Numerical Algorithms (2024) 96:1391–1421
https://doi.org/10.1007/s11075-023-01730-7

ORIGINAL PAPER

The effect of splitting strategy on qualitative property preservation

Siqi Wei[1] · Raymond J. Spiteri[2]

Received: 31 August 2023 / Accepted: 11 December 2023 / Published online: 11 January 2024

Abstract
It is common for mathematical models of physical systems to possess qualitative properties such as positivity, monotonicity, or conservation of underlying physical behavior. When these models consist of differential equations, it is also common for them to be solved via splitting, i.e., splitting the differential equations into parts that are integrated separately. All splitting strategies are not created equal; however, in this work, we study the effect of two splitting strategies on qualitative property preservation applied to the basic susceptible-infected-recovered (SIR) model from epidemiology and the effect of backward integration of operator-splitting methods on positivity preservation in the Robertson test problem. We find that qualitative property preservation does depend on the splitting strategy even if the sub-integrations are performed exactly. Accordingly, the specific choice of splitting strategy used may be informed by requirements of qualitative property preservation. The choice of operator-splitting method also depends on the specific properties of the exact solution of the sub-systems.

Keywords Operator splitting · SIR model · Production-destruction system · Qualitative property preservation · High order

1 Introduction

The influence of mathematical models on modern daily life has increased in accordance with the dramatic increase in modern computing power. Many of these models

Siqi Wei and Raymond J. Spiteri are contributed equally to this work.

✉ Siqi Wei
siqi.wei@usask.ca

Raymond J. Spiteri
spiteri@cs.usask.ca

[1] Department of Mathematics and Statistics, University of Saskatchewan, Saskatoon, SK, Canada

[2] Department of Computer Science, University of Saskatchewan, Saskatoon, SK, Canada

are based on differential equations, e.g., numerical weather prediction, investment portfolio behavior, and trajectory prediction of astronomical bodies. It is not hard to imagine that these models are often large and complex and accordingly have no analytical mathematical solution. Thus, not only do their solutions need to be approximated numerically, but it is also often necessary for the model to be split into multiple parts in order to facilitate their numerical solution. For example, splitting methods are used to solve advection-diffusion-reaction problems [1] and large-scale chemical reaction systems [2].

A production-destruction system (PDS) is a system of ordinary differential equations (ODEs) that is often used in biology and chemistry to describe the production and destruction mechanism between variables.

Definition 1 *A production-destruction system of N constituents can be written as a system of differential equations of the form*

$$\frac{\mathrm{d}}{\mathrm{d}t} y_i(t) = \sum_{j=1}^{N} p_{ij}(\mathbf{y}(t)) - \sum_{j=1}^{N} d_{ij}(\mathbf{y}(t)), \qquad i = 1, 2, \ldots, N, \tag{1}$$

where $\mathbf{y} = [y_1, y_2, \ldots, y_N]^T$ is the vector of constituents. For solutions to be non-negative, we assume that the production terms p_{ij} and the destruction terms d_{ij} are non-negative. The production term p_{ij} is the rate at which constituent j transforms into constituent i, and the destruction term d_{ij} is the rate at which constituent i transforms into constituent j. We assume $p_{ij}(y), d_{ij}(y) \geq 0$ for $y_i(t) \geq 0$, $i = 1, 2, \ldots, N$ and all $t \geq 0$. A sufficient condition for $y(t)$ to be non-negative is then $\lim_{y_i \to 0^+} d_{ij}(y) = 0$, $i = 1, 2, \ldots, N$.

Definition 2 *A PDS (1) is called non-negative if non-negative initial values $y_i(0) \geq 0$ for $i = 1, 2, \ldots, N$ imply non-negative solutions $y_i(t) \geq 0$ for $i = 1, 2, \ldots, N$ for all $t > 0$.*

Definition 3 *A PDS (1) is called conservative if $p_{ij}(\mathbf{y}) = d_{ji}(\mathbf{y})$ for all $i, j = 1, 2, \ldots, N$. The system is called fully conservative if in addition $p_{ii}(\mathbf{y}) = d_{ii}(\mathbf{y}) = 0$ for all $i = 1, 2, \ldots, N$.*

For the purposes of this analysis and without loss of generality, we only consider fully conservative PDSs.

Definition 4 *Given a numerical method to solve (1), let $\mathbf{y}_n$ denote the approximation of $\mathbf{y}(t_n)$ at a time step t_n. The numerical method is called*

- *unconditionally conservative if the sum of all components of $\mathbf{y}_n$ is constant for all $n \in \mathbb{N}$ and all $\Delta t > 0$.*
- *unconditionally non-negative if all components of $\mathbf{y}_{n+1}$ is non-negative for all $\Delta t > 0$ whenever all components of $\mathbf{y}_n$ is non-negative.*

In this paper, we are interested in two examples of PDSs: the susceptible-infected-recovered (SIR) model [3] and the stiff Robertson test problem [4, 5].

Models of PDSs have qualitative properties associated with their variables such as positivity or conservation of underlying physical behavior. Models such as the SIR model further requires monotonicity of certain variables based on the model assumptions. Violation of such properties at best undermines confidence in the model or its solution/predictions; at worst, the modelling process can be invalidated.

Real-world models of complex phenomena such as the spread of disease throughout a population tend to also become complex themselves as the number of processes included or the demands on the predictions increase. Models based on differential equations must be solved numerically. Real-world problems quickly become too unwieldy solve monolithically, either due to their complexity or size, and splitting is a divide-and-conquer approach to obtain numerical solutions more efficiently (or at all) [1, 6–8].

It is well known that numerical solutions generally do not preserve known properties of the exact solution. There are some notable exceptions, however, such as the preservation of linear invariants for linear multi-step and Runge–Kutta methods [9], and a great deal of research has gone into preserving qualitative properties such as positivity, monotonicity, and symplecticity, to name but a few; see, e.g., [1, 10, 11] and references therein. Such methods are also referred to as *structure-preserving* methods.

Many studies focus on the positivity-preserving property of a method. In [12], the authors consider graph-Laplacian ODEs and propose some second-order methods that unconditionally preserve positivity as well as a third-order method that preserves positivity under mild restrictions. These methods are based on Magnus integrators [13]. The authors of [12] propose a splitting strategy for the original system written in an extended space that applies to stiff or non-separable problems and uses the Strang splitting method. The overall method is second order and unconditionally conserves mass (akin to total population in the SIR model) and positivity. Patankar–Runge–Kutta methods (and their modified versions) have been developed to solve production-destruction systems (PDSs) monolithically while preserving the positivity and mass conservation [14, 15]. It turns out that such methods can be interpreted as approximations to the methods proposed in [12].

Although methods that preserve qualitative properties may involve splitting, e.g., [10, 16, 17], few studies systematically consider the effect of splitting strategies used in practice on qualitative property preservation. Given that splitting is so common due to its necessity or utility in practice, we systematically explore the effect of the choice of splitting strategy on the preservation of qualitative properties of the numerical solution of a differential equation. In this study, we limit the strategies considered to the process-based splitting as well as dynamic linearization.

The importance of the choice of splitting strategy on qualitative property preservation is shown indirectly in [12] in the context of insisting on writing the system in graph-Laplacian form. The effect of the choice of splitting strategy on the solution itself is shown more directly and dramatically in [18], where it is shown that two-dimensional rotations can be integrated exactly in time with the use of a splitting strategy based on shear rotations but not with the use of standard directional splitting.

The remainder of this paper is organized as follows. In Section 2, we give a description of operator-splitting methods including N_{op}-additively split methods for $N_{\text{op}} > 2$ and relevant background, definitions, and the qualitative properties of interest on the

SIR model and Robertson test problem. We describe two specific splitting strategies applied to the SIR model in Section 2.2, a process-based splitting based on the production and destruction terms and one based on dynamic linearization, which essentially performs a local linearization at every step of a numerical method. We further describe the generalization of process-based splitting to PDSs in Section 2.3. In Section 3, we give the main theoretical results regarding qualitative property preservation from the splitting strategies applied to the SIR model and the Robertson test problem. We find that not all splitting strategies are created equal when it comes to qualitative property preservation. How well an operator-splitting method preserves the desired qualitative properties depends on the splitting strategy (process-based or dynamical linearization), the operator-splitting method, and the form of the exact solution of the sub-systems. In Section 4, we offer some numerical experiments to support the theoretical results reported in the previous sections. Finally, in Section 5, we summarize our results and offer some conclusions.

2 Theoretical background

In this section, we describe the relevant theoretical background for the study of qualitative property preservation by operator splitting in the context of the production-destruction systems (PDSs). Accordingly, we introduce the necessary background on operator-splitting methods and the qualitative properties of interest. We examine two ways to split the SIR model (process-based and dynamic linearization) and a process-based splitting of the Robertson test problem in detail.

2.1 Operator-splitting methods

In this section, we introduce the operator-splitting (OS) methods as presented in [10]. We consider the initial value problem (IVP) for a 2-additive ordinary differential equation

$$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}t} = \mathcal{F}(t,\mathbf{y}) = \mathcal{F}^{[1]}(t,\mathbf{y}) + \mathcal{F}^{[2]}(t,\mathbf{y}), \qquad \mathbf{y}(0) = \mathbf{y}_0. \tag{2}$$

Let $\varphi^{[\ell]}_{\Delta t}$ be the exact flow of the sub-system

$$\frac{\mathrm{d}\mathbf{y}^{[\ell]}}{\mathrm{d}t} = \mathcal{F}^{[\ell]}(t,\mathbf{y}^{[\ell]})$$

for $\ell = 1, 2$. Compositions of $\varphi^{[\ell]}_{\Delta t}$ can be used to construct numerical solutions to (2). The most commonly known methods are the first-order Godunov (or Lie–Trotter) splitting method,

$$\Phi^{G}_{\Delta t} := \varphi^{[2]}_{\Delta t} \circ \varphi^{[1]}_{\Delta t},$$

and the second-order Strang splitting method,

$$\Phi^{S}_{\Delta t} := \varphi^{[1]}_{\Delta t/2} \circ \varphi^{[2]}_{\Delta t} \circ \varphi^{[1]}_{\Delta t/2}.$$

To construct a general s-stage operator-splitting method, we consider splitting coefficients $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_s]$, where $\boldsymbol{\alpha}_k = [\alpha_k^{[1]}, \alpha_k^{[2]}],\ k = 1, 2, \ldots, s$. An s-stage operator-splitting method that solves (2) can be written as

$$\Psi_{\Delta t} := \prod_{k=1}^{s} \Phi^{\{k\}}_{\boldsymbol{\alpha}_k \Delta t} = \Phi^{\{s\}}_{\boldsymbol{\alpha}_s \Delta t} \circ \Phi^{\{s-1\}}_{\boldsymbol{\alpha}_{s-1} \Delta t} \circ \cdots \circ \Phi^{\{1\}}_{\boldsymbol{\alpha}_1 \Delta t}, \tag{3}$$

where $\Phi^{\{k\}}_{\boldsymbol{\alpha}_k \Delta t} := \varphi^{[2]}_{\alpha_2^{[k]} \Delta t} \circ \varphi^{[1]}_{\alpha_1^{[k]} \Delta t}$. To achieve an order-$p_{OS}$ operator-splitting method, the coefficients $\boldsymbol{\alpha}$ must satisfy a system of polynomial equations derived from the Baker–Campbell–Hausdorff (BCH) formula [10]. For method up to order $p_{OS} = 3$, the order conditions are:

$$\begin{aligned}
p_{OS} = 1: &\qquad \sum_{k=1}^{s} \alpha_k^{[1]} = 1, \qquad \sum_{k=1}^{s} \alpha_k^{[2]} = 1,\\
p_{OS} = 2: &\quad \sum_{i=1}^{s} \alpha_i^{[2]} \left(\sum_{k=1}^{i} \alpha_k^{[1]} \right) = \tfrac{1}{2},\\
p_{OS} = 3: &\ \sum_{i=1}^{s-1} \alpha_i^{[2]} \left(\sum_{k=i+1}^{s} \alpha_k^{[1]} \right)^2 = \tfrac{1}{3}, \ \sum_{i=1}^{s} \alpha_i^{[1]} \left(\sum_{k=i}^{s} \alpha_k^{[2]} \right)^2 = \tfrac{1}{3}.
\end{aligned}$$

The application of OS methods is often limited to first- and second-order because methods of order three or higher require backward-in-time sub-steps for each operator during the integration [19]. In the case of the SIR model, backward-in-time integration tends to add challenges to preserving monotonicity of the numerical solution.

The family of two-stage, second-order operator-splitting methods admits a one-parameter set of solutions, of which the well-known Strang splitting method is a member. This family can be described using a free parameter $\beta \neq 1$. We denote such a method as OS22β, whose coefficients are given in Table 1. By varying the values of β, we can derive second-order OS methods with backward-in-time integration in only one or both of the operators. For $0 \leq \beta \leq 0.5$, both operators are integrated forward-in-time only. For $\beta > 0.5$, operator 1 requires backward integration at one sub-step. For $\beta > 1$ or $\beta < 0$, operator 2 requires backward integration at one sub-step. This makes OS22β a good template to examine the effect of backward integration on the properties (P1)–(P4).

We note that OS22$\beta(1 - \sqrt{2}/2)$ is the "best" two-stage, second-order OS method in the sense that it has the minimum local error measure for this class of methods [20]. For the purposes of comparison with higher-order methods, we also present numerical solutions from the third-order Ruth (R3) method, whose coefficients are given in

Table 1 Coefficients $\alpha_k^{[i]}$ for OS22β

k	$\alpha_k^{[1]}$	$\alpha_k^{[2]}$
1	$\frac{2\beta-1}{2(\beta-1)}$	$1-\beta$
2	$-\frac{1}{2(\beta-1)}$	β

Table 2 Coefficients $\alpha_k^{[i]}$ for the R3 method

k	$\alpha_k^{[1]}$	$\alpha_k^{[2]}$
1	$7/24$	$2/3$
2	$3/4$	$-2/3$
3	$-1/24$	1

Table 2, and the fourth-order Yoshida (Y4) method, whose coefficients are given in Table 3.

2.1.1 Operator-splitting for N_{op}-additive problems

Consider the IVP for an N_{op}-additive ODE

$$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}t} = \mathcal{F}(t,\mathbf{y}) = \sum_{\ell=1}^{N_{\mathrm{op}}} \mathcal{F}^{[\ell]}(t,\mathbf{y}), \qquad \mathbf{y}(0) = \mathbf{y}_0. \tag{4}$$

The IVP (4) can be solved using a generalized Godunov or Strang N_{op}-splitting method,

$$\Phi_{\Delta t}^{G-N_{\mathrm{op}}} := \varphi_{\Delta t}^{[N_{\mathrm{op}}]} \circ \cdots \varphi_{\Delta t}^{[2]} \circ \varphi_{\Delta t}^{[1]}, \tag{5}$$

$$\Phi_{\Delta t}^{S-N_{\mathrm{op}}} := \varphi_{\Delta t/2}^{[1]} \circ \varphi_{\Delta t/2}^{[2]} \circ \cdots \circ \varphi_{\Delta t}^{[N_{\mathrm{op}}]} \cdots \circ \varphi_{\Delta t/2}^{[2]} \circ \varphi_{\Delta t/2}^{[1]}. \tag{6}$$

Remark 1 *We note that the Strang splitting method is a composition of the Godunov splitting method with its adjoint over $\Delta t/2$. One of the approaches to generate high-order N_{op}-split operator-splitting methods is to use composition methods. We can compose basic low order N_{op}-split operator-splitting methods with different step sizes to generate high-order methods* [10]. *For example, let $\gamma_1 = \gamma_3 = \frac{1}{2-2^{1/3}}$, and $\gamma_2 = -\frac{2^{1/3}}{2-2^{1/3}}$. We can generate a Yoshida-like fourth-order N_{op}-split method by composing the Strang splitting method* (6)*:*

$$\Phi_{\Delta t}^{Y-N_{op}} := \Phi_{\gamma_3\Delta t}^{S-N_{op}} \circ \Phi_{\gamma_2\Delta t}^{S-N_{op}} \circ \Phi_{\gamma_1\Delta t}^{S-N_{op}}. \tag{7}$$

Table 3 Coefficients $\alpha_k^{[i]}$ for the Y4 method, where $\theta = \frac{1}{2-2^{1/3}}$

k	$\alpha_k^{[1]}$	$\alpha_k^{[2]}$
1	$\theta/2$	θ
2	$(1-\theta)/2$	$1-2\theta$
3	$(1-\theta)/2$	θ
4	$\theta/2$	0

Table 4 Coefficients $\alpha_k^{[i]}$ for the N_{op}-split Yoshida method, where $\theta = \frac{1}{2-2^{1/3}}$

k	$\alpha_k^{[1]}$	$\alpha_k^{[2]}$	$\cdots$	$\alpha_k^{[N_{\text{op}}-1]}$	$\alpha_k^{[N_{\text{op}}]}$
1	$\theta/2$	$\theta/2$	$\cdots$	$\theta/2$	θ
2	0	0	$\cdots$	$\theta/2$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$N_{\text{op}}-1$	0	$\theta/2$	$\cdots$	0	0
N_{op}	$(1-\theta)/2$	$(1-2\theta)/2$	$\cdots$	$(1-2\theta)/2$	$1-2\theta$
$N_{\text{op}}+1$	0	0	$\cdots$	$(1-2\theta)/2$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$2N_{\text{op}}-1$	$(1-\theta)/2$	$\theta/2$	$\cdots$	$\theta/2$	θ
$2N_{\text{op}}$	0	0	$\cdots$	$\theta/2$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$3N_{\text{op}}-3$	0	$\theta/2$	$\cdots$	0	0
$3N_{\text{op}}-2$	$\theta/2$	0	$\cdots$	0	0

The explicit coefficients of the N_{op}-split Yoshida method can be found in Table 4.

2.2 The SIR model

The susceptible-infected-recovered (SIR) model is a basic compartmental model first introduced by Kermack and McKendrick [3] in 1927. It is used for modeling of the spread of infectious diseases. Each living member of a general population is assigned to compartments susceptible (S), infectious (I), or recovered (R) according to whether they have never had the disease, have the disease, or no longer have the disease. The mathematical model can be described using the following differential equations

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} S(t) = -aS(t)I(t), \\ \dfrac{\mathrm{d}}{\mathrm{d}t} I(t) = aS(t)I(t) - bI(t), \\ \dfrac{\mathrm{d}}{\mathrm{d}t} R(t) = bI(t), \end{cases} \tag{8}$$

for all $t > 0$, $a, b > 0$ with the initial condition

$$S(0) = S_0, \;\; I(0) = I_0, \;\; R(0) = R_0.$$

Despite its simplicity, the SIR model can be used to demonstrate general trends in the evolution of its constituent compartments in response to new data (e.g., suggesting changes in parameter values) and potential interventions (e.g., mandatory masking, limits on gathering sizes, or lockdowns).

2.2.1 Process-based splitting of the SIR model with exact sub-integration

We first solve the SIR model with an operator-splitting strategy that splits the right-hand side of (8) according to the physical processes between the variables:

$$\begin{cases} \frac{\mathrm{d}S^{[1]}}{\mathrm{d}t} = 0, \\ \frac{\mathrm{d}I^{[1]}}{\mathrm{d}t} = -bI^{[1]}(t), \\ \frac{\mathrm{d}R^{[1]}}{\mathrm{d}t} = bI^{[1]}(t), \end{cases} \tag{9a}$$

$$\begin{cases} \frac{\mathrm{d}S^{[2]}}{\mathrm{d}t} = -aS^{[2]}(t)I^{[2]}(t), \\ \frac{\mathrm{d}I^{[2]}}{\mathrm{d}t} = aS^{[2]}(t)I^{[2]}(t), \\ \frac{\mathrm{d}R^{[2]}}{\mathrm{d}t} = 0. \end{cases} \tag{9b}$$

Remark 2 *We note that* (9a) *describes the transformation of population between I and R and* (9b) *describes the transformation of population between S and I. In this case, the process-based splitting coincides with a linear-nonlinear splitting of the original system of ODE. Linear-nonlinear splitting is a common splitting strategy for systems such as reaction-diffusion systems* [21].

At each OS stage k, sub-systems (9a) and (9b) are solved sequentially with time step-sizes $\alpha_k^{[1]}\Delta t$ and $\alpha_k^{[2]}\Delta t$. For such a splitting strategy, each sub-integration can be performed exactly. The exact solutions to (9a) and (9b) at OS stage k are

$$\begin{cases} S_{n,k}^{[1]} = S_{n,k-1}^{[2]}, \\ I_{n,k}^{[1]} = e^{-b\alpha_k^{[1]}\Delta t} I_{n,k-1}^{[2]}, \\ R_{n,k}^{[1]} = R_{n,k-1}^{[2]} + (1 - e^{-b\alpha_k^{[1]}\Delta t})I_{n,k-1}^{[2]}, \end{cases} \tag{10a}$$

$$\begin{cases} S_{n,k}^{[2]} = \frac{[S_{n,k}^{[1]}+I_{n,k}^{[1]}]S_{n,k}^{[1]}}{I_{n,k}^{[1]}\exp[a(S_{n,k}^{[1]}+I_{n,k}^{[1]})(\alpha_k^{[2]}\Delta t)]+S_{n,k}^{[1]}}, \\ I_{n,k}^{[2]} = \frac{[S_{n,k}^{[1]}+I_{n,k}^{[1]}]I_{n,k}^{[1]}}{S_{n,k}^{[1]}\exp[-a(S_{n,k}^{[1]}+I_{n,k}^{[1]})(\alpha_k^{[2]}\Delta t)]+I_{n,k}^{[1]}}, \\ R_{n,k}^{[2]} = R_{n,k}^{[1]}. \end{cases} \tag{10b}$$

The algorithm to advance the numerical solution of (9a) and (9b) using OS22β with exact sub-integration (10) from given S_n, I_n, R_n values at time t_n to values

S_{n+1}, I_{n+1}, R_{n+1} at time $t_{n+1} = t_n + \Delta t$ has the following form.

$$\begin{cases} S_{n,1}^{[1]} = S_n, \\ I_{n,1}^{[1]} = e^{-b\alpha_1^{[1]}\Delta t} I_n, \\ R_{n,1}^{[1]} = R_n + (1 - e^{-b\alpha_1^{[1]}\Delta t}) I_n, \end{cases} \tag{OS22β-process-based.1}$$

$$\begin{cases} S_{n,1}^{[2]} = \frac{[S_{n,1}^{[1]}+I_{n,1}^{[1]}]S_{n,1}^{[1]}}{I_{n,1}^{[1]}\exp[a(S_{n,1}^{[1]}+I_{n,1}^{[1]})(\alpha_1^{[2]}\Delta t)]+S_{n,1}^{[1]}}, \\ I_{n,1}^{[2]} = \frac{[S_{n,1}^{[1]}+I_{n,1}^{[1]}]I_{n,1}^{[1]}}{S_{n,1}^{[1]}\exp[-a(S_{n,1}^{[1]}+I_{n,1}^{[1]})(\alpha_1^{[2]}\Delta t)]+I_{n,1}^{[1]}}, \\ R_{n,1}^{[2]} = R_{n,1}^{[1]}, \end{cases} \tag{OS22β-process-based.2}$$

$$\begin{cases} S_{n,2}^{[1]} = S_{n,1}^{[2]}, \\ I_{n,2}^{[1]} = e^{-b\alpha_2^{[1]}\Delta t} I_{n,1}^{[2]}, \\ R_{n,2}^{[1]} = R_{n,1}^{[2]} + (1 - e^{-b\alpha_2^{[1]}\Delta t}) I_{n,1}^{[2]}, \end{cases} \tag{OS22β-process-based.3}$$

$$\begin{cases} S_{n+1} = S_{n,2}^{[2]} = \frac{[S_{n,2}^{[1]}+I_{n,2}^{[1]}]S_{n,2}^{[1]}}{I_{n,2}^{[1]}\exp[a(S_{n,2}^{[1]}+I_{n,2}^{[1]})(\alpha_2^{[2]}\Delta t)]+S_{n,2}^{[1]}}, \\ I_{n+1} = I_{n,2}^{[2]} = \frac{[S_{n,2}^{[1]}+I_{n,2}^{[1]}]I_{n,2}^{[1]}}{S_{n,2}^{[1]}\exp[-a(S_{n,2}^{[1]}+I_{n,2}^{[1]})(\alpha_2^{[2]}\Delta t)]+I_{n,2}^{[1]}}, \\ R_{n+1} = R_{n,2}^{[2]} = R_{n,2}^{[1]}. \end{cases} \tag{OS22β-process-based.4}$$

2.2.2 Dynamic linearization of the SIR model with exact sub-integration

To solve an ODE $\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}t} = \mathbf{f}(t, \mathbf{y})$ using dynamic linearization, we first write the ODE as

$$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}t} = \mathbf{f}(t, \mathbf{y}) = \mathbf{J}\mathbf{y} + (\mathbf{f}(t, \mathbf{y}) - \mathbf{J}\mathbf{y}),$$

where $\mathbf{J} = \partial\mathbf{f}/\partial\mathbf{y}$ is the Jacobian matrix. Then, the ODE is split as

$$\frac{\mathrm{d}\mathbf{y}^{[1]}}{\mathrm{d}t} = \mathbf{J}\mathbf{y}, \text{ and } \frac{\mathrm{d}\mathbf{y}^{[2]}}{\mathrm{d}t} = \mathbf{f}(t, \mathbf{y}) - \mathbf{J}\mathbf{y}.$$

Unless $\mathbf{f}(t, \mathbf{y})$ is linear and has constant coefficients, $\mathbf{J}$ is generally a function of t and the solution $\mathbf{y}(t)$. In the method of dynamic linearization, $\mathbf{J}$ is evaluated and then frozen at the beginning of each time step. Specifically, solving the SIR model using dynamic linearization from t_n to t_{n+1}, we evaluate the Jacobian matrix at $(t_n, \mathbf{y}_n)$ as

$$\mathbf{J}_n = \begin{bmatrix} -aI(t_n) & -aS(t_n) & 0 \\ aI(t_n) & aS(t_n) - b & 0 \\ 0 & b & 0 \end{bmatrix}.$$

For a splitting based on dynamic linearization of (8), the sub-systems

$$\begin{cases} \frac{\mathrm{d}S^{[1]}}{\mathrm{d}t} = -aI(t_n)S^{[1]}(t) - aS(t_n)I^{[1]}(t), \\ \frac{\mathrm{d}I^{[1]}}{\mathrm{d}t} = aI(t_n)S^{[1]}(t) + aS(t_n)I^{[1]}(t) - bI^{[1]}(t), \\ \frac{\mathrm{d}R^{[1]}}{\mathrm{d}t} = bI^{[1]}(t), \end{cases} \tag{11a}$$

$$\begin{cases} \frac{\mathrm{d}S^{[2]}}{\mathrm{d}t} = -aS^{[2]}(t)I^{[2]}(t) + aI(t_n)S^{[2]}(t) + aS(t_n)I^{[2]}(t), \\ \frac{\mathrm{d}I^{[2]}}{\mathrm{d}t} = aS^{[2]}(t)I^{[2]}(t) - aI(t_n)S^{[2]}(t) - aS(t_n)I^{[2]}(t), \\ \frac{\mathrm{d}R^{[2]}}{\mathrm{d}t} = 0, \end{cases} \tag{11b}$$

can again be integrated exactly and the solutions X_{n+1} for $X \in \{S, I, R\}$ are derived using a desired OS method. The exact solutions to (11a) and (11b) can be generated by using a computer algebra system such as Maple. Due to the complexity of these solutions, we do not present them here.

2.3 Robertson test problem and general PDSs

The Robertson test problem is a stiff system of three non-linear ODEs that describes the chemical reaction between three variables. It is given as follows

$$\begin{aligned} \frac{\mathrm{d}X}{\mathrm{d}t} &= aY(t)Z(t) - bX(t), \\ \frac{\mathrm{d}Y}{\mathrm{d}t} &= bX(t) - aY(t)Z(t) - cY(t)^2, \\ \frac{\mathrm{d}Z}{\mathrm{d}t} &= cY(t)^2, \end{aligned} \tag{12}$$

where a, b, c are positive constants and the initial conditions $X(0) = X_0$, $Y(0) = Y_0$, $Z(0) = Z_0$ are all positive.

2.3.1 Process-based splitting of the Robertson test problem with exact sub-integration

As proposed in (2.2.1), we split the Robertson problem according to processes into the following two sub-systems:

$$\begin{cases} \frac{\mathrm{d}X^{[1]}}{\mathrm{d}t} = aY^{[1]}(t)Z^{[1]}(t) - bX^{[1]}(t), \\ \frac{\mathrm{d}Y^{[1]}}{\mathrm{d}t} = bX^{[1]}(t) - aY^{[1]}(t)Z^{[1]}(t), \\ \frac{\mathrm{d}Z^{[1]}}{\mathrm{d}t} = 0, \end{cases} \tag{13a}$$

$$\begin{cases} \dfrac{\mathrm{d}X^{[2]}}{\mathrm{d}t} = 0, \\ \dfrac{\mathrm{d}Y^{[2]}}{\mathrm{d}t} = -c(Y^{[2]}(t))^2, \\ \dfrac{\mathrm{d}Z^{[2]}}{\mathrm{d}t} = c(Y^{[2]}(t))^2. \end{cases} \tag{13b}$$

Similar to the SIR problem, at each OS stage k, sub-systems (13a) and (13b) are solved exactly with time step-sizes $\alpha_k^{[1]}\Delta t$ and $\alpha_k^{[2]}\Delta t$. The exact solutions to (13a) and (13b) at OS stage k are

$$\begin{cases} X_{n,k}^{[1]} = \dfrac{-\exp(-(aZ_{n,k-1}^{[2]}+b)\alpha_k^{[1]}\Delta t)(aY_{n,k-1}^{[2]}Z_{n,k-1}^{[2]} - bX_{n,k-1}^{[2]}) + aZ_{n,k-1}^{[2]}(X_{n,k-1}^{[2]}+Y_{n,k-1}^{[2]})}{aZ_{n,k-1}^{[2]}+b}, \\ Y_{n,k}^{[1]} = \dfrac{\exp(-(aZ_{n,k-1}^{[2]}+b)\alpha_k^{[1]}\Delta t)(aY_{n,k-1}^{[2]}Z_{n,k-1}^{[2]} - bX_{n,k-1}^{[2]}) + b(X_{n,k-1}^{[2]}+Y_{n,k-1}^{[2]})}{aZ_{n,k-1}^{[2]}+b}, \\ Z_{n,k}^{[1]} = Z_{n,k-1}^{[2]}, \end{cases} \tag{14a}$$

$$\begin{cases} X_{n,k}^{[2]} = X_{n,k}^{[1]}, \\ Y_{n,k}^{[2]} = \dfrac{Y_{n,k}^{[1]}}{cY_{n,k}^{[1]}\alpha_k^{[2]}\Delta t + 1}, \\ Z_{n,k}^{[2]} = \dfrac{cY_{n,k}^{[1]}(Y_{n,k}^{[1]} + Z_{n,k}^{[1]})\alpha_k^{[2]}\Delta t + Z_{n,k}^{[1]}}{cY_{n,k}^{[1]}\alpha_k^{[2]}\Delta t + 1}, \end{cases} \tag{14b}$$

where $\{\alpha_k^\ell\}_{k=1,2,\ldots,s}^{\ell=1,2}$ and X_n, Y_n, Z_n are the numerical approximations of the variables X, Y, Z at $t = t_n$. We note that $X_{n,0}^{[2]} = X_n$, $Y_{n,0}^{[2]} = Y_n$, and $Z_{n,0}^{[2]} = Z_n$. To advance from t_n to t_{n+1}, apply (14a) and (14b) consecutively over all s stages of the operator-splitting method, and let $X_{n+1} = X_{n,s}^{[2]}$, $Y_{n+1} = Y_{n,s}^{[2]}$, $Z_{n+1} = Z_{n,s}^{[2]}$.

2.3.2 Process-based splitting of the production-destruction systems

A natural way to solve a fully conservative PDS (1) with N constituents using operator splitting is to split the system into $\frac{N(N-1)}{2}$ sub-systems,

$$\begin{cases} \frac{\mathrm{d}y_i}{\mathrm{d}t} = p_{ij}(\mathbf{y}) - d_{ij}(\mathbf{y}), \\ \frac{\mathrm{d}y_j}{\mathrm{d}t} = d_{ij}(\mathbf{y}) - p_{ij}(\mathbf{y}), \\ \frac{\mathrm{d}y_k}{\mathrm{d}t} = 0, \text{ for } k \neq i, j \end{cases} \tag{15}$$

for $i = 1, 2, \ldots, N-1$ and $j = i+1, i+2, \ldots, N$.

Remark 3 *1. We note that each sub-system describes the rate at which constituents i and j are transformed from one to the other. If all constituents i and j have two-way connections, a process-based splitting strategy will have $N(N-1)/2$ operators. For example, a general PDS with three constituents would be split into three sub-systems as depicted in* Fig. 1. *In the case of the SIR model, the transformations*

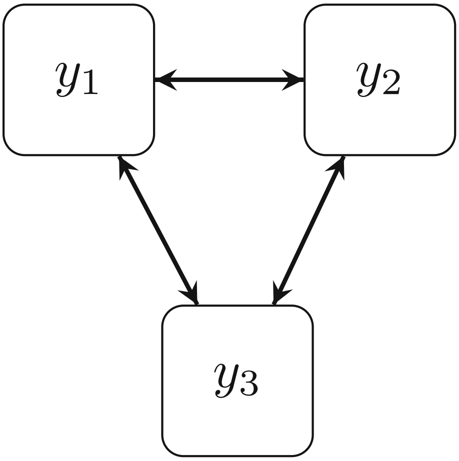

Fig. 1 Flowchart of a PDS with three constituents

are unidirectional and only between S and I and I and R as shown in Fig. 2*. Therefore, the resulting split system consists of only two sub-systems (each treated as one operator on two constituents at a time).*

2. *The choice on how to split a system of differential equations usually depends on the goals of the simulation, e.g., on the properties to be preserved or the physical or computational characteristics of the solution. Splitting the PDSs as described in* (15) *produces much simpler sub-systems and generally increases the chances of obtaining an exact solution for each sub-system (if desired). There are several available high-order* 2*- and* 3*-split operator-splitting methods available, e.g.,* [20, 22]*. N_{op}-split operator-splitting methods include Godunov* (5)*, Strang* (6)*, and Yoshida* (7)*. We are unaware of general N_{op}-split operator-splitting methods beyond these.*

2.4 Qualitative properties

2.4.1 Qualitative properties for the SIR model

Numerical solutions to the SIR model must share important properties with the true solution in order for them to have physical interpretations. We denote the numerical solutions $X_n \approx X(t_n)$, $X \in \{S, I, R\}$, for $t_n = n\Delta t$, $n \in \mathbb{N} := \{0, 1, \dots\}$.

1. The dynamics of an epidemic often dominate the dynamics of birth, death, and population immigration. Accordingly, it is justified to omit the effects of births, deaths, and immigration in a simple SIR model. Hence, the total population is

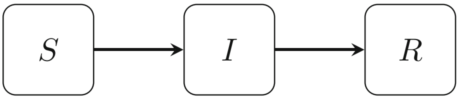

Fig. 2 Flowchart of the SIR model

conserved. The conservation of total population can be derived from the differential equations. By adding the equations of system (8), it is easy to obtain

$$\frac{\mathrm{d}S}{\mathrm{d}t}+\frac{\mathrm{d}I}{\mathrm{d}t}+\frac{\mathrm{d}R}{\mathrm{d}t}=0 \quad \Rightarrow \quad S(t)+I(t)+R(t)=N_0 \text{ for all } t>0.$$

We demand the same property from the numerical solutions; i.e., given initial conditions $S_0 + I_0 + R_0 = N_0$, we have

$$S_n + I_n + R_n = N_0 \text{ for all } n \in \mathbb{N} \tag{P1}$$

Failure to satisfy (P1) would undermine the credibility of any results. That being said, satisfaction of (P1) in itself does not guarantee reliable solutions; i.e., (P1) is a necessary but not sufficient indicator of solution quality.

2. Because the functions S, I, R denote population densities, their values should remain non-negative. Hence, we require the same from the numerical solution; i.e., given initial condition $X_0 \geq 0$, we have

$$X_n \geq 0, \text{ for all } n \in \mathbb{N} \text{ and } X \in \{S, I, R\}. \tag{P2}$$

3. We assume that infected or recovered individuals develop immunity; therefore, the function S is non-increasing in time. We require that the numerical solution satisfies

$$S_n \geq S_{n+1} \text{ for all } n \in \mathbb{N}. \tag{P3}$$

4. We assume that recovered individuals do not move to another compartment. Hence, R must be an non-decreasing function in time. We require the same from the numerical solution:

$$R_n \leq R_{n+1} \text{ for all } n \in \mathbb{N}. \tag{P4}$$

2.4.2 Qualitative properties for general PDSs

General production-destruction systems do not require monotonicity in its variables. We are interested in preserving the conservation and positivity properties in the numerical solution as defined in (4). In the context of the Robertson test problem introduced in Section 2.3.1, because the original ODE (12) is unconditionally conservative, therefore for all $\Delta t > 0$, for each $n \in \mathbb{N}$:

$$X_n + Y_n + Z_n = X_0 + Y_0 + Z_0. \tag{Robertson P1}$$

Because each variable X, Y, Z represents a chemical concentration, for all $\Delta t > 0$, for each $n \in \mathbb{N}$:

$$X_n, Y_n, Z_n \geq 0. \tag{Robertson P2}$$

3 Main results: effect of splitting strategy on qualitative property preservation of production-destruction systems

In this section, we give the main results on the qualitative property preservation of the two different splitting strategies considered applied to the SIR model and process-based splitting applied to the Robertson test problem. We also extend to the positivity-preserving property of OS methods to general production-destruction systems.

3.1 Conservation property of operator-splitting methods

In this section, we discuss the effect of operator-splitting methods in preserving the conservation property of a production-destruction system.

Theorem 1 *Assume that each sub-system* (15) *of a process-based splitting strategy of a PDS* (1) *has an exact solution. Then, the numerical approximation obtained using the operator-splitting methods and exact sub-integration is unconditionally conservative.*

Proof When solving (1) using operator-splitting methods, let $\mathbf{y}_{n,k}^{[\ell]}$ be the numerical solution of solving the sub-system (15) exactly over a fraction of $\alpha_k^{[\ell]} \Delta t$. Because the sum of the derivatives in (15) is equal to zero, the total sum of components of $\mathbf{y}_{n,k}^{[\ell]}$ does not change regardless of the value of $\alpha_k^{[\ell]}$. Hence, the sum of the components of numerical solution $\mathbf{y}_{n+1}$ at $t = t_{n+1}$ is equal to the sum of the components of numerical solution $\mathbf{y}_n$ at $t = t_n$. Hence, the operator-splitting methods are unconditionally conservative. □

Remark 4 *We note that the conservation property relies on two facts: 1. each sub-system is unconditionally conservative, i.e., the sum of the derivatives equals zero, and 2. the underlying numerical method to solve the sub-systems is also unconditionally conservative. In this case, the operator-splitting methods do not affect the conservation property. For the same reason, if the SIR model is split using dynamic linearization* (11a) *and* (11b) *and solved using operator-splitting methods with exact sub-integration, the numerical solution is still unconditionally conservative. In conclusion, property* (P1) *is satisfied for the SIR model with both process-based splitting and dynamical linearization splitting, and property* (Robertson P1) *is satisfied for the Robertson test problem.*

Remark 5 *One may contemplate eliminating one variable using the conservation property of the system for the SIR model and solving a system with one less unknown. However, it is unclear that such an approach would significantly simplify the solution of the SIR model or (even more so) other larger and more complex systems, e.g., production-destruction systems.*

3.2 Process-based splitting of the SIR model

In this section, we focus on the effect of negative time-stepping on the qualitative properties (P2)–(P4). If we do not need to employ any backward stepping, the properties (P2)–(P4) are satisfied trivially as stated in (2).

Theorem 2 *Assume that all* $\alpha_k^{[\ell]} \geq 0$. *Then, properties* (P2)–(P4) *are satisfied when we solve the SIR model using* (10).

Proof Property (P2) is satisfied because the exact solutions (10a) and (10b) are both positive for all intermediate stages of S, I, R if $\Delta t \geq 0$ and $\alpha_k^{[\ell]} \geq 0$.

Property (P3) is satisfied because $\frac{\mathrm{d}S^{[1]}}{\mathrm{d}t} = 0$ and $\frac{\mathrm{d}S^{[2]}}{\mathrm{d}t} \leq 0$ when all intermediate variables of S, I, R are all positive. Hence, when the subsystems (9a) and (9b) are solved exactly, the desired montonicity property (P3) is preserved.

Similarly, Property (P4) is satisfied because $\frac{\mathrm{d}R^{[1]}}{\mathrm{d}t} \geq 0$ and $\frac{\mathrm{d}R^{[2]}}{\mathrm{d}t} = 0$ when all intermediate variables of S, I, R are all positive. □

3.2.1 Solving the SIR model using OS22β with negative coefficients

Lemma 1 *If* $S_n \geq 0$ *and* $I_n \geq 0$, *then* $S_{n+1} \geq 0$ *and* $I_{n+1} \geq 0$ *for all* $\Delta t > 0$ *when the SIR model is solved using* (OS22β-process-based.1).

Proof Equation (10a) implies that at each stage k, after solving the first sub-system (9a) over $\alpha_k^{[1]}\Delta t$, $S_{n,k}^{[1]} \geq 0$ and $I_{n,k}^{[1]} \geq 0$ if $S_{n,k-1}^{[2]} \geq 0$ and $I_{n,k-1}^{[2]} \geq 0$.

Equation (10b) implies that at each stage k, after solving the second sub-system (9b) over $\alpha_k^{[2]}\Delta t$, $S_{n,k}^{[2]} \geq 0$ and $I_{n,k}^{[2]} \geq 0$ if $S_{n,k}^{[1]} \geq 0$ and $I_{n,k}^{[1]} \geq 0$ because exponential functions are positive.

Therefore, if $S_n \geq 0$ and $I_n \geq 0$, we can recursively conclude that $S_{n,k}^{[\ell]} \geq 0$ and $I_{n,k}^{[\ell]} \geq 0$ for $\ell = 1, 2$ and all $k = 1, 2, \ldots, s$. Hence, $S_{n+1} = S_{n,s}^{[2]} \geq 0$ and $I_{n+1} = I_{n,s}^{[2]} \geq 0$. □

Proposition 1 *Property* (P4) *holds for the SIR model for all* $\Delta t > 0$ *if* $S_n \geq 0$ *and* $I_n \geq 0$ *in* (OS22β-process-based.1).

Proof We consider the following two cases:

- $\beta \in (-\infty, 0.5]$.
  When $\beta \in (-\infty, 0.5]$, $\alpha_k^{[1]} \geq 0$ for $k = 1, 2$.
  Hence, $1 - e^{-b\alpha_k^{[1]}\Delta t} \geq 0$ for $k = 1, 2$. Furthermore, the proof of (1) implies that $I_n \geq 0$ and $I_{n,1}^{[2]} \geq 0$. Therefore,
  $$R_{n+1} - R_n = I_n(1 - e^{-b\alpha_1^{[1]}\Delta t}) + I_{n,1}^{[2]}(1 - e^{-b\alpha_2^{[1]}\Delta t}) \geq 0,$$
  and so $R_{n+1} \geq R_n$ for all n.
- $\beta \in (0.5, 1) \cup (1, \infty)$.
  When $\beta \in (0.5, 1) \cup (1, \infty)$, solving the SIR model with (OS22β-process-based) yields
  $$R_{n+1} - R_n = I_n(1 - e^{-b\alpha_1^{[1]}\Delta t}) + \frac{I_n(1 - e^{-b\alpha_2^{[1]}\Delta t})(S_n + I_n e^{-b\alpha_1^{[1]}\Delta t})e^{-b\alpha_1^{[1]}\Delta t}}{S_n e^{-a(S_n + I_n e^{-b\alpha_1^{[1]}\Delta t})\alpha_1^{[2]}\Delta t} + I_n e^{-b\alpha_1^{[1]}\Delta t}}.$$

We first show that

$$R_{n+1} - R_n \geq I_n(1 - e^{-b\alpha_1^{[1]}\Delta t}) + \frac{I_n(1 - e^{-b\alpha_2^{[1]}\Delta t})(S_n + I_n e^{-b\alpha_1^{[1]}\Delta t})e^{-b\alpha_1^{[1]}\Delta t}}{S_n + I_n e^{-b\alpha_1^{[1]}\Delta t}}. \tag{16}$$

When $\beta \in (0.5, 1), \alpha_1^{[1]} < 0, \alpha_1^{[2]} > 0$, and $\alpha_2^{[1]} > 0$, and the following inequalities hold:

$$\begin{cases} I_n(1 - e^{-b\alpha_1^{[1]}\Delta t}) < 0, & \text{because } \alpha_1^{[1]} < 0; \\ \frac{I_n(1-e^{-b\alpha_2^{[1]}\Delta t})(S_n+I_n e^{-b\alpha_1^{[1]}\Delta t})e^{-b\alpha_1^{[1]}\Delta t}}{S_n e^{-a(S_n+I_n e^{-b\alpha_1^{[1]}\Delta t})\alpha_1^{[2]}\Delta t}+I_n e^{-b\alpha_1^{[1]}\Delta t}} > 0, & \text{because } \alpha_2^{[1]} > 0; \\ e^{-a(S_n+I_n e^{-b\alpha_1^{[1]}\Delta t})\alpha_1^{[2]}\Delta t} \leq 1, & \text{because } \alpha_1^{[2]} > 0. \end{cases}$$

Hence, (16) holds.
When $\beta \in (1, \infty), \alpha_1^{[1]} > 0, \alpha_1^{[2]} < 0$, and $\alpha_2^{[1]} < 0$, and the following inequalities hold:

$$\begin{cases} I_n(1 - e^{-b\alpha_1^{[1]}\Delta t}) > 0, & \text{because } \alpha_1^{[1]} > 0; \\ \frac{I_n(1-e^{-b\alpha_2^{[1]}\Delta t})(S_n+I_n e^{-b\alpha_1^{[1]}\Delta t})e^{-b\alpha_1^{[1]}\Delta t}}{S_n e^{-a(S_n+I_n e^{-b\alpha_1^{[1]}\Delta t})\alpha_1^{[2]}\Delta t}+I_n e^{-b\alpha_1^{[1]}\Delta t}} < 0, & \text{because } \alpha_2^{[1]} < 0; \\ e^{-a(S_n+I_n e^{-b\alpha_1^{[1]}\Delta t})\alpha_1^{[2]}\Delta t} \geq 1, & \text{because } \alpha_1^{[2]} < 0. \end{cases}$$

Hence, (16) holds.
Finally,

$$\begin{aligned} R_{n+1} - R_n &\geq I_n(1 - e^{-b\alpha_1^{[1]}\Delta t}) + \frac{I_n(1 - e^{-b\alpha_2^{[1]}\Delta t})(S_n + I_n e^{-b\alpha_1^{[1]}\Delta t})e^{-b\alpha_1^{[1]}\Delta t}}{S_n + I_n e^{-b\alpha_1^{[1]}\Delta t}} \\ &= I_n(1 - e^{-b\alpha_1^{[1]}\Delta t}) + I_n(1 - e^{-b\alpha_2^{[1]}\Delta t})e^{-b\alpha_1^{[1]}\Delta t} \\ &= I_n - I_n e^{-b\alpha_1^{[1]}\Delta t} + I_n e^{-b\alpha_1^{[1]}\Delta t} - I_n e^{-b(\alpha_1^{[1]}+\alpha_2^{[1]})\Delta t} \\ &= I_n(1 - e^{-b(\alpha_1^{[1]}+\alpha_2^{[1]})\Delta t}). \end{aligned}$$

We note that $\alpha_1^{[1]} + \alpha_2^{[1]} = 1$ from the order conditions. Therefore, $R_{n+1} - R_n > 0$.

In conclusion, $R_{n+1} \geq R_n$ for all $\Delta t \geq 0$ when the SIR model is solved using (OS22β-process-based). □

Corollary 1 *Property* (P2) *holds for all* $\Delta t > 0$ *when the SIR model is solved using (OS22β-process-based).*

Proof Proposition 1 implies that, for all $\Delta t \geq 0$, $R_{n+1} \geq 0$ if $R_n \geq 0$. Equation (1) and the initial condition $S_0, I_0, R_0 \geq 0$ imply that $S_n, I_n, R_n \geq 0$ for all n. □

Remark 6 *We note that the intermediate stages $R^{[\ell]}_{n,k}$ may not all be positive, but all solution values at $t = t_{n+1}$ are non-negative.*

Furthermore, the proof of Proposition 1 can be generalized to cases where $\alpha^{[1]}_k + \alpha^{[1]}_{k+1} \geq 0$, which is a property both R3 and Y4 satisfy. In particular, it can easily be verified that in R3, we have $\alpha^{[1]}_2 + \alpha^{[1]}_3 > 0$, and in Y4, we have $\alpha^{[1]}_1 + \alpha^{[1]}_2 > 0$ and $\alpha^{[1]}_3 + \alpha^{[1]}_4 > 0$.

Proposition 2 *Property (P3) is a result of property (P2) for the SIR model for all $\beta \neq 1$ in (OS22β-process-based).*

Proof

- If $\beta \in [0, 1)$, then $\alpha^{[2]}_k \geq 0$ for $k = 1, 2$. Because all intermediate stages $S^{[\ell]}_{n,k}$ and $I^{[\ell]}_{n,k}$ are non-negative, $\frac{\mathrm{d}S^{[2]}}{\mathrm{d}t} = -aS^{[2]}I^{[2]} < 0$. Therefore,

$$S_{n+1} = S^{[2]}_{n,2} \leq S^{[1]}_{n,2} = S^{[2]}_{n,1} \leq S^{[1]}_{n,1} = S_n.$$

  Hence, property (P3) holds for all $\Delta t \geq 0$.
- If $\beta < 0$, then $\alpha^{[1]}_1, \alpha^{[1]}_2, \alpha^{[2]}_1 > 0$ and $\alpha^{[2]}_2 < 0$.
  To show that $S_{n+1} \geq S_n$, it is sufficient to show that $\frac{S_{n+1}}{S_n} \leq 1$.

$$\begin{aligned}\frac{S_{n+1}}{S_n} &= \frac{S^{[2]}_{n,2}}{S^{[1]}_{n,1}} = \frac{S^{[2]}_{n,2}}{S^{[1]}_{n,2}} \cdot \frac{S^{[1]}_{n,2}}{S^{[1]}_{n,1}} = \frac{S^{[2]}_{n,2}}{S^{[1]}_{n,2}} \cdot \frac{S^{[2]}_{n,1}}{S^{[1]}_{n,1}} \\ &= \frac{[S^{[1]}_{n,2} + I^{[1]}_{n,2}]}{I^{[1]}_{n,2}\exp[a(S^{[1]}_{n,2} + I^{[1]}_{n,2})(\alpha^{[2]}_2\Delta t)] + S^{[1]}_{n,2}} \cdot \frac{[S^{[1]}_{n,1} + I^{[1]}_{n,1}]}{I^{[1]}_{n,1}\exp[a(S^{[1]}_{n,1} + I^{[1]}_{n,1})(\alpha^{[2]}_1\Delta t)] + S^{[1]}_{n,1}}.\end{aligned}$$

  To show that $\frac{S_{n+1}}{S_n} \leq 1$, it is sufficient to show that

$$\begin{aligned}(I^{[1]}_{n,2}\exp[a(S^{[1]}_{n,2} + I^{[1]}_{n,2})(\alpha^{[2]}_2\Delta t)]+S^{[1]}_{n,2})(I^{[1]}_{n,1}\exp[a(S^{[1]}_{n,1}+I^{[1]}_{n,1})(\alpha^{[2]}_1\Delta t)]+S^{[1]}_{n,1}) \\ - (S^{[1]}_{n,2} + I^{[1]}_{n,2})(S^{[1]}_{n,1} + I^{[1]}_{n,1}) > 0. \qquad (18)\end{aligned}$$

  Expanding and simplifying the left-hand side of (18), we get

$$\begin{aligned}&\underbrace{I^{[1]}_{n,2}I^{[1]}_{n,1}\exp[a(S^{[1]}_{n,2} + I^{[1]}_{n,2})(\alpha^{[2]}_2\Delta t)]\exp[a(S^{[1]}_{n,1} + I^{[1]}_{n,1})(\alpha^{[2]}_1\Delta t)] - I^{[1]}_{n,2}I^{[1]}_{n,1}}_{\text{part 1}} \\ &\underbrace{\begin{aligned}&+I^{[1]}_{n,2}S^{[1]}_{n,1}\exp[a(S^{[1]}_{n,2} + I^{[1]}_{n,2})(\alpha^{[2]}_2\Delta t)] - I^{[1]}_{n,2}S^{[1]}_{n,1} \\ &\quad +I^{[1]}_{n,1}S^{[1]}_{n,2}\exp[a(S^{[1]}_{n,1} + I^{[1]}_{n,1})(\alpha^{[2]}_1\Delta t)] - I^{[1]}_{n,1}S^{[1]}_{n,2}\end{aligned}}_{\text{part 2}}. \qquad (19)\end{aligned}$$

  Before we show that both parts 1 and 2 are positive, we derive some useful equations and inequalities:

The order-1 condition is

$$\alpha_1^{[2]} + \alpha_2^{[2]} = 1. \tag{20}$$

Because $\alpha_1^{[1]} > 0$ and $\alpha_2^{[1]} > 0$, $R_{n,1}^{[1]} < R_{n,2}^{[1]}$. Therefore, (P1) implies that

$$S_{n,1}^{[1]} + I_{n,1}^{[1]} > S_{n,2}^{[1]} + I_{n,2}^{[1]} \geq 0. \tag{21}$$

Dividing the S and I terms in (10), we get

$$\frac{S_{n,2}^{[1]}}{I_{n,2}^{[1]}} = \frac{S_{n,1}^{[1]}}{I_{n,1}^{[1]}} \exp[\frac{b\Delta t}{2(1-\alpha_2^{[2]})} - a(S_{n,1}^{[1]} + I_{n,1}^{[1]})(1-\alpha_2^{[2]})\Delta t].$$

Hence,

$$S_{n,2}^{[1]} I_{n,1}^{[1]} = S_{n,1}^{[1]} I_{n,2}^{[1]} \exp[\frac{b\Delta t}{2(1-\alpha_2^{[2]})} - a(S_{n,1}^{[1]} + I_{n,1}^{[1]})(1-\alpha_2^{[2]})\Delta t]. \tag{22}$$

We now consider the exponential terms in part 1:

$$\begin{aligned}
&\exp[a(S_{n,2}^{[1]} + I_{n,2}^{[1]})(\alpha_2^{[2]}\Delta t)] \exp[a(S_{n,1}^{[1]} + I_{n,1}^{[1]})(\alpha_1^{[2]}\Delta t)] \\
=\ &\exp[a\Delta t(S_{n,2}^{[1]} + I_{n,2}^{[1]})\alpha_2^{[2]} + a\Delta t(S_{n,1}^{[1]} + I_{n,1}^{[1]})(1-\alpha_2^{[2]})] \\
=\ &\exp[a\Delta t\alpha_2^{[2]}(S_{n,2}^{[1]} + I_{n,2}^{[1]} - (S_{n,1}^{[1]} + I_{n,1}^{[1]})) + a\Delta t(S_{n,1}^{[1]} + I_{n,1}^{[1]})] > 1,
\end{aligned}$$

using (21) and $\alpha_2^{[2]} < 0$. Therefore, part 1 is positive.

We now consider part 2. First, we note that (21) and $\alpha_2^{[2]} < 0$ imply

$$\exp[a(S_{n,2}^{[1]} + I_{n,2}^{[1]})\alpha_2^{[2]}\Delta t] > \exp[a(S_{n,1}^{[1]} + I_{n,1}^{[1]})\alpha_2^{[2]}\Delta t] > 0. \tag{23}$$

Substituting (20), (22), and (23) into part 2, we get

$$\begin{aligned}
\text{part 2} &> I_{n,2}^{[1]} S_{n,1}^{[1]} \exp[a(S_{n,1}^{[1]} + I_{n,1}^{[1]})(\alpha_2^{[2]}\Delta t)] - I_{n,2}^{[1]} S_{n,1}^{[1]} \\
&\quad + I_{n,1}^{[1]} S_{n,2}^{[1]} \exp[a(S_{n,1}^{[1]} + I_{n,1}^{[1]})((1-\alpha_2^{[2]})\Delta t)] - I_{n,1}^{[1]} S_{n,2}^{[1]} \\
&= I_{n,2}^{[1]} S_{n,1}^{[1]} \left[\exp[a(S_{n,1}^{[1]} + I_{n,1}^{[1]})(\alpha_2^{[2]}\Delta t)] - 1\right] \\
&\quad + I_{n,2}^{[1]} S_{n,1}^{[1]} \exp\left[\frac{b\Delta t}{2(1-\alpha_2^{[2]})} - a(S_{n,1}^{[1]} + I_{n,1}^{[1]})(1-\alpha_2^{[2]})\Delta t\right]
\end{aligned}$$

$$\begin{aligned}
&\times \left[\exp[a(S^{[1]}_{n,1} + I^{[1]}_{n,1})((1 - \alpha^{[2]}_2)\Delta t)] - 1\right] \\
&= I^{[1]}_{n,2} S^{[1]}_{n,1} \left\{ \exp\left[\frac{b\Delta t}{2(1 - \alpha^{[2]}_2)}\right] - \exp\left[\frac{b\Delta t}{2(1 - \alpha^{[2]}_2)} - a(S^{[1]}_{n,1} + I^{[1]}_{n,1})(1 - \alpha^{[2]}_2)\Delta t\right] \right. \\
&\quad \left. + \exp[a(S^{[1]}_{n,1} + I^{[1]}_{n,1})(\alpha^{[2]}_2 \Delta t)] - 1 \right\} \\
&= I^{[1]}_{n,2} S^{[1]}_{n,1} \left\{ \exp\left[\frac{b\Delta t}{2(1 - \alpha^{[2]}_2)}\right] \left(1 - \exp[a(S^{[1]}_{n,1} + I^{[1]}_{n,1})(\alpha^{[2]}_2 - 1)\Delta t]\right) \right. \\
&\quad \left. - \left(1 - \exp[a(S^{[1]}_{n,1} + I^{[1]}_{n,1})(\alpha^{[2]}_2 \Delta t)]\right) \right\} \\
&> I^{[1]}_{n,2} S^{[1]}_{n,1} \left\{ \exp[\frac{b\Delta t}{2(1 - \alpha^{[2]}_2)}] \left(1 - \exp[a(S^{[1]}_{n,1} + I^{[1]}_{n,1})(\alpha^{[2]}_2 - 1)\Delta t]\right) \right. \\
&\quad \left. - \left(1 - \exp[a(S^{[1]}_{n,1} + I^{[1]}_{n,1})(\alpha^{[2]}_2 - 1)\Delta t]\right) \right\} \\
&= I^{[1]}_{n,2} S^{[1]}_{n,1} \left\{ \left(\exp[\frac{b\Delta t}{2(1 - \alpha^{[2]}_2)}] - 1\right) \left(1 - \exp[a(S^{[1]}_{n,1} + I^{[1]}_{n,1})(\alpha^{[2]}_2 - 1)\Delta t]\right) \right\} \\
&> 0,
\end{aligned}$$

because $\alpha^{[2]}_2 < 0$.
Now because parts 1 and 2 are both positive, $S_{n+1} \leq S_n$.

- If $\beta > 1$, then $\alpha^{[1]}_1, \alpha^{[2]}_2 > 0$ and $\alpha^{[2]}_1, \alpha^{[1]}_2 < 0$. Similar to the case where $\beta < 0$, to show property (P3), it is enough to show that $\frac{S_{n+1}}{S_n} \leq 1$, which is equivalent to showing that (18) holds.
  Equation (18) can be split into two parts as in (19), and again we show that both parts 1 and 2 are positive.
  We note that when $\beta > 1$, (20) and (22) still hold. Furthermore, because $\alpha^{[1]}_2 < 0$, $I^{[1]}_{n,2} > I^{[2]}_{n,1}$. Hence,

$$S^{[1]}_{n,2} + I^{[1]}_{n,2} = S^{[2]}_{n,1} + I^{[1]}_{n,2} > S^{[2]}_{n,1} + I^{[2]}_{n,1} = S^{[1]}_{n,1} + I^{[1]}_{n,1} \geq 0. \tag{24}$$

Now the exponential terms in part 1 can be written as

$$\exp[a\Delta t\alpha^{[2]}_2(S^{[1]}_{n,2} + I^{[1]}_{n,2} - (S^{[1]}_{n,1} + I^{[1]}_{n,1})) + a\Delta t(S^{[1]}_{n,1} + I^{[1]}_{n,1})] > 1,$$

using (24) and $\alpha^{[2]}_2 > 0$. Hence, part 1 is positive.
We now consider part 2. Using (20), (22), and (24), we get

$$\text{part 2} \;= I^{[1]}_{n,2} S^{[1]}_{n,1} \left\{ \exp[\frac{b\Delta t}{2(1 - \alpha^{[2]}_2)}] - \exp[\frac{b\Delta t}{2(1 - \alpha^{[2]}_2)} - a(S^{[1]}_{n,1} + I^{[1]}_{n,1})(1 - \alpha^{[2]}_2)\Delta t] \right.$$

$$
\begin{aligned}
&\quad + \exp[a(S_{n,2}^{[1]} + I_{n,2}^{[1]})(\alpha_2^{[2]}\Delta t)] - 1\Big\} \\
&> I_{n,2}^{[1]} S_{n,1}^{[1]} \Bigg\{ \exp[\frac{b\Delta t}{2(1-\alpha_2^{[2]})}] - \exp[\frac{b\Delta t}{2(1-\alpha_2^{[2]})} - a(S_{n,1}^{[1]} + I_{n,1}^{[1]})(1-\alpha_2^{[2]})\Delta t] \\
&\quad + \exp[a(S_{n,1}^{[1]} + I_{n,1}^{[1]})(\alpha_2^{[2]}\Delta t)] - 1\Big\} \\
&= I_{n,2}^{[1]} S_{n,1}^{[1]} \Bigg\{ -\exp[\frac{b\Delta t}{2(1-\alpha_2^{[2]})}] \Big(\exp[-a(S_{n,1}^{[1]} + I_{n,1}^{[1]})(1-\alpha_2^{[2]})\Delta t] - 1\Big) \\
&\quad + \Big(\exp[a(S_{n,1}^{[1]} + I_{n,1}^{[1]})(\alpha_2^{[2]}\Delta t)] - 1\Big)\Big\} \\
&> I_{n,2}^{[1]} S_{n,1}^{[1]} \Bigg\{ -\exp[\frac{b\Delta t}{2(1-\alpha_2^{[2]})}] \Big(\exp[-a(S_{n,1}^{[1]} + I_{n,1}^{[1]})(1-\alpha_2^{[2]})\Delta t] - 1\Big) \\
&\quad + \Big(\exp[-a(S_{n,1}^{[1]} + I_{n,1}^{[1]})(1-\alpha_2^{[2]})\Delta t] - 1\Big)\Big\} \\
&= I_{n,2}^{[1]} S_{n,1}^{[1]} \Big(\exp[-a(S_{n,1}^{[1]} + I_{n,1}^{[1]})(1-\alpha_2^{[2]})\Delta t] - 1\Big) \Bigg(1 - \exp[\frac{b\Delta t}{2(1-\alpha_2^{[2]})}]\Bigg) \\
&> 0,
\end{aligned}
$$

because $\alpha_2^{[2]} > 1$. Having shown parts 1 and 2 are both positive, we have the desired property (P3).

□

Remark 7 *We note that when the SIR model is solved using OS22β and the sub-systems* (9a) *and* (9b) *are solved exactly, all properties* (P1)–(P4) *are satisfied for all $\Delta t > 0$. As discussed in* [17, 23]*, if the nonlinear sub-system* (9b) *is solved using a Runge–Kutta method, properties* (P2)–(P4) *are only satisfied with a time-step restriction. This is due to the fact that the Runge–Kutta method does not preserve positivity or monotonicity of the sub-system.*

3.3 Dynamic linearization of the SIR model

In this section, we show that property (P2) does not hold for the SIR model when dynamic linearization is applied even when all $\alpha_k^{[\ell]} \geq 0$. Because (P3) and (P4) are usually consequences of (P2) with potentially more restricted step-sizes, we do not discuss step-size restriction on (P3) and (P4) in this section.

Due to the complexity of the exact solution of (11a) and (11b), we illustrate that property (P2) does not hold for the following set of parameters $\{a = 0.0005, b = 0.05, S_0 = 800, I_0 = 200, R_0 = 0\}$.

Proposition 3 *There exists a step-size $\Delta t^* > 0$ such that* (P2) *does not hold for $t \geq \Delta t^*$ when* (11) *is solved with an s-stage operator-splitting method with $\alpha_s^{[1]} \geq 0$.*

Proof Using the parameter values $\{a = 0.0005, b = 0.05, S_0 = 800, I_0 = 200, R_0 = 0\}$. It is enough to show that there is some $\Delta t^* > 0$ such that not all of $S_1, I_1, R_1 \geq 0$.

We show that when Δt is sufficiently large, R_1 is either negative or greater than N_0, implying that one of S_1 and I_1 must be negative.

Solving (11a) exactly over $\alpha_s^{[1]}\Delta t$, we get

$$\begin{aligned}
R_1 &= R_{0,s}^{[2]} = R_{0,s}^{[1]}\\
&= (-1.348901771 I_{0,s-1}^{[2]} - 1.10636 S_{0,s-1}^{[2]})e^{0.021922\alpha_s^{[1]}\Delta t}\\
&\quad +(0.348861832 I_{0,s-1}^{[2]} + 0.106336 S_{0,s-1}^{[2]})e^{0.22808\alpha_s^{[1]}\Delta t} + S_{0,s-1}^{[2]} + I_{0,s-1}^{[2]} + R_{0,s-1}^{[2]}\\
&= (-1.348901771 I_{0,s-1}^{[2]} - 1.10636 S_{0,s-1}^{[2]})e^{0.021922\alpha_s^{[1]}\Delta t}\\
&\quad +(0.348861832 I_{0,s-1}^{[2]} + 0.106336 S_{0,s-1}^{[2]})e^{0.22808\alpha_s^{[1]}\Delta t} + N_0,
\end{aligned}$$

where $N_0 = S_0 + I_0 + R_0$ is the conserved total population.

Regardless of the values of $S_{0,s-1}^{[2]}$ and $I_{0,s-1}^{[2]}$, $R_1 \to \infty$ or $R_1 \to -\infty$ as $\Delta t \to \infty$.

Hence, for sufficiently large Δt, $R_1 > 1000$ or $R_1 < 0$, implying that (P2) fails when Δt is sufficiently large. □

Remark 8 *We note that when $\alpha_k^{[\ell]} \geq 0$, the property (P2) does not hold because the exact sub-integration of R is no longer non-negative for all $\Delta t > 0$. This is a critical difference between the solution of the SIR model using process-based splitting and the solution using dynamic linearization.*

Finally, if $\alpha_s^{[1]} < 0$, it can be shown from the graph of R_1 that depending on the values of $S_{0,s-1}^{[2]}$ and $I_{0,s-1}^{[2]}$, $R_1 < 0$ for some choices of $\Delta t > 0$.

3.4 Positivity-preservation for the Robertson test problem

Proposition 4 *The numerical solution to the Robertson test problem (12) is unconditionally positive for all $\Delta t > 0$ when process-based splitting is used provided that all OS coefficients $\{\alpha_k^{[\ell]}\}_{k=1,2,\ldots,s}^{\ell=1,2}$ are non-negative.*

Proof Consider the exact solution (14a) to sub-system (13a). Assume that $X_{n,k-1}^{[2]}$, $Y_{n,k-1}^{[2]}$, $Z_{n,k-1}^{[2]} \geq 0$. Because $\alpha_k^{[\ell]} \geq 0$, $\exp(-(aZ_{n,k-1}^{[2]} + b)\alpha_k^{[1]}\Delta t) \leq 1$ for all $\Delta t > 0$. Now, we consider the following two cases:

- Case 1: If $aY_{n,k-1}^{[2]}Z_{n,k-1}^{[2]} - bX_{n,k-1}^{[2]} > 0$, then

$$X_{n,k}^{[1]} > \frac{-(aY_{n,k-1}^{[2]}Z_{n,k-1}^{[2]} - bX_{n,k-1}^{[2]}) + aZ_{n,k-1}^{[2]}(X_{n,k-1}^{[2]} + Y_{n,k-1}^{[2]})}{aZ_{n,k-1}^{[2]} + b} = X_{n,k-1}^{[2]} \geq 0,$$

$$Y_{n,k}^{[1]} = \frac{\exp(-(aZ_{n,k-1}^{[2]} + b)\alpha_k^{[1]}\Delta t)(aY_{n,k-1}^{[2]}Z_{n,k-1}^{[2]} - bX_{n,k-1}^{[2]}) + b(X_{n,k-1}^{[2]} + Y_{n,k-1}^{[2]})}{aZ_{n,k-1}^{[2]} + b} \geq 0,$$

again because both the numerator and denominator are positive.

- Case 2: If $aY^{[2]}_{n,k-1}Z^{[2]}_{n,k-1} - bX^{[2]}_{n,k-1} < 0$, then
$$X^{[1]}_{n,k} \geq 0,$$
because both the numerator and denominator are positive, and
$$Y^{[1]}_{n,k} > \frac{(aY^{[2]}_{n,k-1}Z^{[2]}_{n,k-1} - bX^{[2]}_{n,k-1}) + b(X^{[2]}_{n,k-1} + Y^{[2]}_{n,k-1})}{aZ^{[2]}_{n,k-1} + b} = Y^{[2]}_{n,k-1} \geq 0,$$
because both the numerator and denominator are positive.

It is obvious that $Z^{[1]}_{n,k} = Z^{[2]}_{n,k-1} \geq 0$ in both cases. The exact solution (14b) of sub-system (13b) is obviously non-negative if $X^{[1]}_{n,k}, Y^{[1]}_{n,k}, Z^{[1]}_{n,k} \geq 0$ and $\alpha^{[2]}_k \geq 0$. Therefore, $X_{n+1}, Y_{n+1}, Z_{n+1} \geq 0$ if $X_n, Y_n, Z_n \geq 0$. Because the initial conditions $X_0, Y_0, Z_0 \geq 0$, we have $X_n, Y_n, Z_n \geq 0$ for all $n = 1, 2, \ldots$. □

We now focus our attention on OS22β and study the effect of backward integration on the Robertson test problem. We use the OS22β method for analysis because it admits backward integration in either one or both of the sub-systems. The generalization of the results to other operator-splitting methods can be done using similar argument.

Proposition 5 *We solve the Robertson problem* (12) *using OS22β with* (13a) *and* (13b) *integrated exactly. If $\alpha^{[\ell]}_k < 0$ for some k and ℓ, then one of X_1, Y_1, Z_1 is negative for Δt sufficiently large.*

Proof We note that for $\beta \in [0, 0.5]$, all coefficients of OS22β are non-negative. Therefore, we only need to discuss the following three cases: $\beta \in (0.5, 1)$, $\beta \in (1, \infty)$, and $\beta \in (-\infty, 0)$.

- $\beta \in (0.5, 1)$
For $\beta \in (0.5, 1)$, $\alpha^{[1]}_1 < 0$ and $\alpha^{[2]}_1, \alpha^{[1]}_2, \alpha^{[2]}_2 \geq 0$, we show that if $aY_0Z_0 - bX_0 < 0$, then $X_1 < 0$ for Δt sufficiently large.
  - If $aY_0Z_0 - bX_0 < 0$, after integrating the first operator (13a) over $\alpha^{[1]}_1\Delta t$,
$$\begin{cases} X^{[1]}_{0,1} = \dfrac{-\exp(-(aZ_0 + b)\alpha^{[1]}_1\Delta t)(aY_0Z_0 - bX_0) + aZ_0(X_0 + Y_0)}{aZ_0 + b}, \\ Y^{[1]}_{0,1} = \dfrac{\exp(-(aZ_0 + b)\alpha^{[1]}_1\Delta t)(aY_0Z_0 - bX_0) + b(X_0 + Y_0)}{aZ_0 + b}, \\ Z^{[1]}_{0,1} = Z_0. \end{cases} \tag{25}$$
We note that $Y^{[1]}_{0,1} \to -\infty$ as $\Delta t \to \infty$ because all initial conditions X_0, Y_0, $Z_0 \geq 0$ and $\alpha^{[1]}_1 < 0$. Hence, we can choose Δt large enough such that

$Y_{0,1}^{[1]} < 0$ and $a(Y_{0,1}^{[1]} + Z_{0,1}^{[1]}) + b < 0$. After integrating the second operator (13b) over $\alpha_1^{[2]}\Delta t$,

$$\begin{cases} X_{0,1}^{[2]} = X_{0,1}^{[1]}, \\ Y_{0,1}^{[2]} = \dfrac{Y_{0,1}^{[1]}}{cY_{0,1}^{[1]}\alpha_1^{[2]}\Delta t + 1}, \\ Z_{0,1}^{[2]} = \dfrac{cY_{0,1}^{[1]}(Y_{0,1}^{[1]}+Z_{0,1}^{[1]})\alpha_1^{[2]}\Delta t+Z_{0,1}^{[1]}}{cY_{0,1}^{[1]}\alpha_1^{[2]}\Delta t+1} = -\dfrac{Y_{0,1}^{[1]}}{cY_{0,1}^{[1]}\alpha_1^{[2]}\Delta t+1}+Y_{0,1}^{[1]}+Z_{0,1}^{[1]}. \end{cases} \tag{26}$$

We note that $X_{0,1}^{[2]} > 0$, and, for Δt sufficiently large, $Y_{0,1}^{[2]} > 0$ and $Z_{0,1}^{[2]} < 0$. Moreover, as $\Delta t \to \infty$, $Y_{0,1}^{[2]} \to 0$ and $aZ_{0,1}^{[2]} + b = a(-\frac{Y_{0,1}^{[1]}}{cY_{0,1}^{[1]}\alpha_1^{[2]}\Delta t+1}) + a(Y_{0,1}^{[1]} + Z_{0,1}^{[1]}) + b < 0$. After integrating the first operator (13a) over $\alpha_2^{[1]}\Delta t$,

$$X_{0,2}^{[1]} = \frac{-\exp(-(aZ_{0,1}^{[2]}+b)\alpha_2^{[1]}\Delta t)(aY_{0,1}^{[2]}Z_{0,1}^{[2]}-bX_{0,1}^{[2]})+aZ_{0,1}^{[2]}(X_{0,1}^{[2]}+Y_{0,1}^{[2]})}{aZ_{0,1}^{[2]}+b}. \tag{27}$$

We note that $(aY_{0,1}^{[2]}Z_{0,1}^{[2]} - bX_{0,1}^{[2]}) < 0$ and $\exp(-(aZ_{0,1}^{[2]} + b)\alpha_2^{[1]}\Delta t) \to \infty$ as $\Delta t \to \infty$. Hence, the numerator of $X_{0,2}^{[1]}$ is positive for Δt sufficiently large. Because the denominator of $X_{0,2}^{[1]}$ is negative, $X_{0,2}^{[1]} < 0$. Therefore, $X_1 = X_{0,2}^{[2]} = X_{0,2}^{[1]} < 0$.

- If $aY_0Z_0 - bX_0 > 0$, because $\alpha_1^{[1]} < 0$, $\exp(-(aZ_0 + b)\alpha_1^{[1]}\Delta t) \to \infty$ as $\Delta t \to \infty$. Referring to the expression of $X_{0,1}^{[1]}$ in (25), it is obvious that $X_{0,1}^{[1]} < 0$ and $Y_{0,1}^{[1]}, Z_{0,1}^{[1]} > 0$ for Δt sufficiently large. Referring to the expressions (26) of integrating (13b) over $\alpha_1^{[2]}\Delta t$, in this case $X_{0,1}^{[2]} < 0$ and $Y_{0,1}^{[2]}, Z_{0,1}^{[2]} > 0$. Moreover $Y_{0,1}^{[2]} \to 0$ as $\Delta t \to \infty$. Therefore, we can choose Δt sufficiently large such that $X_{0,1}^{[2]} + Y_{0,1}^{[2]} < 0$. Now refer to (27) for the solution of $X_{0,2}^{[1]}$ after integrating (13a) over $\alpha_2^{[1]}\Delta t$. Because $Z_{0,1}^{[2]} > 0$ and $\alpha_2^{[1]} > 0$, $\exp(-(aZ_{0,1}^{[2]} + b)\alpha_2^{[1]}\Delta t) \to 0$ as $\Delta t \to \infty$. Because $X_{0,1}^{[2]} + Y_{0,1}^{[2]} < 0$, for Δt sufficiently large, $X_{0,2}^{[1]} < 0$. Therefore, $X_1 = X_{0,2}^{[2]} = X_{0,2}^{[1]} < 0$.

- $\beta \in (1, \infty)$
  For $\beta \in (1, \infty)$, $\alpha_1^{[1]}, \alpha_2^{[2]} \geq 0$ and $\alpha_1^{[2]}, \alpha_2^{[1]} < 0$. After integrating the first operator (13a) over $\alpha_1^{[1]}\Delta t$, the resulting intermediate values $X_{0,1}^{[1]}$, $Y_{0,1}^{[1]}$, and $Z_{0,1}^{[1]}$ are non-negative for any $\Delta t > 0$. After integrating the second operator (13b) over a negative time-step $\alpha_2^{[1]}\Delta t$, $X_{0,1}^{[2]}$ and $Z_{0,1}^{[2]}$ are non-negative for all $\Delta t > 0$, and $Y_{0,1}^{[2]} < 0$ for Δt sufficiently large. Therefore, $aY_{0,1}^{[2]}Z_{0,1}^{[2]} - bX_{0,1}^{[2]} < 0$. Because $\alpha_2^{[1]} < 0$, $\exp(-(aZ_{0,1}^{[2]} + b)\alpha_2^{[1]}\Delta t) \to \infty$ as $\Delta t \to \infty$. Therefore, $X_{0,2}^{[1]} \to \infty$, $Y_{0,2}^{[1]} \to -\infty$, and $Z_{0,2}^{[1]} > 0$ when $\Delta t \to \infty$. Moreover, because $X_{0,2}^{[1]}+Y_{0,2}^{[1]}+Z_{0,2}^{[1]}$

remains constant, for Δt sufficiently large, $Y_{0,2}^{[1]} + Z_{0,2}^{[1]} < 0$. Now consider $Z_1 = Z_{0,2}^{[2]}$. Because $\alpha_2^{[1]} > 0$, $Y_{0,2}^{[1]} < 0$, $Z_{0,2}^{[1]} > 0$, $Y_{0,2}^{[1]} + Z_{0,2}^{[1]} < 0$, the numerator of Z_1 is positive and the denominator of Z_1 is negative for Δt sufficiently large. Hence, $Z_1 < 0$.

- $\beta \in (-\infty, 0)$:
  For $\beta \in (-\infty, 0)$, $\alpha_1^{[1]}, \alpha_1^{[2]}, \alpha_2^{[1]} \geq 0$ and $\alpha_2^{[2]} < 0$. As shown in the proof of (4), because the initial conditions $X_0, Y_0, Z_0 \geq 0$, the intermediate values $X_{0,2}^{[1]}$, $Y_{0,2}^{[1]}$, and $Z_{0,2}^{[1]}$ after integrating the first operator (13a) over $\alpha_2^{[1]}\Delta t$ are all non-negative for any $\Delta t > 0$. Because $\alpha_2^{[2]} < 0$, the exact solution of the second operator (14b) indicate that $Y_1 = Y_{0,2}^{[2]} < 0$ for Δt sufficiently large.

□

Remark 9 *We note that although the Robertson test problem has exact solutions* (14a) *and* (14b) *for each of the sub-systems, the exact solutions are not always positive when $\alpha_k^{[\ell]} < 0$. In fact, the exact solutions blow up when $\alpha_k^{[\ell]} < 0$ and $\Delta t \to \infty$. This instability in the exact solution is a main difference between the Robertson test problem and the SIR model. Moreover, although we only care about the positivity of the variables at the end of each time-step t_n, it is beneficial to keep the intermediate variables $X_{n,k}^{[\ell]}$, $Y_{n,k}^{[\ell]}$, $Z_{n,k}^{[\ell]}$ positive because this would reduce the chance of blow up in the next sub-step.*

Remark 10 *Assume that each sub-system* (15) *of a process-based splitting strategy of a PDS* (1) *has a positive exact solution. Then, the numerical results obtained using the Godunov or Strang splitting method with $N(N-1)/2$ operators and exact sub-integration is unconditionally positive because the exact solution of each sub-system* (15) *is unconditionally positive. However, if we use a generalized Yoshida method (7), the positivity is not guaranteed because the exact solution can be negative when integrated backward in time, and this negativity might not be compensated by the subsequent forward integration.*

4 Numerical experiments

In this section, we give the results of some numerical experiments to support the theoretical results reported in the previous sections.

We also performed experiments with the SIR model using the modified Patankar–Runge–Kutta method MPRK22 from [24] as well as the splitting method ES2 from [12] for differential equations that are not split additively. True to the theory, both methods produced results that satisfy (P1)–(P2). However, the resulting accuracies appeared to be significantly worse than solving (9) using Strang splitting ((OS22β-process-based) with $\beta = 1/2$) for a given time step-size.

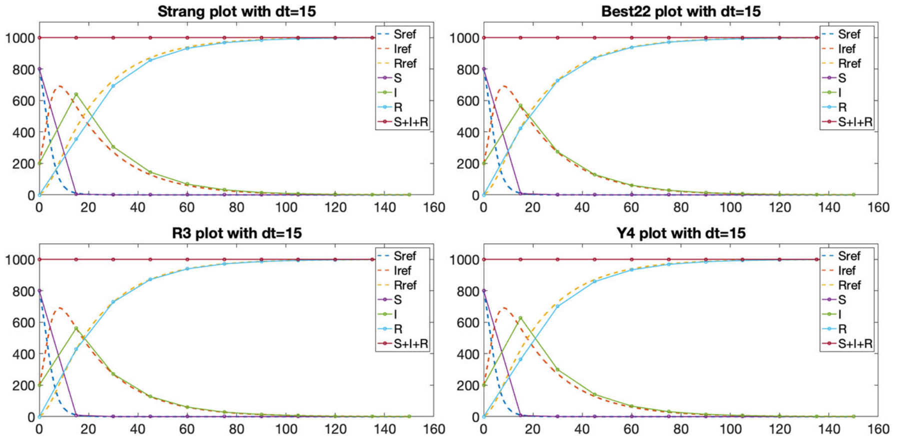

Fig. 3 Plots of the numerical solution for various operator-splitting methods applied to a process-based splitting. Properties (P1)–(P4) all hold

4.1 Process-based splitting

The SIR model is solved using (OS22β-process-based). The splitting methods examined are Strang, OS22$\beta(1-\sqrt{2}/2)$ ("Best22"), R3, Y4, OS22$\beta(-0.25)$, OS22$\beta(0.75)$, and OS22$\beta(1.5)$. Figures 3 and 4 display the numerical results for $\Delta t = 15$. In agreement with the theory, we see that properties (P1)–(P4) all hold. Of note, we see that the presence of positive coefficients in the OS methods is neither necessary nor sufficient for qualitative property preservation. We further note that the numerical result for larger Δt is similar the case when $\Delta t = 15$, as proved in Section 3.2.

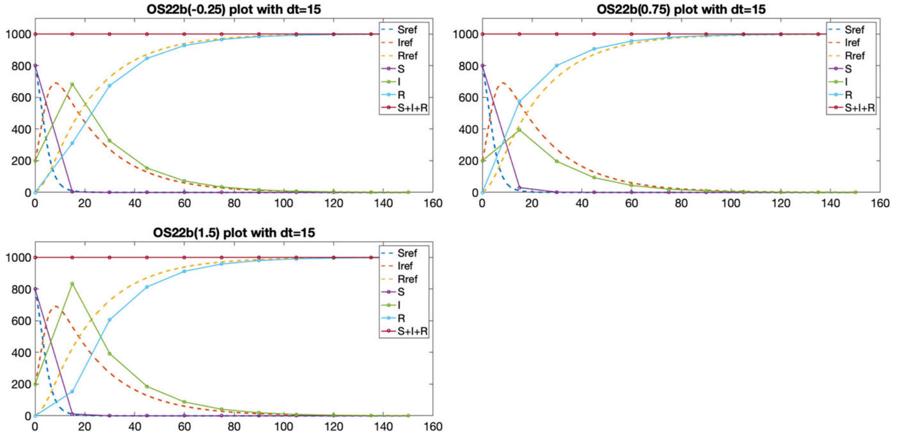

Fig. 4 Plots of the numerical solution for various OS22β operator-splitting methods applied to a process-based splitting. Properties (P1)–(P4) all hold

4.2 Dynamic linearization

The SIR model is now solved using dynamic linearization (11). The splitting methods examined again are Strang, OS22$\beta(1-\sqrt{2}/2)$, R3, Y4, and OS22$\beta(-0.25)$. Table 5 summarizes the smallest step-size such that each property fails for each method. In agreement with the theoretical results, property (P1) holds for all step-sizes, and there is a step-size Δt beyond which each qualitative property (P2)–(P4) does not hold.

We note that both Strang and OS22$\beta(1-\sqrt{2}/2)$ only have positive coefficients. Nonetheless, properties (P2)–(P4) fail for Δt sufficiently large. That is, the mere absence of negative coefficients is not sufficient to guarantee the success of a splitting method depending on the goals.

That being said, all of R3 and Y4 have negative coefficients in both operators, and OS22$\beta(-0.25)$ has a negative coefficient in only the second operator. Again the properties (P2)–(P4) fail for sufficiently large Δt, in agreement with the theory. It seems, however, for this model, the presence of negative coefficients may lead to qualitative property preservation breaking down sooner, i.e., for smaller Δt, than for the case where negative coefficients are absent.

OS22$\beta(1-\sqrt{2}/2)$ is the method with the smallest splitting error among all OS22β methods. We see from Table 5 that it can take a step-size that is almost 50% larger than Strang before any of the properties (P2)–(P4) cease to hold.

Finally, we note from the results of Strang, OS22$\beta(1-\sqrt{2}/2)$, and OS22$\beta(-0.25)$ that properties (P3) and (P4) are not a consequence of (P2). Any of the three properties may cease to fail first.

4.3 Robertson test problem

In this section, we solve the Robertson test problem (12) using process-based splitting. The operator-splitting methods examined here are OS22β(−0.2), OS22β(0.7), and OS22β(1.2). The sub-systems (13a) and (13b) are solved exactly. For the numerical experiments, we use the parameter values $a = 1\text{e}4$, $b = 0.04$, and $c = 3\text{e}7$ with initial conditions $X(0) = X_0 = 1 - 2\texttt{eps}$, $Y(0) = Y_0 = \texttt{eps}$, $Z(0) = Z_0 = \texttt{eps}$. In Table 6, we present the stepsize Δt when one of the three variables X_1, Y_1, Z_1 false to be positive. As proved in Section 3.4, when one of the $\alpha_k^{[\ell]}$ is negative, (Robertson P2) false for Δt sufficiently large.

Remark 11 *We note that Table 6 indicates that when Δt is sufficiently large, at least one of X_1, Y_1, or Z_1 is negative. However, the step sizes that preserve positivity of X_1, Y_1, Z_1 do not guarantee positivity of X_n, Y_n, Z_n for all $n > 1$. In the case of OS22β(0.7), both step sizes $\Delta t = 0.12$ and $\Delta t = 0.13$ fail to produce a positive solution to the Robertson problem over the full interval* $[0, 1e + 10]$.

On the other hand, when process-based splitting is employed and all operator splitting coefficients are positive, positivity of the variables X_n, Y_n, Z_n for $n \geq 1$ is satisfied unconditionally, as claimed in Proposition 4. Figure 5 presents the numerical solution of the Robertson problem solved using process-based splitting with OS22$\beta(1-\sqrt{2}/2)$ for $\Delta t = 10$. Positivity of all three variables and the conservation of the sum of X, Y, Z are satisfied as expected. We note that the numerical results for larger Δt are similar to

Table 5 The smallest step-size such that each property fails for each method, when solving the SIR model using dynamic linearization (11)

Method	Strang			OS22$\beta(1-\sqrt{2}/2)$			R3			Y4		OS22$\beta(-0.25)$		
Stepsize Δt	10.64	10.65	10.69	14.98	14.99	15.10	3.96	3.97	4.10	5.59	5.60	8.69	8.70	8.80
Property 1	T	T	T	T	T	T	T	T	T	T	T	T	T	T
Property 2	T	T	F	T	T	F	T	F	F	T	F	T	T	F
Property 3	T	F	F	T	F	F	T	T	F	T	F	T	F	F
Property 4	T	T	F	T	T	F	T	F	F	T	T	T	T	F

Table 6 In each case, for Δt is sufficiently large, one of $X_1, Y_1, Z_1 < 0$ when the Robertson test problem is solved using process-based splitting

Method	OS22$\beta(-0.2)$		OS22$\beta(0.7)$		OS22$\beta(1.2)$	
Stepsize Δt	0.0015	0.002	0.12	0.13	1e-7	2e-7
$X_1 \geq 0$	T	T	T	F	T	T
$Y_1 \geq 0$	T	T	T	T	T	T
$Z_1 \geq 0$	T	F	T	T	T	F

the case when $\Delta t = 10$. Although operator-splitting methods with positive coefficients are unconditionally positive for the Robertson test problem, they are less accurate as the MPRK22 method for a given step size for this problem.

Remark 12 *Furthermore, the differential equations of the Robertson problem imply that Z should be monotonically increasing. When using process-based splitting and an operator-splitting method with positivie coefficients, the monotonicity of Z is satisfied unconditionally for the same reason that R, in the SIR model, is monotonically increasing as presented in Theorem 2. When using operator splitting method with negative coefficients, this property fails for Δt sufficiently large as shown in Fig. 6.*

5 Summary and conclusions

Mathematical modelling is omni-present in modern daily life. These models are typically large, complex, and require solutions to be approximated by numerical methods.

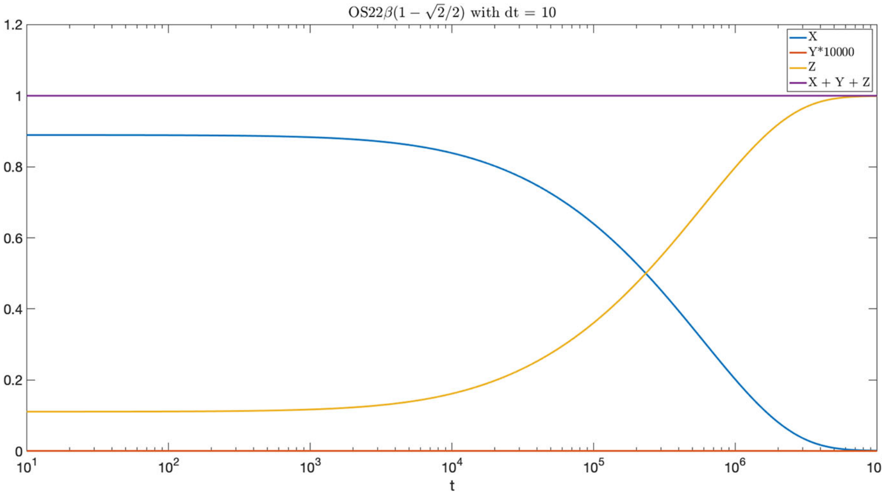

Fig. 5 Positivity of all three variables and the conservation of the sum of X, Y, Z are satisfied when the Robertson problem is solved using process-based splitting with operator splitting method OS22$\beta(1-\sqrt{2}/2)$

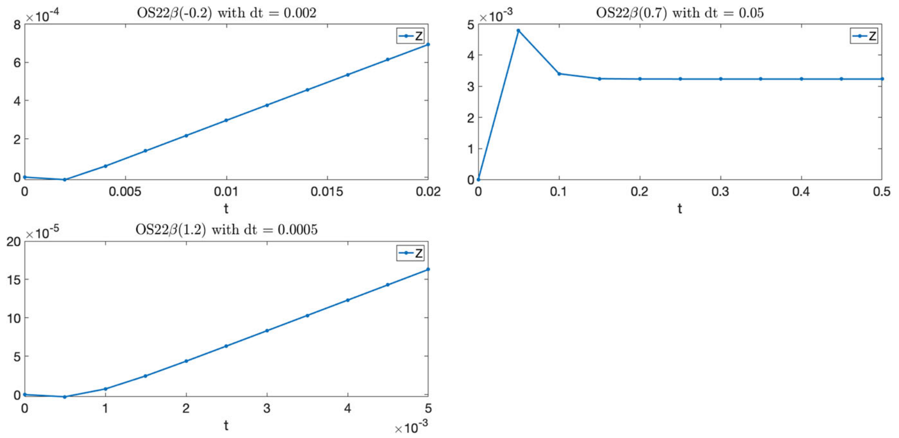

Fig. 6 When solving the Robertson problem with operator-splitting methods with negative coefficients, Z fails to be monotonically increasing for Δt sufficiently large

Often, the problems are posed as differential equations that are so large that they must be split into pieces that are solved separately. Furthermore, the numerical solutions may be required to satisfy certain qualitative properties in order to be physically meaningful.

The SIR model is a basic model of infectious disease spread that can be used to illustrate how qualitative properties, such as positivity, monotonicity, or conservation of total population, are affected by the choice of splitting strategy, i.e., despite the fact that the sub-systems are integrated exactly. Accordingly, an analysis such as this can inform which splitting strategies are most amenable to qualitative property preservation.

We have demonstrated that a process-based splitting, which for the SIR model also happens to correspond to a splitting based on linear/nonlinear terms, unconditionally preserves positivity, monotonicity, and total population. This result has some applicability to understanding qualitative property preservation of the more general class of production-destruction systems. For PDSs, total population and positivity are unconditionally preserved under process-based splitting.

On the other hand, the popular and powerful dynamic linearization method is only conditionally stable; i.e., there is a step-size beyond which at least one of the qualitative properties (P1)–(P4) cease to hold. In practice, these step-sizes may be so large as to yield inaccurate solutions, in which case smaller step-sizes would be required anyway, and the conditional nature of qualitative property preservation may largely be irrelevant. As usual, the impact of the presence or absence of restrictions due to stability depends on the goals of the simulation.

Comparing the two splitting strategies applied to the SIR model, we conclude that the process-based splitting is preferred over dynamic linearization because the exact solutions of sub-systems of the process-based splitting are unconditionally positive and conservative for all $\Delta t > 0$. By comparing the results of the SIR model and the Robertson test problem, we conclude that when choosing a particular operator-splitting method, if the exact solution to the sub-systems preserves the desired qualitative properties for $\Delta t < 0$, then it is safe to use operator-splitting methods involving backward integration. Otherwise, one should expect a step-size restriction to preserve positivity when using operator-splitting methods with negative coefficients.

Author contribution S.W. and R.S. wrote to the main manuscript text. S.W. prepared Figs. 1-4. All authors reviewed the manuscript.

Funding This work was supported by the National Sciences and Engineering Research Council of Canada through its Discovery Grant program.

Availability of data and materials The data that support the findings of this study are available from the corresponding author, S.W., upon reasonable request.

Declarations

Ethics approval and consent to participate Not applicable

Consent for publication Not applicable

Competing interests The authors declare no competing interests.

References

1. Hundsdorfer, W., Verwer, J.G.: Numerical solution of time-dependent advection-diffusion-reaction equations, vol. 33. Springer, Berlin (2003)
2. Lukassen, A.A., Kiehl, M.: Operator splitting for chemical reaction systems with fast chemistry. J. Comput. Appl. Math. **344**, 495–511 (2018)
3. Kermack, W.O., McKendrick, A.G.: A contribution to the mathematical theory of epidemics. Proc. R. Soc. Lond. A **115**, 700–721 (1927)
4. Robertson, H.: The solution of a set of reaction rate equations. Numer. Anal. Introduction. 178182 (1966)
5. Hairer, E., Norsett, S.P., Wanner, G.: Solving ordinary differential equations II: stiff and differential-algebraic problems. Solving ordinary differential equations II: stiff and differential-algebraic problems. Springer, New York (1993). https://books.google.ca/books?id=m7c8nNLPwaIC
6. Marchuk, G.I.: On the theory of the splitting-up method. In: Numerical Solution of Partial Differential Equations-II, pp. 469–500. Academic Press, Maryland (1971)
7. Yanenko, N.N.: The Method of Fractional Steps. The Solution of Problems of Mathematical Physics in Several Variables, p. 160. Springer, New York,: Translated from the Russian by T. Holt, Cheron. English translation edited by M (1971)
8. Glowinski, R., Osher, S.J., Yin, W. (eds.): Splitting methods in communication, imaging, science, and engineering. Scientific Computation, p. 820. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-41589-5
9. Shampine, L.F.: Conservation laws and the numerical solution of ODEs. Comput. Math. Appl. **12**(5–6), 1287–1296 (1986)
10. Hairer, E., Wanner, G., Lubich, C.: Geometric numerical integration: structure preserving algorithms for ordinary differential equations, vol. 31. Springer, Berlin (2006)
11. Spijker, M.N.: Stepsize conditions for general monotonicity in numerical initial value problems. SIAM J. Numer. Anal. **45**(3), 1226–1245 (2007). https://doi.org/10.1137/060661739
12. Blanes, S., Iserles, A., Macnamara, S.: Positivity-preserving methods for ordinary differential equations. ESAIM. Math. Model. Numer. Anal. **56**(6), 1843–1870 (2022). https://doi.org/10.1051/m2an/2022042

13. Magnus, W.: On the exponential solution of differential equations for a linear operator. Commun. Pur. Appl. Math. **7**, 649–673 (1954). https://doi.org/10.1002/cpa.3160070404
14. Kopecz, S., Meister, A.: Unconditionally positive and conservative third order modified Patankar-Runge-Kutta discretizations of production-destruction systems. BIT. Numer. Math. **58**(3), 691–728 (2018). https://doi.org/10.1007/s10543-018-0705-1
15. Izgin, T., Kopecz, S., Meister, A.: On the stability of unconditionally positive and linear invariants preserving time integration schemes. SIAM J. Numer. Anal. **60**(6), 3029–3051 (2022). https://doi.org/10.1137/22M1480318
16. Boris, J.P., et al.: Relativistic plasma simulation-optimization of a hybrid code. In: Proceedings: Fourth Conference on Numerical Simulation of Plasmas, pp. 3–67 (1970)
17. Wei, S., Spiteri, R.J.: Qualitative property preservation of high-order operator splitting for the sir model. Appl. Numer. Math. **172**, 332–350 (2022)
18. Bernier, J., Casas, F., Crouseilles, N.: Splitting methods for rotations: application to Vlasov equations. SIAM Journal on Scientific Computing **42**(2), 666–697 (2020). https://doi.org/10.1137/19M1273918
19. Goldman, G., Kaper, T.J.: Nth-order operator splitting schemes and nonreversible systems. SIAM J. Numer. Anal. **33**(1), 349–367 (1996)
20. Auzinger, W., Hofstätter, H., Ketcheson, D., Koch, O.: Practical splitting methods for the adaptive integration of nonlinear evolution equations. Part I: Construction of optimized schemes and pairs of schemes. BIT Numer. Math. 57(1), 55–74 (2017)
21. MacNamara, S., Strang, G.: Operator splitting. In: Splitting Methods in Communication, Imaging, Science, and Engineering. Sci. Comput., pp. 95–114. Springer, Cham (2016)
22. Spiteri, R.J., Tavassoli, A., Wei, S., Smolyakov, A.: Practical 3-splitting beyond Strang (2023)
23. Csomós, P., Takács, B.: Operator splitting for space-dependent epidemic model. Appl. Numer. Math. **159**, 259–280 (2021)
24. Burchard, H., Deleersnijder, E., Meister, A.: A high-order conservative patankartype discretisation for stiff systems of production-destruction equations. Appl. Numer. Math. **47**(1), 1–30 (2003)