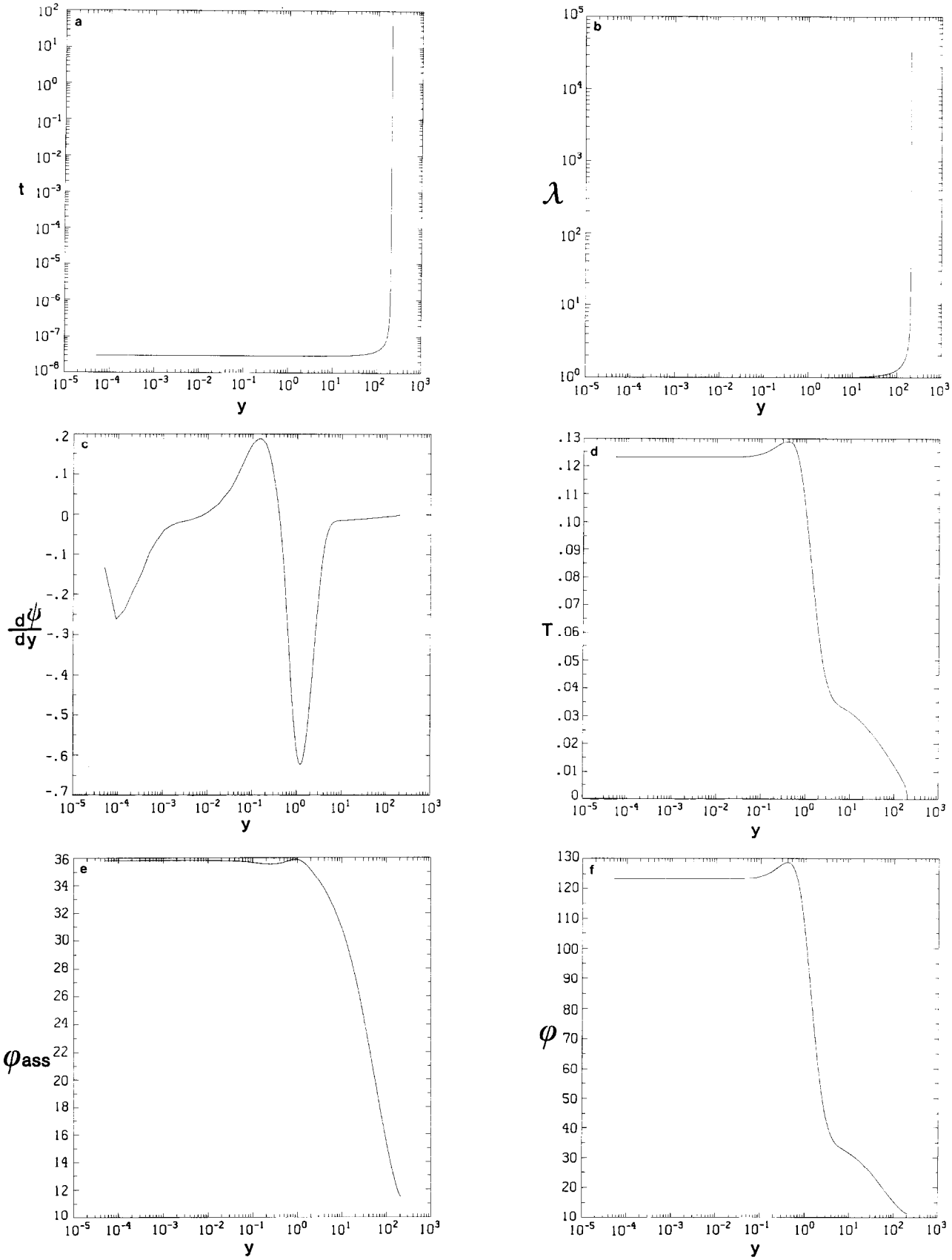

FIG. 4. Evolution curves of representative parameters as a function of the optical depth parameter y for the final version of the model including the semi-Lagrangian and non-linear term modifications of Section 3.6 for the Kompaneetz bremsstrahlung and double Compton terms all included, model $\alpha = 1$: (a) time, (b) the expansion scale parameter λ, (c) $d\psi/dy$, (d) temperature, (e) assymptotic value of ϕ, ϕ_{ass}, (f) heating factor ϕ, (g) step size Δy during integration, (h) Hubble parameter H, and (i)–(j) a sequence of spectra at selected values of y (i) at small values of y, $y \leqslant 1$; (j) at much later time, $y \gg 1$. The figures are reproduced from Ref. [6], where a more complete set for $\alpha = 1/2$ and $\alpha = 1$ may be found.

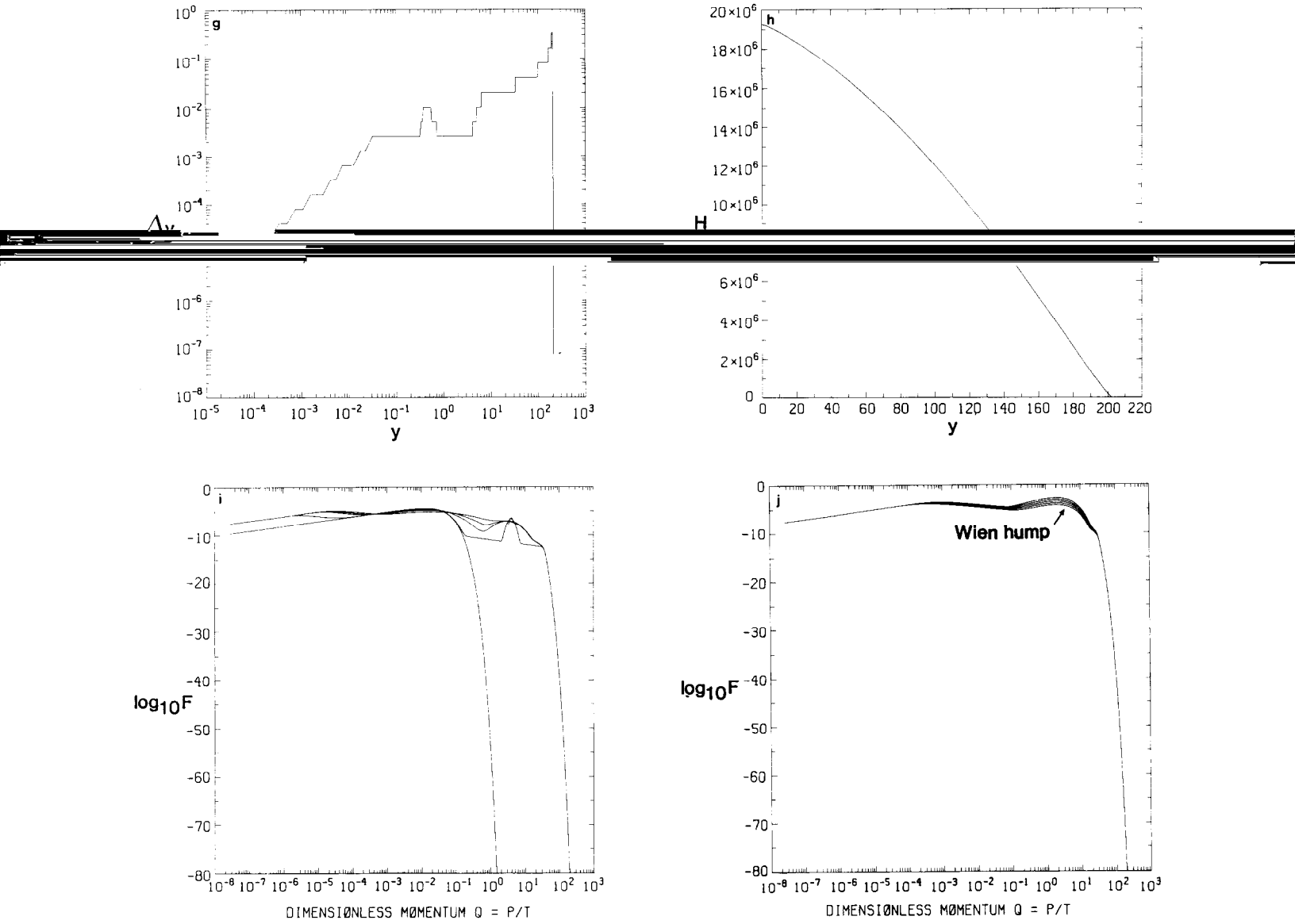

FIG. 4—*Continued*

errors in that region. At the same time it should be noted that in Fig. 4 the temperature curve briefly increases before dropping, and in a number of other curves small deviations are stronger in the corresponding region. This may almost certainly be attributed to the truncation error, which appears, under certain circumstances, in the high end of the photon spectrum in the form of very small and fast oscillations about the zero value.

It is quite easy to show that this temperature rise is a purely numerical effect and not a physical one. The internal energy of the photon gas is given by[4]

$$U = TS, \tag{3.7.2}$$

so that the change in internal energy dU per unit time is given by

$$\frac{dU}{dt} = T\frac{dS}{dt} + S\frac{dT}{dt}. \tag{3.7.3}$$

For the isolated system discussed here $dU/dt \equiv 0$, so rearranging Eq. (3.7.3) leads to the expression

$$\frac{dT}{dt} = -\frac{T}{S}\frac{dS}{dt}. \tag{3.7.4}$$

Now, the second law of thermodynamics requires that $dS/dt \geqslant 0$, which implies that $dT/dt \leqslant 0$, and, consequently,

[4] In the case discussed here, internal energy of electrons and protons is negligible.

the temperature should be monotonically decreasing or constant as a function of opacity, time, or expansion scale.

In order to overcome this problem, the first-order derivative term contribution in Eq. (3.6.12) needs to be absorbed into the semi-Lagrangian term, to ensure that the truncation error remains small everywhere. This may well be possible, but in view of the overall accuracy of scheme (3.6.12) it appears that such a complication is not warranted at this stage.

4. SUMMARY AND A COMMENT ON THE HARDWARE

This article has dealt with the analysis of the Kompaneetz equation with sources and sinks. We have compared a number of implicit differencing schemes in the attempt to solve the above equation. These include non-conservative, conservative, and conservative semi-Lagrangian schemes. The non-conservative scheme exhibits a rather severe drift in the particle number (in the absence of sources and sinks) and consequently cannot be integrated due to this inconsistency. Higher order non-conservative schemes exhibit instabilities, particularly in the region where the spatial variable is small, and consequently were not useful. The second differencing scheme considered was a conservative scheme, similar to that used by Lightman [13–14]. This scheme exhibits several problems not apparent in Lightman's case, because his model describes a non-expanding medium and the phase space density function $f \ll 1$ everywhere initially, which is not true in our case. The latter problem was rectified by a modification of the difference expression for the non-linear term, while the former required the advective term (due to expansion) to be treated using the semi-Lagrangian approach. The final version of the differencing scheme was found to be: (1) consistent, when the pure Planck spectrum is used as the initial conditions, and (2) stable when each of the steps is iterated several times, and this led to identical results to within the rounding errors. This means that the scheme converges, and the limit is found to be the Planck spectrum, or the Bose–Einstein spectrum with vanishing chemical potential ξ.

In the future, a model of this type may be used to compute a constrained fit to the cosmological background radiation with distortions [8, 22], if the data is confirmed and is good enough for such a purpose. At the same time some more attention may be given to the truncation problem mentioned at the end of the previous section, as it occasionally results in slow integration in regions relatively far away, below the constraining curves given in GM. In particular, when the massive neutrinos have a mass close to the 1 MeV limit and the branching ratio is close to unity, the problem is generally most noticeable. This may be important if an attempt is made to fit distorted data.

Finally, before we conclude, we would like to make several comments regarding the hardware and precision of the numeric calculations.

The computation required for the model discussed here was done exclusively on Digital Equipment Corporation's VAX series computers. The initial development of the model was in default precision on these machines, but because the photons injected into the background spectrum can be located over a large region of photon momenta, then it was necessary to extend this precision of calculations beyond this level. The default *double precision* mode on DEC's machines unfortunately does not help because there is no change in the dynamic range of the floating arithmetic so that the *G_FLOATING* mode had to be selected. It should be pointed out that this extended precision does not make a significant difference to the computation results, but only allows extension of the range over which the spectrum may be computed from about $q = 80$ to a value of q exceeding 600. This problem can be further overcome by modifying the code to extend it to the *H_FLOATING* mode which has a floating point range of about $10^{\pm 5000}$, but requires twice the array storage capacity as the *G_FLOATING* mode, which has the dynamic range of approximately $10^{\pm 300}$. It turns out that the machines developed over a decade ago, such as the VAX-11/780, which is often used as a standard for evaluating scientific computing performance, lack part of the hardware of the machines of a more recent vintage such as the micro VAX II or the VAX-11/8650. As a result the model developed in this section has much better runtime characteristics on the latter two machines than on a standard VAX-11/780, where computing time increases by probably more than an order of magnitude. Since a typical run required to evaluate an entry in the tables of GM takes typically 1–2 h on a VAX-11/8650 (with a large amount of variation depending on the initial parameters), then it becomes clear that the model will not run in the extended precision in realistic time on the older type VAX computers.

REFERENCES

1. A. Arakawa and Y. Mintz, *The UCLA Atmospheric General Circulation Model*, Appendix to Chapter V, p. V-35, notes distributed at the workshop March 25–April 4, 1974, Department of Meteorology, University of California, Los Angeles, 1974.
2. J. R. Bates and A. McDonald, *Mon. Weather Rev.* **110**, 1831 (1982).
3. J. Crank and P. Nicholson, *Proc. Cambridge Philos. Soc.* **43**, 50 (1947).
4. A. Cuyt and L. Wuytack, *Non-linear Methods in Numerical Analysis*, (North-Holland, Amsterdam, 1987).
5. D. Gottlieb and S. A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications* (SIAM, Philadelphia, 1987).
6. H. Granek, *Neutrinos In The Eearly Universe*, Doctoral thesis, University of Melbourne, 1988 (unpublished).
7. H. Granek and B. H. J. McKellar (GM), Constraints on the heavy neutrino decays in the early universe, *Int. J. Mod. Phys. A* **6**, 2387 (1991).

8. S. Hayakawa, T. Matsumoto, H. Matsuo, H. Murakami, S. Sato, A. E. Lange, and P. L. Richards, *P. A. S. J.* **39**, 941 (1987).
9. F. B. Hildebrand, *Finite-Difference Equations and Simulations* (Prentice-Hall, Englewood Cliffs, NJ, 1968).
10. A. F. Illarionov and R. A. Sunyaev, *Sov. Astron.* **18**, 413 (1975) (*Astron. Zh.* **51**, 698 (1974)).
11. E. Isacson and H. B. Keller, *Analysis of Numerical Methods* (Wiley, New York, 1966).
12. A. S. Kompaneetz, *Sov. Phys. JETP* **4**, 730 (1957) (*J. Exp. Teor. Fiz.* **31**, 876 (1956)).
13. A. P. Lightman, *Astrophys. J.* **244**, 392 (1981).
14. A. P. Lightman, private communication. This mainly concerns the numerical procedures used to obtain the results in Lightman, 1981 [13].
15. A. R. Mitchell, *Computational Methods in Partial Differential Equations* (Wiley, New York, 1969).
16. R. D. Richtmyer and K. W. Morton, *Difference Methods for Initial-Value Problems*, 2nd ed. (Interscience, New York, 1967).
17. P. J. Roache, *Computational Fluid Dynamics* (Hermosa, Albuquerque, NM, 1972).
18. R. Sedgewick, *Algorithms*, 2nd ed. (Addison–Wesley, Reading, MA, 1988).
19. J. Silk and A. Stebbins, *Astrophys. J.* **269**, 1 (1983).
20. J. von Neumann and R. D. Richtmyer, *J. Appl. Phys.* **21**, 232 (1950).
21. S. Weinberg, *Gravitation and Cosmology* (Wiley, New York, 1972).
22. T. Matsumoto, S. Hayakawa, H. Matsuo, H. Murakami, S. Sato, A. E. Lange, and P. L. Richards, *Astrophys. J.* **329**, 567 (1988).