On the positivity of low order explicit Runge-Kutta schemes applied in splitting methods*

Alf Gerisch and Rüdiger Weiner †

December 20, 1999

Abstract

Splitting methods are a frequently used approach for the solution of large stiff initial value problems of ordinary differential equations with an additively split right-hand side function. Such systems arise, for instance, as method of lines discretizations of evolutionary partial differential equations in many applications. We consider the choice of explicit Runge-Kutta (RK) schemes in implicit-explicit splitting methods. Our main objective is the preservation of positivity in the numerical solution of linear and nonlinear positive problems while maintaining a sufficient degree of accuracy and computational efficiency. A 3-stage second order explicit RK method is proposed which has optimized positivity properties. This method compares well with standard s-stage explicit RK schemes of order s, $s = 2, 3$. It has advantages in the low accuracy range and this range is interesting for an application in splitting methods. Numerical results are presented.

Keywords: Differential Equations, Splitting Methods, Positivity, Explicit Runge-Kutta Methods.
Subject Classification: MSC(1991) 65L05, 65L06, 65M20.

1 Introduction

Splitting methods are a frequently used approach for the solution of large stiff initial value problems (IVPs) of ordinary differential equations (ODEs) with an additively split right-hand side function,

$$y'(t) = f_1(t, y(t)) + f_2(t, (y(t)),\ t \geq t_0,\ y(t_0) = y_0, \tag{1}$$

see e.g. [3, 8, 4]. Semi-discretizations of many evolutionary partial differential equations (PDEs) by the method of lines (MOL) result in such problems. In this case the functions f_1 and f_2 often correspond to the spatial discretization of terms of different type of a given PDE, e.g. advection and diffusion/reaction.

*This manuscript appears as Report on Numerical Mathematics, No. 35, Universität Halle–Wittenberg, see `http://www.mathematik.uni-halle.de/reports/reports.html`.

†Institut für Numerische Mathematik, Fachbereich Mathematik und Informatik, Martin–Luther Universität Halle–Wittenberg, Postfach, 06099 Halle (Saale), Germany. E-mail: {gerisch, weiner}@mathematik.uni-halle.de

Splitting methods are based on splitting the right-hand side of a given ODE into a sum of two parts which are each at least easier to handle in time integration schemes. In [4] an implicit-explicit splitting scheme is considered of the form

$$y_{n+1} = \Phi_1\left(\frac{\tau}{2}, t_n + \frac{\tau}{2}\right)\Phi_2\left(\tau, t_n\right)\Phi_1\left(\frac{\tau}{2}, t_n\right)y_n. \tag{2}$$

Here, an approximation y_n of $y(t_n)$ is advanced by a time step τ to yield y_{n+1} as an approximation of $y(t_n+\tau)$. Φ_1 and Φ_2 are approximate evolution operators of f_1 and f_2, respectively, such that (for $i = 1, 2$) $v := \Phi_i(\tau, \tilde{t})u$ approximates the solution $y(\tilde{t} + \tau)$ of $y'(t) = f_i(t, y(t)), t \geq \tilde{t}$ with initial condition $y(\tilde{t}) = u$. If the operators Φ_i are at least second order accurate approximations of the exact evolution operators associated with the f_i then the order of consistency of the approximation (2) is two.

The above scheme is applied to an ODE system with a right-hand side function which splits into a part f_1 corresponding to advection and a part f_2 representing diffusion and reaction in [4]. Therefore, in order to avoid a time step restriction by stability, the operator Φ_2 is chosen as an implicit method, the linearly implicit trapezoidal splitting method [3, 8]. Further, the ODE system is often of a type such that a non-negative initial condition evolves in time without becoming negative (e.g. [4, 9]). This is the case, for example, if a PDE models concentrations or densities and the semi-discretization is done positivity preserving. This qualitative property of the exact solution should carry over to the approximate numerical solution and because of the sequential character of (2) we have to ensure it simply for both approximate evolution operators. Here we assume that this property holds for Φ_2. An explicit method is chosen for Φ_1 because f_1 stems from the discretization of an advection operator in the application [4] and explicit methods are in general more efficient for such problems. In this paper we are concerned with the selection of an efficient explicit method (Φ_1) which is sufficiently accurate for an application in (2) and preserves non-negative initial conditions of an IVP for reasonably large time step sizes.

In the remainder of this paper we consider the solution of the IVP

$$y'(t) = f(t, y(t)),\ t \geq t_0,\ y(t_0) = y_0, \tag{3}$$

where $n \in \mathbb{N},\ t_0 \in \mathbb{R},\ y_0 \in \mathbb{R}^n$ and $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ are given. Let f has the property

$$f \text{ is continuous and (3) has a unique solution for all } t_0 \in \mathbb{R} \text{ and all } y_0 \in \mathbb{R}^n. \tag{4}$$

The relations $\geq,\ >$, etc. are meant for each component of matrices or vectors throughout this paper.

Definition 1 (positive ODE system, IVP)
The ODE system in (3) as well as the IVP (3) are called positive *if f has the property (4) and $y(t) \geq 0$ holds for all $t \geq t_0$ whenever $y_0 \geq 0$.*

Lemma 1 ([6])
Let f satisfy condition (4). The IVP (3) associated with this function is positive if and only if for all t and any vector $v \in \mathbb{R}^n$ and all $i = 1(1)n$, holds

$$v_i = 0,\ v_j \geq 0 \textit{ for all } j \neq i \quad \Rightarrow \quad f_i(t, v) \geq 0.$$

We denote with $\mathcal{P}$ the set of functions f for which the corresponding IVP (3) is positive. For $g : \mathbb{R} \to \mathbb{R}^n$ continuous and $g(t) \geq 0$ for all t we denote with $\mathcal{L}_g^+$ the subset of linear problems of $\mathcal{P}$ where $f(t,y) = Py + g(t),\ P \in \mathbb{R}^{n,n}$.

Corollary 1
Let $g : \mathbb{R} \to \mathbb{R}^n$ be a continuous function such that $g(t) \geq 0$ for all t. $f \in \mathcal{L}_g^+$ if and only if $f(t,y) = Py + g(t),\ P \in \mathbb{R}^{n,n}$ and $P - \mathrm{diag}P \geq 0$.

Positive ODE systems arise in a great variety of applications, e.g. when modelling chemical reactions, in the semi-discretization of air pollution [9] and biomathematical models [4]. The quantity $y(t)$ usually describes the concentration or density of some species. In such a situation we are naturally interested in obtaining non-negative numerical approximations y_m of the solution $y(t_m)$ at discrete time points t_m by an appropriate numerical method. This requirement is not met in general. We consider explicit Runge-Kutta (ERK) methods for the solution of (3) in this paper; for multi-step methods see for instance [7, 1].
An s-stage Runge-Kutta (RK) method for the solution of (3) can be characterized by a coefficient matrix $A \in \mathbb{R}^{s,s}$ (with $a_{ij} = 0$ for all $j \geq i$ for explicit methods), a weight vector $b \in \mathbb{R}^s$ and a knot vector $c \in \mathbb{R}^s$ defined by $c_i := \sum_{j=1}^s a_{ij}$. In short, such a scheme is often represented by its Butcher array

$$\begin{array}{c|c} c & A \\ \hline & b \end{array}$$

or by the pair (A, b) and it advances a given approximation y_n of $y(t_n)$ by a time step τ via

$$\begin{aligned} y_{n+1}^i &= y_n + \tau \sum_{j=1}^s a_{ij} f(t_n + c_j\tau, y_{n+1}^j),\ i = 1(1)s, \\ y_{n+1} &= y_n + \tau \sum_{i=1}^s b_i f(t_n + c_i\tau, y_{n+1}^i). \end{aligned}$$

Any explicit RK method can be written as a convex combination of forward Euler steps, [12]. Let $\alpha_{ij} \geq 0$ be given for $i = 2(1)s+1$ and $j = 1(1)i-1$ such that $\sum_{j=1}^{i-1} \alpha_{ij} = 1$ and denote $a_{s+1,j} := b_j$ for $j = 1(1)s$. Then holds with $y^{(1)} := y_n$,

$$y^{(i)} := \sum_{j=1}^{i-1} (\alpha_{ij} y^{(j)} + \tau\beta_{ij} f(t_n + c_j\tau, y^{(j)})), \quad i = 2(1)s+1, \tag{5}$$

that $y_{n+1} = y^{(s+1)}$. Here β_{ij} is defined as $\beta_{ij} := a_{ij} - \sum_{l=j+1}^{i-1} \alpha_{il} a_{lj}$.

Definition 2 (positive method, [6])
Let there be given a one-step method, $\mathcal{F} \subset \mathcal{P}$ and $0 < H \leq \infty$. The method is called positive *on $\mathcal{F}$ with threshold H if the numerical approximations obtained by the method are uniquely defined and are non-negative whenever the method is applied to the IVP (3) with any $f \in \mathcal{F},\ t_0 \in \mathbb{R},\ y_0 \geq 0$ and with step size at most H. If this holds with $H = \infty$ then the method is called* unconditionally positive, *otherwise* conditionally positive.

In this paper we say that a method taken from a class of methods has *optimal positivity* on a certain problem class if it is a positive method on this problem class with a step size restriction H and all other methods from the given class have, for positivity on the problem class, a step size restriction $\tilde{H} \leq H$.
Results on positivity of numerical methods applied to linear problem sets $\mathcal{L}_g^+$ can be found in [1], regarding nonlinear problems we refer to [9, 6]. We present some of these results in Section 2. The motivation for this work stems from the application of time-split methods (2) to IVPs (1) arising as MOL discretization of coupled hyperbolic-parabolic PDE systems. The MOL approach already introduces a spatial error in the solution process. Therefore and because (2) is at most second order accurate, we are searching for explicit methods Φ_1 with low to modest order of accuracy only (order 2 or 3). More important are the positivity properties and in Section 3 we study a class of explicit Runge-Kutta methods from this point of view. The method obtained and standard methods are then compared and evaluated in Section 4. We use the linear, scalar advection equation and the biomathematical model describing tumor angiogenesis from [4] as test cases. Finally, we discuss our results and draw some conclusions in Section 5.

2 General results on positivity of ERK methods

Definition 3 (absolute monotonicity and threshold factor, [10])
A rational function $R(z)$ is said to be absolutely monotonic *at a point $z \in \mathbb{R}$ if R has no pole and is non-negative in z and all its derivatives exist and are non-negative in z. The* threshold factor of *$R(z)$, $T(R)$, is defined as* $T(R) = \sup\{r | r = 0$ *or* $(r > 0,\ R$ *is absolutely monotonic* $\forall z \in [-r, 0])\}$.

The application of an ERK method with step size τ to a problem taken from class $\mathcal{L}_g^+$ results in a scheme of the form

$$y_{n+1} = R(\tau P)y_n + \tau \sum_{i=1}^{s} R_i(\tau P) g(t_n + c_i \tau). \tag{6}$$

Here $R(z)$ is a polynomial approximation to e^z around $z = 0$ and it is called the stability function of the method. The functions $R_i(z)$ are also polynomials.
We introduce subclasses of the problem class $\mathcal{L}_g^+$ and define for $\alpha \geq 0$

$$\mathcal{L}_g^+(\alpha) := \{f \in \mathcal{L}_g^+ | f(t, y) = Py + g(t) \text{ where } p_{ii} \geq -\alpha,\ i = 1(1)n\}.$$

The theory of Bolley and Crouzeix in [1] leads us to the following theorem.

Theorem 1 ([1])
Let $R(z)$ and $R_i(z)$ be the polynomials in (6) of an ERK method, and assume that they are absolutely monotonic on the interval $[-\mu, 0]$ for a value $\mu \geq 0$. Then the method (6) is positive on $\mathcal{L}_g^+(\alpha)$ under the step size restriction $\alpha\tau \leq \mu$.

Absolute monotonicity of the polynomials $R(z)$ and $R_i(z)$ of an ERK method is not sufficient to guarantee positivity of the method when applied to more general problem classes. An important quantity of RK methods with respect to nonlinear problems is the *radius of absolute monotonicity of a RK method*, which we denote by $T(A, b)$ where (A, b) is the RK method at hand (see Section 1).

Definition 4 (Radius of absolute monotonicity of an RK method, [11])
Let an RK method (A, b) be given and denote $\mathbb{1} = (1, 1, \ldots, 1)^{\mathsf{T}}$. $T(A, b) := \zeta$ where $\zeta > 0$ is the largest possible value such that for all $z \in [-\zeta, 0]$ the RK scheme is absolutely monotonic in z, that is

$$\begin{aligned} & (I - zA)^{-1} \text{ exists}, \\ R(z) &= 1 + zb^T(I - zA)^{-1}\mathbb{1} \geq 0, \\ A(z) &= A(I - zA)^{-1} \geq 0, \\ b(z) &= b^T(I - zA)^{-1} \geq 0, \\ e(z) &= (I - zA)^{-1}\mathbb{1} \geq 0. \end{aligned}$$

If no such ζ exists then we set $T(A, b) := 0$.

The first condition is always satisfied for ERK schemes.
This radius is used by Kraaijevanger [11] in the study of contractivity of RK methods and also used in the nonlinear positivity theory for RK methods by Horvath [6]. It holds that the threshold factor of the stability function $R(z)$ of an RK method is greater than or equal to the radius of absolute monotonicity of this method, $T(R) \geq T(A, b)$. Further, $T(A, b) > 0$ is necessary for contractivity resp. positivity of the RK method when applied to certain subclasses of dissipative problem sets. We give the following lemma with statements from [11]; for concepts of reducibility of RK methods we refer to [2].

Lemma 2 ([11])
For irreducible RK methods (A, b) holds:

1. $T(A, b) > 0 \quad \Leftrightarrow \quad A \geq 0, b > 0$ *and* $\forall i, j (A^2_{ij} \neq 0 \Rightarrow A_{ij} \neq 0)$.
2. *Let $r > 0$. Then $T(A, b) \geq r \quad \Leftrightarrow \quad (A, b)$ is absolutely monotonic in $-r$ and $A \geq 0$.*

In [9] the positivity of ERK methods is studied using the reformulation of the method as convex combination of forward Euler steps, see (5).

Lemma 3 (see also [9])
Let (A, b) be a given ERK scheme and assume that the coefficients β_{ij} in (5) are non-negative. Consider a positive ODE $y'(t) = f(t, y(t))$. If $u + \tau f(t, u) \geq 0$ for all $u \geq 0$, all t and all step sizes $0 < \tau \leq \tau_0$ then the ERK method (A, b) is positive for the given ODE under the step size restriction

$$\tau \leq \min_{1 \leq j < i \leq s+1} \frac{\alpha_{ij}}{\beta_{ij}} \tau_0, \quad \text{where } \tfrac{\alpha_{ij}}{\beta_{ij}} := +\infty \text{ for } \beta_{ij} = 0. \tag{7}$$

Proof We show that $\tilde{u} := \alpha_{ij} u + \tau \beta_{ij} f(t_n + c_j \tau, u) \geq 0$ for all step sizes τ satisfying the condition of the lemma. If $\beta_{ij} = 0$ then this is obviously true, so assume henceforth $\beta_{ij} > 0$. If $\alpha_{ij} = 0$ then there exists no $\tau > 0$ satisfying the conditions of the lemma and if $\alpha_{ij} > 0$ then

$$\tilde{u} = \alpha_{ij} \left(u + \tau \frac{\beta_{ij}}{\alpha_{ij}} f(t_n + c_j \tau, u) \right) \geq 0$$

if $\tau \frac{\beta_{ij}}{\alpha_{ij}} \leq \tau_0$, i.e. $\tau \leq \frac{\alpha_{ij}}{\beta_{ij}} \tau_0$. □
We will refer to $\min\limits_{1 \leq j < i \leq s+1} \frac{\alpha_{ij}}{\beta_{ij}}$ as the *positivity factor* of a given ERK method (A, b) in this paper.

3 Positivity of 3-stage ERK methods

The absolute monotonicity of the stability polynomial of an ERK method is crucial with respect to the allowable time step size in order to guarantee positivity of the method when applied to the problem class $\mathcal{L}_0^+(\alpha)$. This can be seen from Theorem 1. The absolute monotonicity of polynomials is studied in [10] and it is stated that s-stage ERK methods of order s have a threshold factor $T(R) = 1$, whereas s-stage ERK methods of order $s-1$ have a threshold factor of at most 2. This means that, at the cost of just one matrix-vector product, the allowable time step size with respect to positivity of the method is doubled. Further, [10] gives the optimal stability polynomials in these two cases.

We are interested in second or third order methods here. Numerical experiments in [10] demonstrate that (on a linear test problem) the 3-stage method of order two performs more efficient with respect to positivity compared to the s-stage methods of order s for $s = 2, 3$ (optimal stability polynomial for positivity on $\mathcal{L}_0^+(\alpha)$ in each case). Therefore we consider 3-stage explicit Runge-Kutta methods of order two with optimal stability polynomial for positivity on $\mathcal{L}_0^+(\alpha)$ in this section. We will use the free parameters in this class of methods to satisfy nonlinear positivity conditions and further order conditions.

Consider a 3-stage ERK method (A, b). The conditions for order two are $\sum_i b_i = 1$ and $\sum_i b_i c_i = \frac{1}{2}$. The stability polynomial of a 3-stage ERK method of order two is $R_{3,2}(z) = 1 + z + \frac{1}{2}z^2 + b_3 a_{32} a_{21} z^3$. On the other hand, the optimal stability polynomial for 3-stage ERK methods of order two with respect to positivity on the problem class $\mathcal{L}_0^+(\alpha)$ is $R_{3,2}^+(z) = 1 + z + \frac{1}{2}z^2 + \frac{1}{12}z^3$. Hence, beside the two order conditions, the parameters of the method have to satisfy $b_3 a_{32} a_{21} = \frac{1}{12}$. Solving for these three conditions results in the methods given by the Butcher array in Figure 1 (left); denote $\gamma := b_3 a_{32}$.

$$
\begin{array}{c|ccc}
 & 0 & & \\
 & \frac{1}{12\gamma} & 0 & \\
 & \frac{1}{b_3}\left(\frac{1}{2} - \frac{b_2}{12\gamma} - \gamma\right) & \frac{\gamma}{b_3} & 0 \\
\hline
 & 1 - b_2 - b_3 & b_2 & b_3
\end{array}
\quad \left(\begin{array}{l} b_2, b_3, \gamma \in \mathbb{R}, \\ b_3, \gamma \neq 0 \end{array}\right)
\qquad
\begin{array}{c|ccc}
 & 0 & & \\
 & \frac{1}{12\gamma} & 0 & \\
 & \frac{1}{b_3}\left(\frac{1}{2} - \frac{b_2}{12\gamma} - \gamma\right) & \frac{\gamma}{b_3} & 0 \\
\hline
 & 1 - b_2 - b_3 & b_2 & \frac{(6\gamma - b_2)^2}{48\gamma^2 - b_2}
\end{array}
\quad \left(\begin{array}{l} b_2, \gamma \in \mathbb{R}, \\ \gamma \neq 0,\ b_2 \neq 6\gamma, \\ b_2 \neq 48\gamma^2 \end{array}\right)
$$

Figure 1: Butcher array of a general 3-stage, second order ERK method with optimal positivity on the class $\mathcal{L}_0^+(\alpha)$ (left) and the same but with an additional order three condition satisfied (right).

We will further restrict the range of values for the parameters in Figure 1 (left) in Subsection 3.1. The aim is to satisfy the conditions of Lemma 2 for $T(A, b) > 0$. Following this we construct the method with $T(A, b) = 2$, and find the optimal method with respect to Lemma 3. It will turn out that both methods are identically. Finally, we consider positivity of the methods on the class $\mathcal{L}_g^+(\alpha)$.

It is also possible to use the free parameters b_2, b_3 and γ in the methods of Figure 1 (left) to satisfy one order three condition. The third order condition $\sum_{i,j} b_i a_{ij} c_j = b_3 a_{32} a_{21} = \frac{1}{6}$ cannot be satisfied because of the condition on the stability polynomial. However, the other third order condition $\sum_i b_i c_i^2 = \frac{1}{3}$ can be satisfied. We have $b_3 c_3 = \frac{1}{2} - b_2 c_2$. Substituting this in the third

order condition yields $\frac{1}{3} = b_2 c_2^2 + \frac{1}{b_3}\left(\frac{1}{2} - b_2 c_2\right)^2$. Employing $c_2 = \frac{1}{12\gamma}$, we arrive after some calculations at the methods given in Figure 1 (right). In Subsection 3.2, we state some results concerning positivity properties of this class of methods.
Finally, in Subsection 3.3 we consider the linear stability properties of the methods from Figure 1.

3.1 Nonlinear positivity of the methods from Figure 1 (left)

The radius of absolute monotonicity $T(A, b)$

Lemma 4
Let (A, b) be an irreducible scheme from Figure 1 (left). It holds $T(A, b) > 0$ if and only if

$$b_3 \in (0, 1), \quad b_2 \in \left(0, \frac{3}{4}\right), \quad b_2 + b_3 < 1, \text{ and } \gamma \in \left(\frac{1}{4} - \sqrt{\frac{1}{16} - \frac{b_2}{12}}, \frac{1}{4} + \sqrt{\frac{1}{16} - \frac{b_2}{12}}\right).$$

Proof By Lemma 2 holds $T(A, b) > 0 \Leftrightarrow A \geq 0, b > 0$ and $\forall i, j (A_{ij}^2 \neq 0 \Rightarrow A_{ij} \neq 0)$. We show that the latter holds if and only if the conditions given in the lemma hold.
$\Rightarrow$
$b > 0 \Rightarrow b_1 = 1 - b_2 - b_3, b_2, b_3 > 0 \Rightarrow b_2, b_3 < 1$ and $b_2 + b_3 < 1$.
$\gamma \neq 0$ and $a_{21} \geq 0 \Rightarrow \gamma > 0 \Rightarrow a_{21} > 0$ and with $b_3 > 0$ also $a_{32} > 0$. Further $A_{ij}^2 \neq 0$ only for $i = 3, j = 1$, that is $A_{31}^2 = a_{21}a_{32} \neq 0$, and this implies that also $a_{31} > 0$ holds.
$a_{31} > 0 \Rightarrow \gamma^2 - \frac{1}{2}\gamma + \frac{b_2}{12} < 0 \Rightarrow b_2 < \frac{3}{4}$ and $\gamma \in \left(\frac{1}{4} - \sqrt{\frac{1}{16} - \frac{b_2}{12}}, \frac{1}{4} + \sqrt{\frac{1}{16} - \frac{b_2}{12}}\right)$.
$\Leftarrow$
The conditions on the b_i imply $b > 0$. $b_2 < \frac{3}{4} \Rightarrow \gamma > 0 \Rightarrow a_{32} > 0$ and $a_{21} > 0$. Further, $a_{31} > 0$ follows with the conditions on γ. □

Remark 1 *We see that all schemes satisfying the conditions of Lemma 4 are irreducible because $b > 0$ and the subdiagonal part of A is positive.*

For 3-stage ERK methods (A, b) we obtain

$$(I - zA)^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -za_{21} & 1 & 0 \\ -za_{31} & -za_{32} & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ za_{21} & 1 & 0 \\ za_{31} + z^2 a_{32} a_{21} & za_{32} & 1 \end{pmatrix}.$$

We already know that $T(A, b) \leq 2$ for the methods under consideration because that is the threshold factor $T(R)$ of the stability function $R(z)$. Let (A, b) be an irreducible 3-stage ERK method with $T(A, b) > 0$. Then (A, b) is absolutely monotonic in z ($z \in [-2, 0]$) if and only if the following conditions are satisfied (see Definition 4):

$$\begin{aligned} A(z) \;&:\; C_1(z) := a_{31} + z a_{32} a_{21} \geq 0, \\ b(z) \;&:\; C_2(z) := b_1 + z(b_2 a_{21} + b_3 a_{31}) + z^2 b_3 a_{32} a_{21} \geq 0, \\ &\;\; C_3(z) := b_2 + z b_3 a_{32} \geq 0, \\ e(z) \;&:\; C_4(z) := 1 + z a_{21} \geq 0, \\ &\;\; C_5(z) := 1 + z(a_{31} + a_{32}) + z^2 a_{32} a_{21} \geq 0. \end{aligned} \tag{8}$$

Lemma 5
Let (A, b) be a scheme from Figure 1 (left) satisfying the conditions of Lemma 4. Then (A, b) is absolutely monotonic in z ($z \in [-2, 0]$) if and only if $C_i(z) \geq 0$, $i = 1(1)5$, and this is the case if and only if

$$\begin{aligned} z &\geq 12\gamma - 6 + \frac{b_2}{\gamma}, \\ 0 &\leq 1 - b_2 - b_3 + \left(\frac{1}{2} - \gamma\right) z + \frac{1}{12} z^2, \\ z &\geq -\frac{b_2}{\gamma}, \\ z &\geq -12\gamma, \\ 0 &\leq z^2 + \left(6 - \frac{b_2}{\gamma}\right) z + 12 b_3. \end{aligned}$$

Proof The statement follows by simplifying and rearranging the conditions $C_i(z) \geq 0$, $i = 1(1)5$, by using the method coefficients given in Figure 1 (left). □

Theorem 2
Let (A, b) be an irreducible scheme from Figure 1 (left). Then $T(A, b) = 2$ if and only if $b_2 = b_3 = \frac{1}{3}$ and $\gamma = \frac{1}{6}$.

Proof Let $T(A, b) = 2$. Then the conditions of Lemma 4 are satisfied and (A, b) is absolutely monotonic in $z = -2$. Hence the conditions of Lemma 5 are satisfied for $z = -2$. The third condition implies $\gamma \leq \frac{b_2}{2}$ and the fourth condition $\gamma \geq \frac{1}{6}$. These two bounds on γ make $b_2 \geq \frac{1}{3}$ necessary. The first condition of Lemma 5 for $z = -2$ implies

$$0 \geq \gamma^2 - \frac{1}{3}\gamma + \frac{b_2}{12} =: p(\gamma).$$

It is $p(0) > 0$ and the discriminant is given by $D = \frac{1}{36} - \frac{b_2}{12}$. In order to have the condition satisfied for some γ we need $D \geq 0$ and this is the case only for $b_2 \leq \frac{1}{3}$. Hence $b_2 = \frac{1}{3}$ is necessary and this immediately implies $\gamma = \frac{1}{6}$. The fifth condition of Lemma 5 for $z = -2$ now implies $b_3 \geq \frac{1}{3}$ whereas the second condition requires $b_3 \leq \frac{1}{3}$. Hence we need $b_3 = \frac{1}{3}$
On the other hand, if (A, b) is the scheme from Figure 1 (left) with $b_2 = b_3 = \frac{1}{3}$ and $\gamma = \frac{1}{6}$ then (A, b) is irreducible and satisfies the conditions of Lemma 4. We already know that $T(A, b) \leq 2$. With Lemma 2 holds $T(A, b) \geq 2$ if (A, b) is absolutely monotonic in $z = -2$ and this is the case if the conditions of Lemma 5 are satisfied for $z = -2$. By inspection we see that this is the case. □

The positivity factor

We will investigate the positivity factor (appearing in Lemma 3) of the class of ERK methods (A, b) from Figure 1 (left) with $T(A, b) > 0$ (see Lemma 4). Our aim is to construct methods with factor 2 and therefore we have to show that $\frac{\alpha_{ij}}{\beta_{ij}} \geq 2$ for all $1 \leq j < i \leq 4$, see also Equation (5).

We have $\alpha_{21} = 1$ and $\beta_{21} = \frac{1}{12\gamma} > 0$. Hence $\frac{\alpha_{21}}{\beta_{21}} \geq 2 \Leftrightarrow \gamma \geq \frac{1}{6}$. It is $\beta_{32} = a_{32} = \frac{\gamma}{b_3} > 0$ and hence $\frac{\alpha_{32}}{\beta_{32}} \geq 2 \Leftrightarrow \alpha_{32} \geq \frac{2\gamma}{b_3}$. Using $\alpha_{32} \leq 1$ this implies $b_3 \geq 2\gamma \geq \frac{1}{3}$.
It is $\beta_{43} = b_3 > 0$ and $\frac{\alpha_{43}}{\beta_{43}} \geq 2 \Leftrightarrow \alpha_{43} \geq 2b_3$. Using $\alpha_{43} \leq 1$ implies $b_3 \leq \frac{1}{2}$ and hence $\gamma \leq \frac{1}{4}$.
It is $\beta_{42} = b_2 - \alpha_{43}\frac{\gamma}{b_3} \geq 0 \Leftrightarrow \alpha_{43} \leq \frac{b_2 b_3}{\gamma}$. This implies $2\gamma \leq b_2$ and hence $b_2 \geq \frac{1}{3}$ and $b_1 \leq \frac{1}{3}$.
It is $\beta_{31} = a_{31} - a_{21}\alpha_{32} \geq 0 \Leftrightarrow \alpha_{32} \leq \frac{a_{31}}{a_{21}} = \frac{1}{b_3}(6\gamma - 12\gamma^2 - b_2)$. Using $b_2 \geq 2\gamma$ implies $\alpha_{32} \leq \frac{2\gamma}{b_3}(2 - 6\gamma)$ and $\gamma \geq \frac{1}{6}$ leads to $\alpha_{32} \leq \frac{2\gamma}{b_3}$. Hence we need $\alpha_{32} = \frac{2\gamma}{b_3}$. Now we can simplify $\beta_{31} = \frac{1}{b_3}\left(\frac{1}{3} - \frac{b_2}{12\gamma} - \gamma\right)$.
Case 1: $\beta_{31} = 0$
This is only possible for $\gamma = \frac{1}{6}$ and $b_2 = \frac{1}{3}$ because we already know $b_2 \geq \frac{1}{3}$. Now we have $\frac{b_2 b_3}{\gamma} = 2b_3$ and therefore $\alpha_{43} = 2b_3$. This also gives $\beta_{42} = 0$. It is $\beta_{41} = b_1 - a_{21}\alpha_{42} - a_{31}\alpha_{43} = \frac{1}{3} - b_3 - \frac{\alpha_{42}}{2} \geq 0 \Leftrightarrow b_3 \leq \frac{1}{3} - \frac{\alpha_{42}}{2}$. Using $b_3 \geq \frac{1}{3}$ leads to $b_3 = \frac{1}{3}$ and $\alpha_{42} = 0$. Hence we obtain the method $(b_2, b_3, \gamma) = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{6}\right)$.
Case 2: $\beta_{31} > 0$
This is only possible if $b_2 < \frac{1}{3}$ and this is in contradiction with $b_2 \geq \frac{1}{3}$.
Hence the only method (A, b) from Figure 1 (left) with $T(A, b) > 0$ and positivity factor 2 is given by

$$(b_2, b_3, \gamma) = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{6}\right), \text{ with } \alpha_{21} = 1, \alpha_{31} = 0, \alpha_{32} = 1, \alpha_{41} = \frac{1}{3}, \alpha_{42} = 0, \text{ and } \alpha_{43} = \frac{2}{3}.$$

Positivity on $\mathcal{L}_g^+(\alpha)$

We will apply the class of ERK methods (A, b) from Figure 1 (left) with $T(A, b) > 0$ (see Lemma 4) to the problem class $\mathcal{L}_g^+(\alpha)$. The polynomials $R_i(z)$, $i = 1, 2, 3$ in (6) of a 3-stage ERK method are given by

$$R_1(z) = b_1 + (b_2 a_{21} + b_3 a_{31})z + b_3 a_{32} a_{21} z^2, \quad R_2(z) = b_2 + b_3 a_{32} z, \quad R_3(z) = b_3, \tag{9}$$

and simplify for methods from Figure 1 (left) to

$$R_1(z) = 1 - b_2 - b_3 + \left(\frac{1}{2} - \gamma\right) z + \frac{1}{12} z^2, \quad R_2(z) = b_2 + \gamma z, \quad R_3(z) = b_3.$$

For optimal positivity of the methods applied to problems from $\mathcal{L}_g^+(\alpha)$ we need, according to Theorem 1, that the threshold factor $T(R_i)$ of the polynomials $R_i(z)$ is as large as possible. It is $T(R) = 2$ and therefore we are interested in methods with $T(R_i) \geq 2$. For the method derived above ($(b_2, b_3, \gamma) = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{6}\right)$) it is easily shown that Theorem 1 applies with $\mu = 2$ (and this is optimal).

3.2 Nonlinear positivity of the methods from Figure 1 (right)

The class of methods in Figure 1 (right) compared to the class in Figure 1 (left) satisfies additionally one of the order three conditions. We were able to identify the admissible range of the parameters (b_2, γ) such that $T(A, b) > 0$ holds for this class. Within this parameter range we determined numerically the method which maximizes $T(A, b)$. This is the method with $b_2 = 0.3572, \gamma = 0.3039$ leading to $T(A, b) = 1.1754$. Figure 3.2 gives a plot of $T(A, b)$ as a function of the free parameters b_2 and γ.

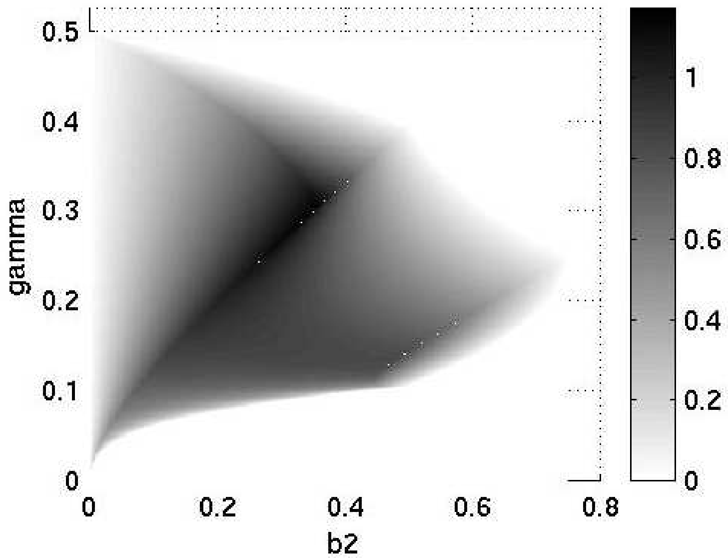

Figure 2: The radius of absolute monotonicity $T(A, b)$ of the schemes (A, b) in Figure 1 (right) as a function of the free parameters b_2 and γ.

Rewriting the ERK method from Figure 1 (right) with $b_2 = 0.3572$ and $\gamma = 0.3039$ as a convex combination of forward Euler steps (see formula (5)) with $\alpha_{3,1} = 0.3213, \alpha_{4,1} = 0.38$ and $\alpha_{4,2} = 0.0000764$ results in a positivity factor of ≈ 1.1754 (see Lemma 3).
These numerically obtained values (optimal $T(A, b)$, positivity factor) for the methods from Figure 1 (right) are slightly better than those which hold for s-stage methods of order s (all values equal one) but they are worse than the values for the optimal method from Figure 1 (left) where all values equal two. Further, the methods from Figure 1 (right) are still of second order only and the advantage of having one of the third order conditions satisfied is expected to be marginal. The numerical performance of the optimized method (with an additional third order condition satisfied) is almost the same as for the optimized method devised in the previous subsection (Figure 1 (left) with $(b_2, b_3, \gamma) = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{6}\right)$).
Therefore we will omit the results of this method in our numerical tests.

3.3 Linear stability

Compared with s-stage ERK methods of order s, the stability functions of the 3-stage methods in Figure 1 have a twice as large absolute monotonicity interval allowing for twice as large time steps with regard to positivity of the method when applied to the problem class $\mathcal{L}_0^+(\alpha)$.
We now turn our attention to the linear stability of the 3-stage ERK methods of order 2 with optimal positivity on the problem class $\mathcal{L}_0^+(\alpha)$. The linear stability region is given in Figure 3. For comparison, we also print the linear stability regions of the s-stage ERK methods of order s for $s = 2, 3$.

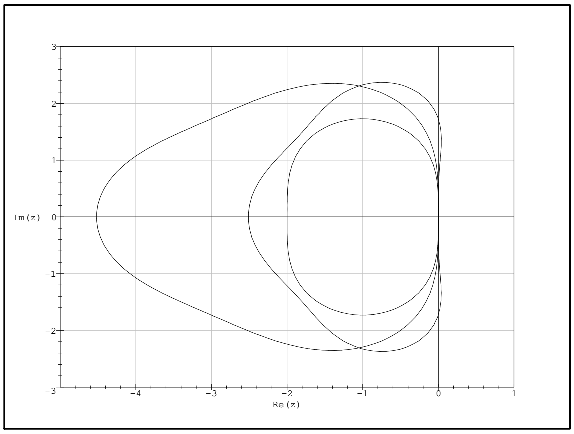

Figure 3: Linear stability regions for the 3-stage ERK methods given in Figure 1 and the s-stage ERK methods of order s for $s = 3$ and $s = 2$ (largest to smallest).

The region of the 3-stage methods from Figure 1 is stretched by a factor of about two in the real direction compared with the regions of the s-stage methods of order s. With respect to the imaginary direction, there is only little stretching compared to the 2-stage, second order methods and a slight disadvantage near the imaginary axis compared to the 3-stage, third order methods. Altogether, the 3-stage, 2nd order methods from Figure 1 have also favourable linear stability properties.

4 Numerical experiments

In the previous section we have constructed 3-stage ERK methods of order 2 with optimal positivity on $\mathcal{L}_0^+(\alpha)$. Using additional conditions which are sufficient for positivity of the methods on other problem sets (non-autonomous, non-linear) we have identified a method with favourable properties. This is the method from Figure 1 (left) with $b_2 = b_3 = 2\gamma = \frac{1}{3}$, which we will refer to as RK32, see Figure 4.

We compare this method with the following schemes: modified Euler (ME, two stages, second order) and Runge-Kutta-Fehlberg method 2(3) (RKF2(3), three stages, third order, see [9, 5]). Both methods have $T(R) = T(A, b) = 1$, and positivity factor 1.

We have selected two test examples. Firstly, the performance of the ERK methods is evaluated with respect to accuracy, positivity and efficiency on a MOL approximation of the scalar, linear advection equation in Section 4.1. Here the ERK methods are applied with constant time step sizes. Secondly, in Section 4.2 we consider the solution of a coupled hyperbolic-parabolic PDE system from [4] with an implicit-explicit splitting scheme of the form (2). The ERK methods are used for the explicit part of this scheme. A time step size control is used in the splitting scheme.

$$
\begin{array}{c|cc}
0 & 0 & \\
1 & 1 & 0 \\
\hline
 & \frac{1}{2} & \frac{1}{2} \\
\hline
 & 1 & 0
\end{array}
\qquad
\begin{array}{c|ccc}
0 & 0 & & \\
1 & 1 & 0 & \\
\frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\
\hline
 & \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \\
\hline
 & \frac{1}{2} & \frac{1}{2} & 0
\end{array}
\qquad
\begin{array}{c|ccc}
0 & 0 & & \\
\frac{1}{2} & \frac{1}{2} & 0 & \\
1 & \frac{1}{2} & \frac{1}{2} & 0 \\
\hline
 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\
\hline
 & \frac{1}{2} & \frac{1}{2} & 0
\end{array}
$$

Figure 4: Butcher arrays for ME, RKF2(3), and RK32 (from left to right). The last row of each array defines an embedded method.

4.1 Scalar, linear advection equation

We consider the scalar, linear advection equation in one space dimension

$$
\begin{aligned}
u_t(t,x) + u_x(t,x) &= 0 \text{ for } x \in (0,1),\ t > 0, \\
u(0,x) &= u_0(x) \text{ for } x \in [0,1], \\
u(t,0) &= u_0(0) \text{ for } t \geq 0.
\end{aligned}
$$

We use two different initial conditions, a block and a smooth profile:

$$
u_0(x) = \begin{cases} 1 & \text{for } 0.3 \leq x \leq 0.6 \\ 0 & \text{otherwise} \end{cases}, \qquad \text{and} \qquad u_0(x) = \sin^2(\pi x).
$$

Our final time is $t_f = \frac{1}{4}$ and we discretize the spatial derivative on an equidistant spatial grid with mesh width $\Delta x = \frac{1}{100}$ employing the positive, 3rd order upwind discretization with van Leer flux limiter as described in [9, 4]. The result of this discretization is a nonlinear, positive, autonomous ODE system.

We test the methods with fixed time step sizes τ and perform $k = 15, 20, \ldots, 40, 50$ steps. This leads to values of $\tau \approx 0.017, \ldots, 0.005$ and Courant (CFL) numbers $\nu := \frac{\tau}{\Delta x} = \frac{25}{k} \approx 1.7, \ldots, 0.5$. The 3-stage methods require 3 right-hand side evaluations per time step and the 2-stage method only 2.

In Table 1 we give the scaled l_2-norm ($\|v\|_2 = (\frac{1}{n}\sum_{i=1}^{n} v_i^2)^{1/2}$ for $v \in \mathbb{R}^n$) and the ∞-norm ($\|v\|_\infty = \max_{i=1(1)n} |v_i|$ for $v \in \mathbb{R}^n$) of the errors of the numerical approximations with respect to a high accuracy solution of the ODE (obtained with the code DOPRI5 (see [5]), time discretization error) and with respect to the exact solution of the PDE (full time-space error). Further, in order to characterize positivity, the value of the smallest component of the solution is given. The spatial error of the semi-discretization (estimated by the difference between exact solution of the PDE and high accuracy solution of the ODE) is approximately 10^{-1} in the scaled l_2-norm and $10^{-0.36}$ in the ∞-norm for the block profile and $10^{-3.1}$ and $10^{-2.5}$, respectively, for the smooth profile.

We call an approximate solution *positive* if the smallest component is greater than -10^{-8}. In the tables this is marked by a horizontal line.

The CFL threshold for positivity is 1 for ME and ≈ 0.7 for the three stage, third order method RKF2(3). RK32 has a CFL limit of ≈ 1.25 for the block profile and of ≈ 1.7 for the smooth profile), and hence larger steps are possible.

k	$\|te\|_2$	$\|fe\|_2$	$\|te\|_\infty$	$\|fe\|_\infty$	$\min_i u_i$
ME					
15	5.96	5.96	6.44	6.44	$-10^{6.4}$
20	2.49	2.49	2.92	2.92	$-10^{2.9}$
25	−1.04	−0.81	−0.55	−0.30	0
30	−1.48	−0.94	−0.90	−0.27	0
35	−1.84	−1.00	−1.21	−0.31	$-10^{-30.4}$
40	−2.02	−1.01	−1.36	−0.32	$-10^{-30.4}$
50	−2.22	−1.03	−1.55	−0.34	$-10^{-30.6}$
RKF3(2)					
15	5.37	5.37	5.84	5.84	$-10^{5.8}$
20	−1.06	−0.84	−0.44	−0.32	$-10^{-1.4}$
25	−1.96	−1.01	−1.31	−0.36	$-10^{-3.3}$
30	−2.56	−1.03	−1.92	−0.36	$-10^{-5.4}$
35	−2.81	−1.04	−2.17	−0.36	0
40	−3.00	−1.04	−2.36	−0.36	$-10^{-30.4}$
50	−3.32	−1.04	−2.69	−0.36	0
RK32					
15	−1.13	−0.84	−0.60	−0.31	$-10^{-6.8}$
20	−1.72	−0.98	−1.06	−0.29	$-10^{-30.2}$
25	−1.93	−1.01	−1.28	−0.31	$-10^{-30.3}$
30	−2.08	−1.02	−1.42	−0.33	$-10^{-30.2}$
35	−2.21	−1.03	−1.54	−0.33	$-10^{-30.2}$
40	−2.32	−1.03	−1.65	−0.34	$-10^{-30.3}$
50	−2.51	−1.03	−1.84	−0.35	$-10^{-30.5}$

k	$\|te\|_2$	$\|fe\|_2$	$\|te\|_\infty$	$\|fe\|_\infty$	$\min_i u_i$
ME					
15	2.98	2.98	3.59	3.59	$-10^{3.6}$
20	0.67	0.67	1.05	1.05	$-10^{1.0}$
25	−1.87	−1.87	−1.40	−1.40	$+10^{-12.2}$
30	−3.00	−2.81	−2.43	−2.29	$+10^{-13.5}$
35	−3.49	−3.01	−2.93	−2.45	$+10^{-14.9}$
40	−3.60	−3.04	−3.03	−2.47	$+10^{-16.2}$
50	−3.79	−3.07	−3.21	−2.48	$+10^{-18.0}$
RKF3(2)					
15	3.00	3.00	3.52	3.52	$-10^{3.5}$
20	−1.58	−1.58	−0.93	−0.93	$-10^{-4.3}$
25	−3.73	−3.13	−3.08	−2.51	$-10^{-5.1}$
30	−4.20	−3.17	−3.48	−2.54	$-10^{-7.5}$
35	−4.34	−3.16	−3.61	−2.53	$-10^{-31.1}$
40	−4.47	−3.16	−3.75	−2.52	$+10^{-29.9}$
50	−4.74	−3.15	−4.04	−2.51	$+10^{-26.4}$
RK32					
15	−2.04	−2.03	−1.45	−1.45	$+10^{-10.7}$
20	−3.29	−2.94	−2.73	−2.37	$+10^{-13.4}$
25	−3.48	−3.00	−2.88	−2.43	$+10^{-14.9}$
30	−3.64	−3.05	−3.05	−2.47	$+10^{-17.1}$
35	−3.78	−3.07	−3.18	−2.48	$+10^{-18.8}$
40	−3.89	−3.09	−3.29	−2.49	$+10^{-20.1}$
50	−4.09	−3.11	−3.51	−2.49	$+10^{-21.7}$

Table 1: Results for the scalar, linear advection equation with block initial condition (left) and smooth initial condition (right). k denotes the number of time steps, $\|te\|_2$ the scaled l_2-norm of the time discretization error ($\|te\|_\infty$ the corresponding l_∞-error), $\|fe\|_2$ and $\|fe\|_\infty$ the norms of the full time-space error, and $\min u_i$ the value of the minimal component in the approximate solution at final time. The norm values are given as logarithms to the base 10.

The space-time error of the solution should balance with the spatial error introduced by approximating the spatial derivatives. Hence we require that the approximations have space-time errors which are smaller than $10^{-0.9}$ and $10^{-0.3}$ for the block solution and 10^{-3} and $10^{-2.4}$ for the smooth solution (scaled l_2- and l_∞-error, respectively). In order to obtain this, we have to limit the CFL number to 0.7 for ME and to 1 for RK32. No further reduction of the CFL number is necessary for RKF2(3).

With the obtained CFL limits for positivity and sufficient accuracy we require at least 70 right-hand side evaluations with ME, 105 with RKF2(3) and 75 with RK32. Therefore, the methods ME and RK32 are equally efficient for this example. However, considering the larger CFL limit of RK32, this method allows for larger time steps. This may pay off if it is employed in time splitting methods, especially in the low accuracy range (because larger time steps are possible for the expensive implicit part of the splitting). This will be discussed in the following section.

4.2 Application in the context of implicit-explicit splitting

Here we apply the ERK methods ME, RKF2(3) and the optimized method RK32 as part of an implicit-explicit time stepping scheme for the solution of the ODE system arising after semi-discretization of a hyperbolic-parabolic system of PDEs describing the process of tumor angiogenesis. We refer to [4] for details of the PDE, the spatial discretization and the employed time splitting method. We present numerical results obtained on a 150×150 and on a 200×200 spatial grid resulting in ODE systems of dimension approximately 40000 and 80000, respectively.
We expect an improved performance of the splitting scheme if the explicit method RK32 instead of ME or RKF2(3) is used within the splitting, at least for low accuracy requirements. The reason is that – although RK32 and ME perform comparable for the linear advection equation considered in the previous section – the larger time steps allowed by RK32 should pay off because we have to solve less linear systems in the implicit part of the splitting scheme in order to reach a given final time.
We use a time step size selection strategy based on estimating the local error of the approximate solution. The local error after each part of a time split step is estimated (and we reject the step if the required local accuracy is not reached for one part). We can use embedding for error estimation in the explicit schemes (explicit Euler for ME, the last stage of RK32 provides a first order embedding, and modified Euler for RKF2(3), see Figure 4). We use, as stated in the introduction, the linearly implicit trapezoidal splitting method as implicit solver in the splitting scheme. There is no obvious and cheap embedded solution provided by this scheme and therefore we use Richardson extrapolation to estimate the local error in this part of the splitting.

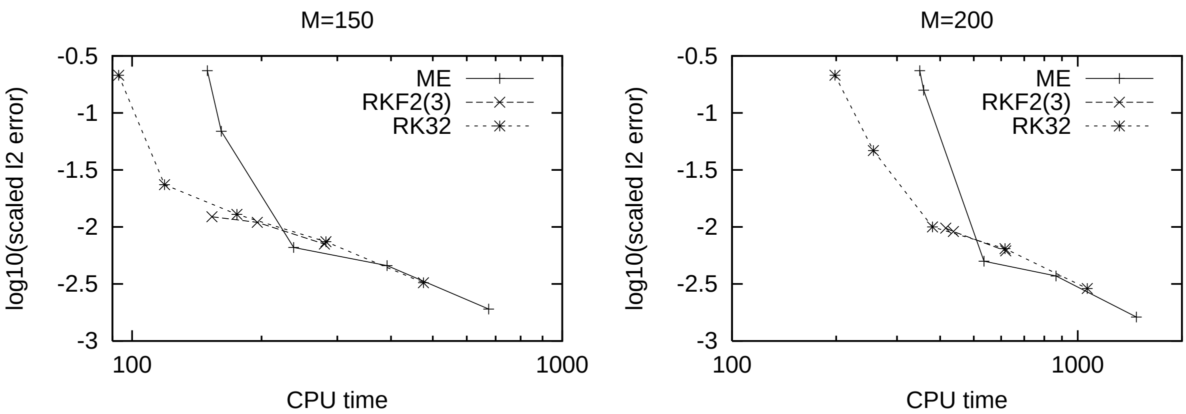

Figure 5: Accuracy vs. computation time plots for the hyperbolic-parabolic example and time splitting with step size control . The grid spacing is $1/150$ (left) and $1/200$ (right). The required accuracy ranges from $10^{-2}, 10^{-2.5}$ to 10^{-4}. The method RKF2(3) returns no solution for accuracies 10^{-2} and $10^{-2.5}$ in acceptable time.

Figure 5 gives accuracy vs. computation time plots for this example at final time 0.9 (the biological assumptions in the model hold only up to this time). The accuracy of the approximate solution is measured against a high accuracy solution of the ODE system (computed with the code DOPRI5). Hence, only time discretization and splitting errors are considered, no spatial errors. Table 2 contains the values of the most negative component of the numerical solution at time 0.9.

$\log_{10}$(tol)	M=150 ME	RKF2(3)	RK32	M=200 ME	RKF2(3)	RK32
-2	-12	$*$	-7	-9	$*$	-7
-2.5	-4	$*$	-3	-6	$*$	-4
-3	-7	-6	-7	-6	-7	-5
-3.5	-20	-7	-20	-14	-6	-8
-4	-21	-20	-21	-22	-20	-22

Table 2: Values of the most negative component in the numerical solution of the hyperbolic-parabolic problem at $t = 0.9$ for the different methods and required accuracies. These values are all negative and we give the logarithm to the base 10 of the absolute value in the table; a $*$ indicates that no solution was returned in acceptable time.

We see that there are no significant differences in the efficiency of the considered methods for high accuracy demands. However, for low demands, the optimized method RK32 demonstrates a better performance. The third order scheme fails to produce an acceptable solution in the low accuracy range. The positivity results of RK32 and ME are very similar, with a slight advantage for RK32 (compare the positivity of solutions which are computed with about the same amount of CPU time).

There should be an even greater advantage of employing the optimized method compared to the standard methods if the computations in the implicit part of the splitting get more involved because of the potentially larger time steps which can be taken.

5 Conclusions

We have discussed the choice of 3-stage ERK methods in Strang-type implicit-explicit splitting methods for the solution of MOL discretizations of evolutionary PDEs. Our main objectives were the positivity of the methods for as large as possible time steps and computational efficiency while maintaining a sufficient degree of accuracy in the solution. As a result we propose the method RK32. This method compares well with standard second or third order ERK methods.

The method RK32 allows for larger time steps in the solution of the scalar, linear advection equation compared to the other methods in order to obtain comparable accuracy and positivity in the solution.

If RK32 is applied in the more complex situation of implicit-explicit splitting methods then there is an efficiency gain compared to the standard methods in the low accuracy range. This efficiency gain is expected to be even larger if the implicit part of the splitting becomes computationally more expensive. We note especially that an acceptable solution can be obtained in less computation time by employing the devised optimized method in the splitting approach than by using the standard methods. This is important in large scale simulations or for parameter estimation. Altogether, RK32 appears to be a very reliable choice as explicit method in implicit-explicit splitting schemes.

References

[1] C. Bolley and M. Crouzeix. Conservation de la positivité lors de la discrétisation des problèmes d'évolution paraboliques. *RAIRO Anal. Numer.*, 12:237–245, 1978.

[2] G. Dahlquist and R. Jeltsch. Shifted Runge–Kutta methods and transplanted differential equations. In K. Strehmel, editor, *Numerical Treatment of Differential Equations, proceedings, Halle 1987*, pages 47–56. Teubner, Leipzig, 1988.

[3] C. Eichler-Liebenow, N. H. Cong, R. Weiner, and K. Strehmel. Linearly implicit splitting methods for higher space-dimensional parabolic differential equations. *Appl. Numer. Math.*, 28:259–274, 1998.

[4] A. Gerisch, D. F. Griffiths, R. Weiner, and M. A. J. Chaplain. Splitting Methods for Mixed Hyperbolic-Parabolic Systems. Technical Report NA/186, University of Dundee, 1998.

[5] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I.* Springer, 1987.

[6] Zoltán Horváth. Positivity of Runge–Kutta and diagonally split Runge–Kutta methods. *Appl. Numer. Math.*, 28:309–326, 1998.

[7] W. Hundsdorfer. Partial implicit BDF2 blends for convection dominated flows. Technical Report MAS-R9831, CWI, 1998.

[8] W. Hundsdorfer. Trapezoidal and midpoint splittings for initial-boundary value problems. *Math. Comput.*, 67:1047–1062, 1998.

[9] W. Hundsdorfer, B. Koren, M. van Loon, and J. G. Verwer. A positive finite-difference advection scheme. *J. Comput. Phys.*, 117:35–46, 1995.

[10] J. F. B. M. Kraaijevanger. Absolute monotonicity of polynomials occuring in the numerical solution of initial value problems. *Numer. Math.*, 48:303–322, 1986.

[11] J. F. B. M. Kraaijevanger. Contractivity of Runge-Kutta methods. *BIT*, 31:482–528, 1991.

[12] Chi-Wang Shu and Stanley Osher. Efficient Implementation of Essentially Non-oscillatory Shock-Capturing Schemes. *J. Comput. Phys.*, 77:439–471, 1988.