IMA Journal of Numerical Analysis (1996) **16,** 47–63

Global error estimation with adaptive explicit Runge–Kutta methods

M. CALVO

Departamento de Matemática Aplicada, Universidad de Zaragoza, 50009 Zaragoza, Spain

D. J. HIGHAM

Department of Mathematics and Computer Science, University of Dundee, DD1 4HN, Scotland

AND

J. I. MONTIJANO AND L. RÁNDEZ

Departamento de Matemática Aplicada, Universidad de Zaragoza, 50009 Zaragoza, Spain

[Received 28 July 1994 and in revised form 1 December 1994]

Users of locally-adaptive software for initial value ordinary differential equations are likely to be concerned with global errors. At the cost of extra computation, global error estimation is possible. Zadunaisky's method and 'solving for the error estimate' are two techniques that have been successfully incorporated into Runge–Kutta algorithms. The standard error analysis for these techniques, however, does not take account of the stepsize selection mechanism. In this paper, some new results are presented which, under suitable assumptions show that these techniques are asymptotically valid when used with an adaptive, variable stepsize algorithm—the global error estimate reproduces the leading term of the global error in the limit as the error tolerance tends to zero. The analysis is also applied to Richardson extrapolation (step halving). Numerical results are provided for the technique of solving for the error estimate with several Runge–Kutta methods of Dormand, Lockyer, McGorrigan and Prince.

1. Introduction

Adaptive software for the initial value system

$$y'(t)=f(t,y(t)), \qquad 0<t\leqslant t_{\text{end}}, \qquad y(0)=y_0, \tag{1.1}$$

produces discrete approximations $y_n \approx y(t_n)$. Typically an error tolerance δ, which is supplied by the user to control the accuracy of the numerical solution, determines dynamically the gridpoints $t_n=t_{n-1}+h_n$, $(t_0=0)$ and the discrete approximations y_n at these points. The user is most likely to be concerned with the *global errors* $\mathrm{GE}_n := y_n - y(t_n)$; however, the relation between GE_n and δ is highly problem-dependent. (For example, choosing $\delta=10^{-6}$ does not automatically ensure that the numerical solution has six correct decimal digits.) It is possible, however, to invest more computational effort in order to compute global error approximations. In this paper we consider global error estimation techniques for explicit Runge-Kutta (ERK) algorithms. Traditionally, such schemes are analysed without reference to the stepsize selection process—stepsizes are assumed to be either constant or bounded above by some maximum value. However, the schemes are usually implemented within adaptive, variable stepsize algorithms that are driven by the error tolerance δ. Hence, we feel that an important question remains unanswered—how does a global error estimation technique, when combined with an adaptive strategy for varying the stepsize, behave as a function of δ? Since a user has direct control over δ (rather than the individual stepsizes) this is a very natural question to ask. Clearly, for a scheme to be asymptotically valid in this sense, it must produce a global error estimate that converges to the true global error as $\delta \to 0$. This is the issue that we address in this paper.

The paper is organized as follows: In the next section we introduce the adaptive ERK methods and describe the three global error estimation techniques that we analyse. Section 3 outlines known results about the behaviour of the global error as a function of δ, both for the standard stepsize changing technique and a recently proposed alternative (Calvo *et al* (1994)). In Section 4 we prove that the global error estimates are asymptotically valid, in the sense that they deliver the leading term in the global error as $\delta \to 0$. We also quantify the effect of so-called r-term estimators in the variable stepsize setting. Finally, in Section 5, we present some numerical experiments that confirm the asymptotic behaviour in the case of solving for the error estimate.

2. Global error estimation techniques

Techniques for global error estimation have been available for many years. A diverse range of approaches has been put forward, with various degrees of mathematical rigour. We refer the reader to Peterson (1986) and Skeel (1986) for overviews of the area. In this work we concentrate on three techniques that have been found to work well in practice and have been rigorously justified in the case of constant stepsizes. (For the third technique, Richardson extrapolation, a variable stepsize proof is also available, Henrici (1962).)

We begin by specifying the class of adaptive ERK methods that we consider. We assume that a pth-order ERK formula is used to obtain a discrete solution $\{t_n, y_n\}$ of (1.1) whose global error is to be estimated. This means that the local error LE_n over a step from (t_{n-1}, y_{n-1}) to (t_n, y_n) satisfies

$$\mathrm{LE}_n := y_n - z(t_n) = h_n^{p+1}\psi(t_{n-1}, y_n - {}_1) + O(h_n^{p+2}), \tag{2.1}$$

where $h_n = t_n - t_{n-1}$ is the stepsize, ψ is a sufficiently smooth function, and $z(t)$ is a piecewise continuous function defined in each interval $(t_{n-1}, t_n]$ as the local solution satisfying $z'(t) = f(t, z(t))$ and $z(t_{n-1}) = y_{n-1}$ together with $z(0) = y_0$.

For the purpose of error control, on the step from (t_{n-1}, y_{n-1}) to (t_n, y_n) a locally-based measure of the error, $e_n = e(h_n; t_{n-1}, y_{n-1})$ is computed, where

$$e_n = h_n^p\tilde{\psi}(t_{n-1}, y_{n-1}) + \mathrm{O}(h_n^{p+1}), \tag{2.2}$$

and $\bar{\psi}$ is a sufficiently smooth function. Recall that the main formula is assumed to have order p. Error estimates of the form (2.2) arise with (a) local-error-per-unit-step control, (b) defect control with a suitably high order interpolant, and (c) extrapolated-local-error-per-step control when the secondary formula has order $p-1$. (See Higham (1991) for more details and references.) The step is accepted if $\|e_n\| \leq \delta$, where $\|\cdot\|$ is some vector norm and δ is the user-supplied error tolerance. Otherwise the step is re-taken from t_{n-1} with a smaller stepsize. The next stepsize h_{n+1} is given by the formula

$$h_{n+1} = \theta\left(\frac{\delta}{\text{estmax}_n}\right)^{1/p} h_n, \tag{2.3}$$

where $\theta \in (0, 1)$ is constant and, in the usual case,

$$\text{estmax}_n = \|e_n\|. \tag{2.4}$$

The alternative formula proposed in Calvo *et al* (1994) has

$$\text{estmax}_n = \max\left\{\|e_n\|, h_n^p \min\left\{\kappa \frac{1}{t_n}\sum_{i=1}^{n}\frac{\|e_i\|}{h_i^{p-1}}, \text{estabs}\right\}\right\}, \tag{2.5}$$

in (2.3), where κ and estabs are constants that depend on the method. This version produces a more robust global error behaviour.

Solving for the error estimate, which is also referred to as 'solving for the correction', assumes that a computable continuous extension $Q(t)$ of the discrete solution $\{t_n, y_n\}$ is available (i.e. $Q(t_n) = y_n$) satisfying $Q(t) \in \mathbb{C}^1$. Putting $\varepsilon(t) = Q(t) - y(t)$, the global error in $Q(t)$ satisfies the so-called error equation or secondary problem given by

$$\varepsilon'(t) = \bar{f}(t, \varepsilon(t)) := Q'(t) - f(t, Q(t) - \varepsilon(t)), \qquad \varepsilon(0) = 0. \tag{2.6}$$

Hence solving the secondary problem numerically provides an approximation to the global error in the main problem.

In general, (2.6) is solved with a different ERK method, but with the same stepsize sequence as that used for the main problem. It is intuitively clear that in order for the resulting global error estimate to be valid, the secondary problem must be solved more accurately than the main one. Suppose that over each step the continuous extension satisfies

$$Q(t_{n-1} + \sigma h_n) - z(t_{n-1} + \sigma h_n) = \mathrm{O}(h_n^{p+s}), \quad \text{for all} \quad 0 < \sigma \leq 1, \tag{2.7}$$

where either $s = 0$ (giving an order $p-1$ extension) or $s = 1$ (giving an order p extension). Dormand *et al* (1989) showed that by exploiting the special structure of the secondary problem is possible to derive customised ERK formulas that achieve the extra order of accuracy with fewer stages that a general formula of order $p + r$. Recall that the local error in the main integration is $\mathrm{O}(h_n^{p+1})$. Dormand *et al* (1989) derive formulas that produce local errors on the secondary problem that are essentially $\mathrm{O}(h_n^{p+r+1})$, where $r \geq 1$ and typically $r = 1$ or 2. They are referred to as r-term estimation formulas, since they have r extra zero terms in their local error expansions.

Zadunaisky's technique dates back to the work of Zadunaisky (1966); see also Pereyra (1984). The basic idea is to introduce a neighbouring problem to (1.1), which, in modern references, is taken to be

$$\hat{y}'(t)=\hat{f}(t,\hat{y}(t)):=f(t,\hat{y}(t))+d(t), \qquad \hat{y}(0)=y_0, \tag{2.8}$$

where $d(t)=d(t,\delta):=Q'(t)-f(t,Q(t))$ is the defect in $Q(t)$. This problem is set up to have solution $\hat{y}(t)=Q(t)$. The neighbouring problem is solved with the same ERK formula and the same stepsize sequence as the main problem. Since the exact solution of (2.8) is known, the global error in the neighbouring computation can be found, and used as an approximation to the global error in the main computation. Intuitively, in order for this technique to be valid the main and neighbouring problems must be 'sufficiently close', which amounts to saying that the defect must be small. Note that with this technique, both the main and the neighbouring problem are solved with the same 'special' formula. Therefore, this technique is less flexible than solving for the error estimate with regard to the choice of the RK formulas. Dormand *et al* (1989) derive special r-term estimation formulas with the property that the local errors in the main and neighbouring computations differ by $O(h_n^{p+r+1})$.

Global error estimation by *Richardson extrapolation*, or step halving, proceeds as follows. The usual error control and stepsize selection method is applied to produce a solution $\{y_n\}$ from stepsizes $\{h_n\}$. Simultaneously, the same ERK formula is applied over pairs of steps with stepsize $h_n/2$ to produce another discrete approximation $\{\tilde{y}_n\}$. The quantity

$$\frac{y_n-\tilde{y}_n}{2^p-1} \tag{2.9}$$

then approximates the global error in $\tilde{y}_n$. (Similarly, multiplying by 2^p in (2.9) produces an approximation to the global error in y_n.) This technique, which does not require the computation of a continuous solution, was chosen by Shampine and Watts (1976) for the GERK code.

3. The global error

In this section we review the relevant results from Calvo *et al* (1994) and Higham (1991) concerning the behaviour of the global error in a variable stepsize algorithm.

We assume that

(i) the stepsize satisfying $\max_n \{h_n\}\to 0$ as $\delta\to 0$, and
(ii) the function $\tilde{\psi}(t,y(t))$ of (2.2) is non-vanishing over $[0,t_{\text{end}}]$.

From the first of these assumptions, standard convergence theory implies that $\max_n\{y_n-y(t_n)\}\to 0$. Assumption 2 ensures that we are controlling a quantity that behaves like h_n^p (and not like some higher power of h_n). It follows that if a function is $O(h_n)$ then it is also $O(\delta^{1/p})$. The second assumption above can be weakened to $\tilde{\psi}(0,y(0))\neq 0$ if the new stepsize changing technique defined by

(2.3) and (2.5) is used; see Calvo *et al* (1994) for details. However, for simplicity, we will assume that the standard step-changing formula (2.3)–(2.4) is used. (Our analysis is easily adapted to lead to the same qualitative results for the new technique.) From (2.2)–(2.4), since h_n and h_{n+1} are $O(\delta^{1/p})$, we have

$$\|e_{n+1}\| = h_{n+1}^p \|\tilde{\psi}(t_n, y_n)\| + O(\delta^{(p+1)/p}) = \frac{\theta^p \delta h_n^p \|\tilde{\psi}(t_n, y_n)\|}{h_n^p \|\tilde{\psi}(t_{n-1}, y_{n-1})\|} + O(\delta^{(p+1)/p}).$$

Hence, on every step,

$$\|e_n\| = \theta^p \delta + O(\delta^{(p+1)/p}), \tag{3.1}$$

which will be needed later.

Results in Higham (1991) show that under these assumptions the numerical solution has a global error that is asymptotically linear in δ. More precisely, if we let $\eta_l(t)$ denote the interpolant defined by

$$\eta_l(t) := z(t) + \frac{(t - t_{n-1})}{h_n} \mathrm{LE}_n, \qquad t \in (t_{n-1}, t_n], \qquad \eta_l(0) = y_0, \tag{3.2}$$

then as $\delta \to 0$, for any fixed t, the global error in $\eta_l(t)$ satisfies

$$\eta_l(t) - y(t) = v(t)\delta + g(t). \tag{3.3}$$

Here, $v(t)$ is $\mathbb{C}^1$ and independent of δ, and $g(t)$ is continuous and piecewise $\mathbb{C}^1$ with zeroth and first derivatives of $O(\delta^{(p+1)/p})$. Note that $\eta_l(t)$ is not computable, in general, but it passes through the mesh data $\{t_n, y_n\}$. Hence this result reveals information about the global error in the discrete solution as a function of δ. The defect in $\eta_l(t)$ will be denoted $d_l(t)$:

$$d_l(t) := \eta_l'(t) - f(t, \eta_l(t)). \tag{3.4}$$

(For definiteness, the derivative $\eta_l'(t)$ at each gridpoint $t = t_n$ is defined by taking the limit from the left.)

Throughout our analysis we assume that the computable extension $Q(t)$ is $\mathbb{C}^1$ with $s = 0$ or $s = 1$ in (2.7). It follows (Higham (1991)) that

$$Q(t) - y(t) = O(\delta) \quad \text{and} \quad d(t) = O(\delta^{(p-1)/p}). \tag{3.5}$$

The aim of this work is to analyse the global error estimation techniques described in Section 2, when incorporated into a variable stepsize algorithm. We will show the under assumptions (i) and (ii), the global error estimates are asymptotically valid as $\delta \to 0$.

4. Analysis of the global error estimation techniques

4.1 *Solving for the error estimate*

We now examine the technique of solving for the error estimate. We suppose that a numerical solution $\{t_n, \varepsilon_n\}$ has been computed for the secondary problem (2.6), and we define an interpolant through this data by

$$\overline{\eta_l}(t) := \bar{z}(t) + \frac{(t - t_{n-1})}{h_n} \overline{\mathrm{LE}}_n, \qquad t \in (t_{n-1}, t_n] \qquad \overline{\eta_l}(0) = 0, \tag{4.1}$$

where the associated local solution $\bar{z}(t)$ satisfies

$$\bar{z}'(t)=\bar{f}(t,\bar{z}(t)) \qquad \text{and} \qquad \bar{z}(t_{n-1})=\bar{\varepsilon}_{n-1}, \qquad \bar{z}(0)=0$$

and $\overline{\mathrm{LE}}_n := \varepsilon_n - \bar{z}(t_n)$ is the local error. We also let $\bar{d}(t) := \overline{\eta_l}'(t) - \bar{f}(t, \overline{\eta_l}(t))$ denote the defect in $\overline{\eta_l}(t)$. Our aim is to show that $\overline{\eta_l}(t)$ is a valid approximation to $\varepsilon(t)$, the global error in the main problem.

We begin with a lemma.

Lemma 4.1 The defect $\bar{d}(t)$ in the interpolant $\overline{\eta_l}(t)$ defined above for an r-term estimation formula when solving for the error estimate satisfies

$$\bar{d}(t) = \bar{\gamma}(t)\delta^{(p+r)/p} + \mathrm{O}(\delta^{(p+r+1)/p}), \tag{4.2}$$

where $\bar{\gamma}(t)$ is $\mathbb{C}^1$ and independent of δ.

Proof. The key property is that an r-term estimation formula has a local error expansion that is zero until the h_n^{p+1+r} term (see equations (12) and (15) of Dormand *et al* (1989)). Moreover, since the elementary differentials of the secondary problem at the point $(t_{n-1}, \varepsilon_{n-1})$ can be written in terms of (t_{n-1}, y_{n-1}) (see Dormand *et al* (1989), p 840), it follows that the local-error-per-unit-step satisfies

$$\frac{\overline{\mathrm{LE}}_n}{h_n} = \bar{\psi}(t_{n-1}, y_{n-1})h_n^{p+r} + \mathrm{O}(\delta^{(p+r+1)/p}), \tag{4.3}$$

where $\bar{\psi}$ and $\mathbb{C}^1$ and independent of δ. From (2.2), we have

$$h_n^p = \frac{\|e_n\|}{\|\bar{\psi}(t_{n-1}, y_{n-1})\|} + \mathrm{O}(\delta^{(p+1)/p}).$$

Using this, along with (3.1), in (4.3) shows that the local-error-per-unit-step has the form

$$\frac{\overline{\mathrm{LE}}_n}{h_n} = \bar{\gamma}(t_{n-1}, y_{n-1})\delta^{(p+r)/p} + \mathrm{O}(\delta^{(p+r+1)/p}), \tag{4.4}$$

where

$$\bar{\gamma}(t,y) = \theta^{p+r}\frac{\bar{\psi}(t,y)}{\|\bar{\psi}(t,y)\|^{(p+r)/p}}$$

is $\mathbb{C}^1$ and independent of δ.

Now from the definition (4.1) of the interpolant, since $(t-t_{n-1})/h_n = \mathrm{O}(1)$,

$$\begin{aligned}
\bar{d}(t) &= \overline{\eta_l}'(t) - \bar{f}(t, \overline{\eta_l}(t)) \\
&= \bar{z}'(t) + \frac{\overline{\mathrm{LE}}_n}{h_n} - \bar{f}(t, \bar{z}(t)) + \mathrm{O}(\overline{\mathrm{LE}}_n) \\
&= \frac{\overline{\mathrm{LE}}_n}{h_n} + \mathrm{O}(\overline{\mathrm{LE}}_n) \\
&= \bar{\gamma}(t_{n-1}, y_{n-1})\delta^{(p+r)/p} + \mathrm{O}(\delta^{(p+r+1)/p}),
\end{aligned}$$

using (4.4). Finally, since $\bar{\gamma}$ is $\mathbb{C}^1$, we may replace the arguments (t_{n-1}, y_{n-1}) by $(t, y(t))$ without affecting the leading term, giving the required expression. (With a slight abuse of notation, we replace $\bar{\gamma}(t, y(t))$ by $\bar{\gamma}(t)$.) □

We can now prove the main theorem concerning the accuracy of the estimate.

THEOREM 4.1 The error in the global error estimate of an r-term estimation formula when solving for the error estimate satisfies

$$\overline{\eta_r}(t) - \varepsilon(t) = \bar{v}(t)\delta^{(p+r)/p} + O(\delta^{(p+r+1)/p}), \tag{4.5}$$

where $\bar{v}(t)$ is $\mathbb{C}^1$ and independent of δ.

Proof. We will use $\omega(t) := \overline{\eta_r}(t) - \varepsilon(t)$ to denote the error in the global error estimate. Taking the defect and expanding about $\varepsilon(t)$ leads to the variational equation

$$\omega'(t) - \bar{f}_\varepsilon(t, \varepsilon(t))\omega(t) = \bar{d}(t) + O(\omega(t)^2). \tag{4.6}$$

Its solution satisfies (see, for example, Ascher *et al* (1988))

$$\omega(t) = \bar{Y}(t)\int_0^t \bar{Y}^{-1}(\mu)\bar{d}(\mu)\,d\mu + O(\omega(t)^2), \tag{4.7}$$

where $\bar{Y}(t)$ is a fundamental matrix of the variational equation of the secondary problem (2.6) with respect to the solution $\varepsilon(t)$. Since $\bar{f}_\varepsilon(t, \varepsilon(t)) = f_y(t, y(t))$, $\bar{Y}(t)$ satisfies

$$\bar{Y}'(t) = f_y(t, y(t))\bar{Y}(t),$$

showing that $\bar{Y}(t)$ is independent of δ.

Now, inserting the expression (4.2) into (4.7) gives

$$\omega(t) = \delta^{(p+r)/p}\bar{Y}(t)\int_0^t \bar{Y}^{-1}(\mu)\bar{\gamma}(\mu)\,\mathrm{d}\mu + O(\delta^{(p+r+1)/p}) + O(\omega(t)^2), \tag{4.8}$$

proving the result. □

Theorem 4.1 shows that the technique is asymptotically valid in the sense that

$$\varepsilon_n - \varepsilon(t_n) = \bar{v}(t_n)\delta^{(p+r)/p} + O(\delta^{(p+r+1)/p}),$$

where $\bar{v}(t)$ is the solution of the linear problem

$$\bar{v}'(t) = f_y(t, y(t))\bar{v}(t) + \theta^{p+r}\frac{\bar{\psi}(t, y(t))}{\|\bar{\psi}(t, y(t))\|^{(p+r)/p}}, \qquad \bar{v}(0) = 0. \tag{4.9}$$

The global error, which is $O(\delta)$, is approximated up to terms of $O(\delta^{(p+r)/p})$. In particular, this confirms that there is a gain in accuracy if r is increased. Furthermore, the theorem shows that for a given t, the leading term in the 'error in the error' settles down to a fixed value, $\bar{v}(t)$, and that $\bar{v}(t)$ is a $\mathbb{C}^1$ function of t. This implies that for sufficiently small δ, each component of $\overline{\eta_r}(t) - \varepsilon(t)$ decreases monotonically to zero with δ—this conclusion could not be drawn from a weaker results such as $\overline{\eta_r}(t) - \varepsilon(t) = O(\delta^{(p+1)/p})$.

4.2 *Zadunaisky's technique*

To investigate Zadunaisky's technique, we first define quantities analogous to those used in the previous section. We let $\{t_n, \hat{y}_n\}$ denote the numerical solution for the neighbouring problem (2.8). The corresponding local solution and local error over a step from t_{n-1} to t_n will be denoted $\hat{z}(t)$ and $\widehat{\mathrm{LE}}_n$, respectively, with

$$\hat{z}'(t) = \hat{f}(t, \hat{z}(t)), \qquad \hat{z}(t_{n-1}) = \hat{y}_{n-1} \quad \text{and} \quad \widehat{\mathrm{LE}}_n := \hat{y}_n - \hat{z}(t_n).$$

An interpolant $\hat{\eta}_I(t)$ to the discrete solution can be defined by

$$\hat{\eta}_I(t) := \hat{z}(t) + \frac{(t - t_{n-1})}{h_n}\widehat{\mathrm{LE}}_n, \qquad t \in (t_{n-1}, t_n], \quad \hat{\eta}_I(0) = y_0,$$

and the corresponding defect $\hat{d}_I(t)$ is given by $\hat{d}_I(t) := \hat{\eta}_I'(t) - \hat{f}(t, \hat{\eta}_I(t))$. We use $\hat{\varepsilon}_I(t) := \hat{\eta}_I(t) - Q(t)$ to denote the global error in $\hat{\eta}_I(t)$ for the neighbouring computation. Similarly $\varepsilon_I(t) := \eta_I(t) - y(t)$ denote the analogous global error for the main problem. Our aim is to compare $\hat{\varepsilon}_I(t)$ and $\varepsilon_I(t)$.

The following lemma captures an essential property of an r-term estimation formula.

LEMMA 4.2 The defect $\hat{d}_I(t)$ in the interpolant $\hat{\eta}_I(t)$ defined above for an r-term Zadunaisky formula satisfies

$$\hat{d}_I(t) = d_I(t) + \hat{\gamma}(t)\delta^{(p+r)/p} + \mathrm{O}(\delta^{(p+r+1)/p}) + \mathrm{O}(\delta\hat{\varepsilon}_I(t)), \tag{4.10}$$

where $\hat{\gamma}(t)$ is $\mathbb{C}^1$ and independent of δ, and we recall that $d_I(t)$ is defined in (3.4).

Proof. With an r-term formula, the expansion of the local error for the neighbouring problem and for the main problem on each step agree up to and including the $\mathrm{O}(h_n^{p+r})$ term (see equation (12) of Dormand *et al* (1989)). It follows that the difference between the local-error-per-unit-step on the two problems satisfies

$$\frac{\widehat{\mathrm{LE}}_n}{h_n} - \frac{\mathrm{LE}_n}{h_n} = \hat{\psi}(t_{n-1}, y_{n-1})h_n^{p+r} + \mathrm{O}(\delta^{(p+r+1)/p}), \tag{4.11}$$

where $\hat{\psi}$ is $\mathbb{C}^1$ and independent of δ. Analysis similar to that in Lemma 4.1 then shows that

$$\frac{\widehat{\mathrm{LE}}_n}{h_n} - \frac{\mathrm{LE}_n}{h_n} = \hat{\gamma}(t_{n-1}, y_{n-1})\delta^{(p+r)/p} + \mathrm{O}(\delta^{(p+r+1)/p}), \tag{4.12}$$

where $\hat{\gamma}$ is $\mathbb{C}^1$ and independent of δ.

Now from the definitions of the interpolants, we have

$$\hat{d}_I(t) = \frac{\widehat{\mathrm{LE}}_n}{h_n} - \frac{(t - t_{n-1})}{h_n}\hat{f}_{\hat{y}}(t, \hat{z}(t))\widehat{\mathrm{LE}}_n + \mathrm{O}(\widehat{\mathrm{LE}}_n^2), \tag{4.13}$$

$$d_I(t) = \frac{\mathrm{LE}_n}{h_n} - \frac{(t - t_{n-1})}{h_n}f_y(t, z(t))\mathrm{LE}_n + \mathrm{O}(\mathrm{LE}_n^2). \tag{4.14}$$

Note that $\hat{f}_{\hat{y}} \equiv f_y$, by construction. Also, using (3.5)

$$\hat{z}'(t) - f\big(t, \hat{z}(t)\big) = d(t) = \mathrm{O}(\delta^{(p-1)/p})$$

and

$$\hat{z}(t_{n-1}) - z(t_{n-1}) = \hat{\varepsilon}_l(t_{n-1}) = \mathrm{O}(\delta).$$

Hence using a standard differential inequality (see, for example, Hairer *et al* (1993), Theorem 10.2, page 58) we have, for $t_{n-1} < t \leqslant t_n$,

$$\hat{z}(t) - z(t) = \mathrm{O}(\delta) + \mathrm{O}(h_n \delta^{(p-1)/p}) = \mathrm{O}(\delta). \tag{4.15}$$

Subtracting (4.14) from (4.13), and using (4.15), leads to

$$\begin{aligned}\hat{d}_l(t) - d_l(t) = \frac{\widehat{\mathrm{LE}}_n}{h_n} - \frac{\mathrm{LE}_n}{h_n} - \frac{(t - t_{n-1})}{h_n} f_y\big(t, z(t)\big)(\widehat{\mathrm{LE}}_n - \mathrm{LE}_n)\\ + \mathrm{O}(\widehat{\mathrm{LE}}_n^2) + \mathrm{O}(\mathrm{LE}_n^2) + \mathrm{O}\big(\delta \hat{\varepsilon}_l(t)\big) + \mathrm{O}(\delta^2).\end{aligned}$$

Now, from (4.12), this becomes

$$\hat{d}_l(t) - d_l(t) = \hat{\gamma}(t_{n-1}, y_{n-1})\delta^{(p+r)/p} + \mathrm{O}(\delta^{(p+r+1)/p}) + \mathrm{O}\big(\delta \hat{\varepsilon}_l(t)\big),$$

and replacing the arguments (t_{n-1}, y_{n-1}) by $\big(t, y(t)\big)$ gives the result. □

We can now prove the main theorem concerning the accuracy of the estimate.

THEOREM 4.2 Suppose $1 \leqslant r < p$. With an r-térm Zadunaisky algorithm the global error for the neighbouring problem satisfies

$$\hat{\varepsilon}_l(t) = \varepsilon_l(t) + \hat{v}(t)\delta^{(p+r)/p} + \mathrm{O}(\delta^{(p+r+1)/p}), \tag{4.16}$$

where $\hat{v}(t)$ is $\mathbb{C}^1$ and independent of δ.

Proof. The technique of proof is similar to that used for Theorem 4.1.

The global error $\hat{\varepsilon}_l(t) = \hat{\eta}_l(t) - Q(t)$ for the neighbouring problem satisfies the variational equation

$$\hat{\varepsilon}_l'(t) - \hat{f}_{\hat{y}}\big(t, \hat{y}(t)\big)\hat{\varepsilon}_l(t) = \hat{d}_l(t) + \mathrm{O}\big(\hat{\varepsilon}_l(t)^2\big). \tag{4.17}$$

By construction, $\hat{y}(t) = Q(t)$ and $\hat{f}_{\hat{y}} \equiv f_y$, so the solution of (4.17) satisfies

$$\hat{\varepsilon}_l(t) = \hat{Y}(t)\int_0^t \hat{Y}^{-1}(\mu)\hat{d}_l(\mu)\,\mathrm{d}\mu + \mathrm{O}\big(\hat{\varepsilon}_l(t)^2\big), \tag{4.18}$$

where the fundamental matrix $\hat{Y}(t)$ is the solution of

$$\hat{Y}'(t) = f_y\big(t, Q(t)\big)\hat{Y}(t), \qquad \hat{Y}(0) = I. \tag{4.19}$$

Now $Q(t) = y(t) + \varepsilon(t)$, so replacing $Q(t)$ by $y(t)$ in (4.19) introduces an $\mathrm{O}\big(\varepsilon(t)\big)$ perturbation. It follows from a standard differential inequality (see, for example, Hairer *et al* (1993), Theorem 10.2, page 57) that

$$\hat{Y}(t) = Y(t) + \mathrm{O}\big(\varepsilon(t)\big), \quad \text{and} \quad \hat{Y}^{-1}(t) = Y^{-1}(t) + \mathrm{O}\big(\varepsilon(t)\big), \tag{4.20}$$

where $Y(t)$ solves $Y'(t) = f_y\big(t, y(t)\big)Y(t)$ and $Y(0) = I$, and hence is independent of δ.

Finally, using (4.20) in (4.18), substituting for $\hat{d}_l(t)$ from Lemma 4.2 and using (3.5) we have

$$\begin{aligned}\hat{\varepsilon}_l(t) &= Y(t)\int_0^t Y^{-1}(\mu)(d_l(\mu) + \hat{\gamma}(\mu)\delta^{(p+r)/p})\,\mathrm{d}\mu + \mathrm{O}(\delta^{(p+r+1)/p})\\ &\quad + \mathrm{O}\big(\delta\hat{\varepsilon}_l(t)\big) + \mathrm{O}\big(\hat{\varepsilon}_l(t)^2\big),\\ &= \varepsilon_l(t) + Y(t)\int_0^t Y^{-1}(\mu)\hat{\gamma}(\mu)\delta^{(p+r)/p}\,\mathrm{d}\mu + \mathrm{O}(\delta^{(p+r+1)/p})\\ &\quad + \mathrm{O}\big(\delta\hat{\varepsilon}_l(t)\big) + \mathrm{O}\big(\hat{\varepsilon}_l(t)^2\big),\end{aligned}$$

giving the result. □

Note that $\varepsilon_l(t_n) := \eta_l(t_n) - y(t_n) = y_n - y(t_n)$, and similarly $\hat{\varepsilon}_l(t_n) = \hat{y}_n - Q(t_n)$. Hence, Theorem 4.2 applies to the computed global error estimate at the gridpoints, and it shows that Zadunaisky's technique has a similar asymptotic validity to solving for the error estimate, cf. Theorem 4.1. The result requires $r < p$, which is a consequence of neglecting $\mathrm{O}(\delta^2)$ terms in the linearization (4.17). However, since r is typically 1 or 2 this limitation has no practical significance.

4.3 *Richardson extrapolation*

We now examine Richardson extrapolation, as described in Section 2. Henrici (1962, page 136) gave a variable stepsize result for this technique under the assumption that $h_n = \vartheta(t_n)h$, where $\vartheta(t)$ is piecewise continuous, $0 < \vartheta(t) \leq 1$ for $t \in [0, t_{\mathrm{end}}]$, and h is constant. We can regard h as the 'maximum' stepsize. Henrici showed that the global error estimate is correct up to $\mathrm{O}(h^{p+1})$ terms. We show below that it is straightforward to analyse Richardson extrapolation when the standard stepsize selection process is used, and the estimate can be shown to be valid in the limit as $\delta \to 0$.

From Higham (1991) the relation (3.3) holds for the true global error, with $v(t)$ the solution of the variational equation

$$v'(t) - f_y\big(t, y(t)\big)v(t) = \theta^p \frac{\psi\big(t, y(t)\big)}{\|\tilde{\psi}\big(t, y(t)\big)\|}, \qquad v(0) = 0. \tag{4.21}$$

For the standard step-changing policy, the functions ψ and $\tilde{\psi}$ are those appearing in the expansions (2.1) and (2.2). With the alternative step-changing policy given by (2.5), the function $\tilde{\psi}$ must be defined slightly differently—see Calvo *et al* (1994). Now consider a second, simultaneous integration using pairs of steps with length $h_n/2$ to generate $\{t_n, \tilde{y}_n\}$. Straightforward analysis (see, for example, Hairer *et al* (1993), Section II.4) shows that the local error at the end of a pair of steps of length $h_n/2$ from $\tilde{y}_{n-1}$ to $\tilde{y}_n$ has the form

$$\tilde{y}_n - \tilde{z}_n(t_n) = \frac{1}{2^p} h_n^{p+1}\psi(t_{n-1}, \tilde{y}_{n-1}) + \mathrm{O}(h_n^{p+2}),$$

cf. (2.1). Here, $\tilde{z}(t)$ is the appropriate local solution: $\tilde{z}'(t) = f\big(t, \tilde{z}(t)\big)$, and

$\tilde{z}(t_{n-1}) = \tilde{y}_{n-1}$. Hence, by analogy with (3.3) and (4.21), we find that the interpolant

$$\tilde{\eta}_I(t) := \tilde{z}_n(t) + \frac{(t - t_{n-1})}{h_n}[\tilde{y}_n - \tilde{z}(t_n)], \qquad t \in (t_{n-1}, t_n], \qquad \tilde{\eta}_I(0) = y_0,$$

satisfies

$$\tilde{\eta}_I(t) - y(t) = \tilde{v}(t)\delta + O(\sigma^{(p+1)/p}),$$

where $\tilde{v}(t)$ solves the variational equation (4.21) with the right-hand side scaled by $1/2^p$. Because (4.21) is linear, we find $\tilde{v}(t) = v(t)/2^p$. It follows that (2.9) gives a global error estimate that is valid up to $O(\delta^{(p+1)/p})$.

5. Discussion and numerical results

Although our analysis has the advantage of dealing directly with the error tolerance, it suffers from the inherent limitation of being relevant only for 'sufficiently small' δ. However, the underlying adaptive algorithm is based on similar asymptotics—the error estimation and stepsize selection mechanism is motivated by small h expansions such as (2.2), and produces a global error that satisfies the relation (3.3). Hence, it could be argued that the global error estimation techniques analysed here possess similar properties to the underlying approximations. In this sense the results are extremely positive, and serve to justify the use of global error estimation in adaptive algorithms.

In order to check the agreement between our asymptotic results on global error estimators and the numerical results obtained in practice, extensive calculations with different RK methods and initial value problems (largely from the DETEST set of problems, Enright & Pryce (1987)) have been performed. For the sake of brevity we present here only numerical results from solving for the error estimate, which we regard as the method of choice.

Let us recall that the asymptotic results for the technique of solving for the error estimate were obtained under the assumptions (i) and (ii) given in Section 3. Assumption (i) is normally satisfied, but as illustrated in Calvo *et al* (1994), there are particular methods and problems for which assumption (ii) does not hold. In this case it was shown in Calvo *et al* (1994) that the standard stepsize changing policy does not guarantee the tolerance proportionality condition (3.3). Since the leading term $\tilde{v}(t)$ of the global error estimation in the numerical solution of the secondary problem satisfies the variational equation (4.9), similar difficulties to the ones encountered for tolerance proportionality appear in connection with the global error estimation when $\tilde{\psi}(t, y(t))$ vanishes in the integration interval. It must be pointed out that although the assumptions for tolerance proportionality (see Calvo *et al* (1994), Higham (1991)) and for the asymptotic validity of the global error estimation are identical, the two properties are independent. In fact, it is possible to construct methods for which the main formula presents tolerance proportionality even when the function $\tilde{\psi}(t, y(t))$ vanishes, and at the same time, $\tilde{v}(t)$ turns out to be unbounded at some point in the integration interval and therefore the global error estimator is not reliable.

In our numerical experiments, to avoid difficulties with the standard stepsize changing technique caused by a vanishing $\tilde{\psi}(t, y(t))$ we used the alternative stepsize changing formula based on (2.5). With this formula, when $\|\tilde{\psi}(t, y(t))\|$ is small (in a sense which depends on the scale of the problem) at a grid point it is replaced by a suitable non-vanishing function; therefore the new technique is equivalent to using another stepsize function $\tilde{\psi}$ such that $\|\tilde{\psi}\| \geq v > 0$, and assumption (ii) is satisfied.

We chose Runge-Kutta methods from those proposed by Dormand *et al* (1989). A typical method of order p consists of an ERK triple of order p (i.e. a pair of discrete formulas of orders p and $p-1$ together with a continuous extension of order $p-1$ or p) and an ERK formula for the secondary problem with order $p+r$. In some cases the same formula is used for the main and the secondary problem although its order is higher for the latter problem.

Next we describe how we test the asymptotic behaviour of the global error estimator. Given a problem (1.1) and a triple of order p together with an r-term error estimator, we compute, for all gridpoints corresponding to a given tolerance δ, the quotients

$$\tilde{v}_n := \frac{\varepsilon(t_n) - \varepsilon_n}{\delta^{(p+r)/p}}. \tag{5.1}$$

If our theory is correct, these values should converge (for $\delta \to 0$) to points on a fixed curve $\tilde{v}(t)$. (In fact, $\tilde{v}(t)$ solves the variational equation (4.9).)

We present results for two test problems from class A of the DETEST set (Enright & Pryce (1987)):

$$\textbf{A4}: \; y'(t) = \frac{y(t)}{4}\left(1 - \frac{y(t)}{20}\right), \qquad y(0) = 1,$$

$$\textbf{A3}: \; y'(t) = y(t)\cos(t), \qquad y(0) = 1.$$

and one from class D:

$$\textbf{D1}: \begin{bmatrix} y_1'(t) \\ y_2'(t) \\ y_3'(t) \\ y_4'(t) \end{bmatrix} = \begin{bmatrix} y_3(t) \\ y_4(t) \\ -y_1(t)/(y_1(t)^2 + y_2(t)^2)^{3/2} \\ -y_2(t)/(y_1(t)^2 + y_2(t)^2)^{3/2} \end{bmatrix},$$

$$\begin{bmatrix} y_1(0) \\ y_2(0) \\ y_3(0) \\ y_4(0) \end{bmatrix} = \begin{bmatrix} 1-\epsilon \\ 0 \\ 0 \\ \sqrt{(1+\epsilon)/(1-\epsilon)} \end{bmatrix}, \qquad \epsilon = 0{\cdot}1,$$

with $0 < t \leq 20$ in each case.

First, we used the second order method RK2(1)3FD of Dormand *et al* (1989). In this case, the second order triple has three stages and uses the FSAL condition. The same discrete method is used for the main problem and for the error

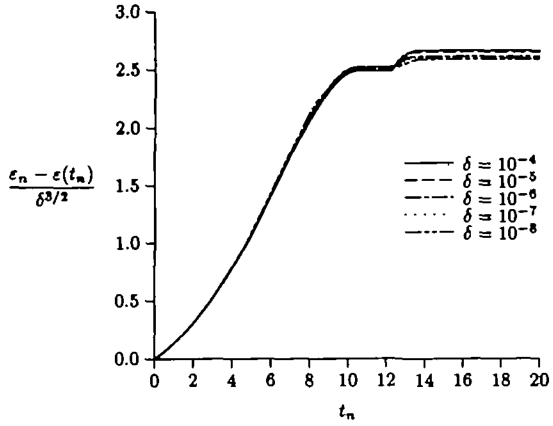

FIG. 1. A4 problem, RK2(1)3FD.

equation, but for the error equation it has order three, and therefore we have one-term global error estimation.

In Fig. 1 we plot for the A4 problem the polygon determined by the points $\{t_n, \tilde{v}_n\}$ for $\delta = 10^{-4}, \ldots, 10^{-8}$. It is clear that as δ tends to zero the polygons converge to a fixed curve. In Figs 2 and 3 we show the corresponding graphs for problems A3 and D1.

It must be remarked that if we use the standard technique for stepsize changing, the figures determined by the points $\{t_n, \tilde{v}_n\}$ show a clear jump when the numerical integration crosses a TP-singular point, i.e. a point t^* where $\tilde{\psi}(t, y(t))$ vanishes. To illustrate this fact, Fig. 4 shows the polygons determined by the points $\{t_n, \tilde{v}_n\}$ for $\delta = 10^{-4}, \ldots, 10^{-8}$ on the A3 problem. In this case, the TP-singular points are those values of t such that $\sin(t) = (\sqrt{5} - 1)/2$, i.e.

$$t^* = 0{\cdot}6662 \cdots + 2k\pi, \qquad t^* = 2{\cdot}475 \cdots + 2k\pi,$$

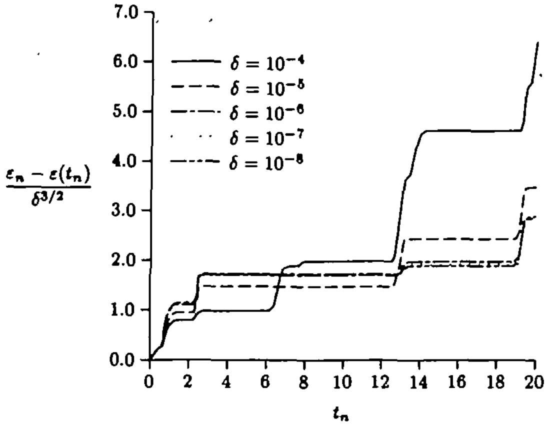

FIG. 2. A3 problem, RK2(1)3FD.

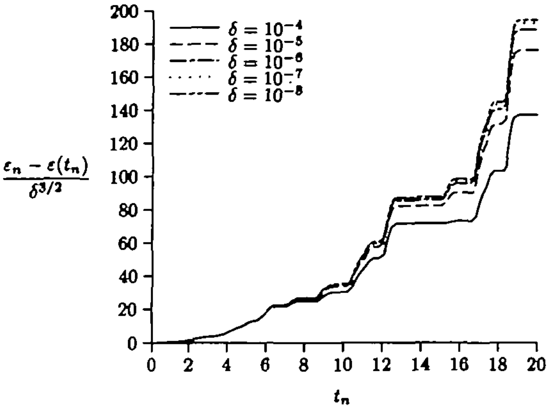

FIG. 3. D1 problem, RK2(1)3FD.

which belong to the interval [0, 20]. At each of these points the polygons show a jump and clearly they do not converge as $\delta \to 0$.

Next we consider the third-order method RK3(2)4FD of Dormand *et al* (1989). This method requires four stages per step and uses the FSAL condition. The same formula is employed as both integrator and estimator. In Figs 5 and 6 we present the polygons determined by the points $\{t_n, \tilde{v}_n\}$ for the problems A4 and A3 respectively, computed with $\delta = 10^{-4}, \ldots, 10^{-12}$.

Finally, we use as the main integrator the well-known pair RK5(4)7FM (or DOPRI5) of Dormand and Prince (Hairer *et al* (1993), page 178) with a continuous extension of order four, and as estimator a seven-stage formula given in Dormand *et al* (1989) which allows two-term estimation. Now, since the order of the method is higher than in the above cases, the global errors are smaller and, moreover, as we use a two-term error estimation, the global errors in the solution of the secondary problem are much smaller. This means that special attention must be paid to roundoff errors. In order to get reliable results, the computations

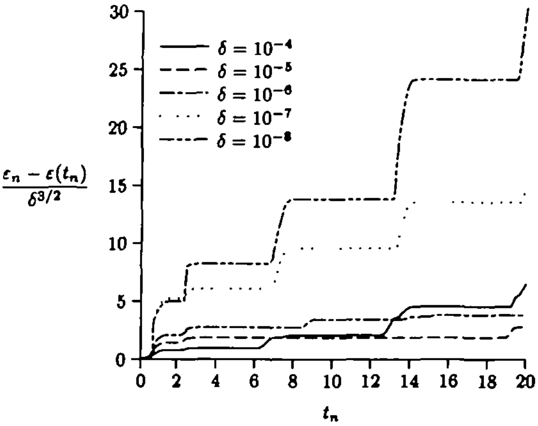

FIG. 4. A3 problem, RK2(1)3FD with standard stepsize control.

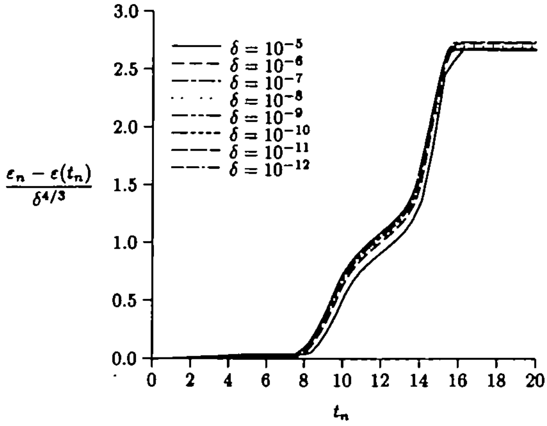

FIG. 5. A4 problem, RK3(2)4FD.

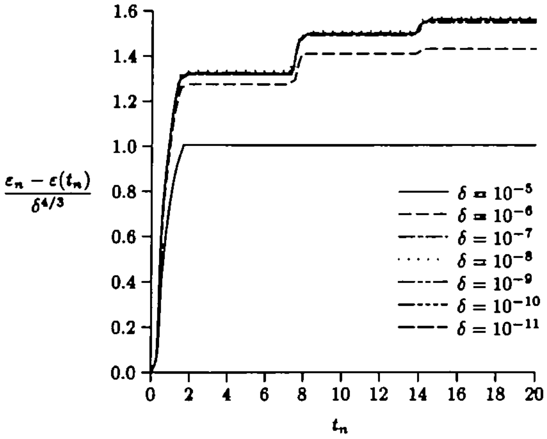

FIG. 6. A3 problem, RK3(2)4FD.

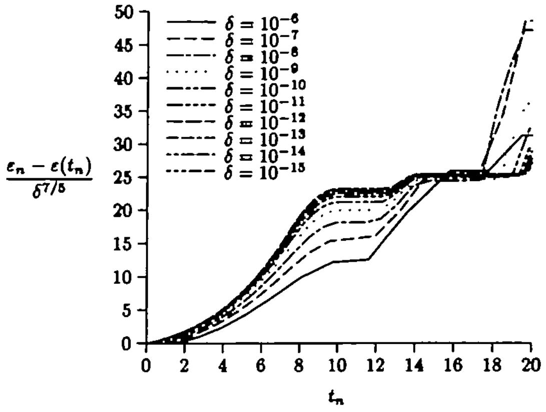

FIG. 7. A4 problem, RK5(4)7FM.

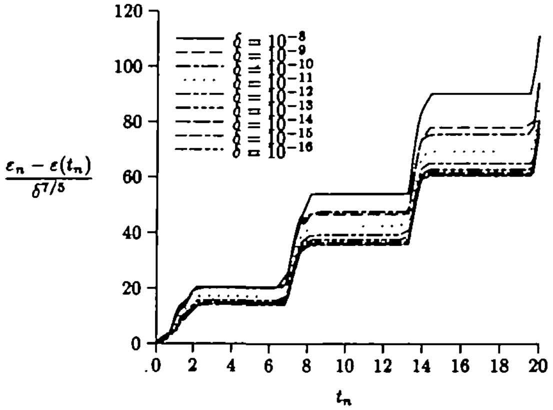

FIG. 8. A3 problem, RK5(4)7FM.

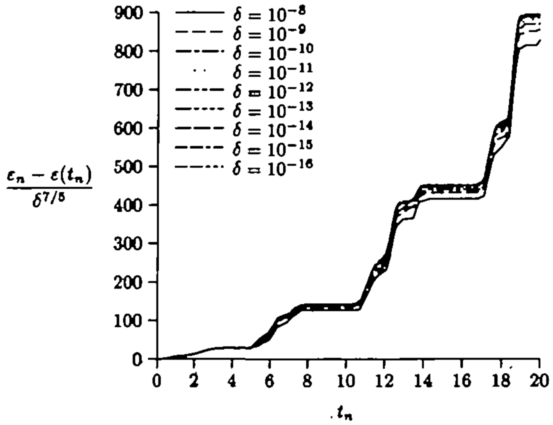

FIG. 9. D4 problem, RK5(4)4FM.

with stringent tolerances were carried out in quadruple precision (32 significant figures). In Figs 7, 8 and 9 we show the polygons determined by $\{t_n, \tilde{v}_n\}$ for the problems A4, A3 and D1, respectively, with tolerances between 10^{-6} and 10^{-16}.

Similar numerical results that support our theoretical predictions have been obtained on a wide range of methods and problems.

Acknowledgements

We thank the referees for their incisive comments. The research work of MC, JIM and LR has been supported by Project DGICYT PB93-0305.

REFERENCES

ASCHER, U. M., MATTHEIJ, R. M. M., & RUSSELL, R. D. 1988 *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, Englewood Cliffs, NJ: Prentice Hall.

CALVO, M., HIGHAM, D. J., MONTIJANO, J. I., & RÁNDEZ, L. 1994 Stepsize selection for tolerance proportionality in explicit Runge-Kutta codes, *Tech. Report NA/152*, University of Dundee.

DORMAND, J. R., LOCKYER, M. A., MCGORRIGAN, N. E., & PRINCE, P. J. 1989 Global error estimation with Runge-Kutta triples, *Comput. Maths. Appl.* **18**, 835-846.

ENRIGHT, W. H., & PRYCE, J. D. 1987 Two FORTRAN packages for assessing initial value methods, *ACM Trans. Math. Software* **13**, 1-27.

HAIRER, E., NØRSETT, S. P., & WANNER, G. 1993 *Solving Ordinary Differential Equations I, Nonstiff Problems*, 2nd ed. Berlin: Springer.

HENRICI, P. 1962 *Discrete Variable Methods in Ordinary Differential Equations*, New York: Wiley.

HIGHAM, D. J. 1991 Global error versus tolerance for explicit Runge-Kutta methods, *IMA J. Numer. Anal.* **11**, 457-480.

PEREYRA, V. 1984 Deferred corrections software and its application to seismic ray tracing, *Computing* **5**, 211-226.

PETERSON, P. J. 1986 Global error estimation using defect correction techniques for explicit Runge-Kutta methods, *Tech. Report 192/86*, University of Toronto.

SHAMPINE, L. F., & WATTS, H. A. 1976 Global error estimation for ordinary differential equations, *ACM Trans. Math. Software* **2**, 172-186.

SKEEL, R. D. 1986 Thirteen ways to estimate global error, *Numer. Math.* **48**, 1-20.

ZADUNAISKY, P. E. 1966 A method for the estimation of errors propagated in the numerical solution of a system of ordinary differential equations, *Proc. Int. Astron. Union, Symp. No. 25 (Thessaloniki, 1964)*. New York: Academic, pp 281-287.