Applied Numerical Mathematics 159 (2021) 259–280

Contents lists available at ScienceDirect

Applied Numerical Mathematics

www.elsevier.com/locate/apnum

Operator splitting for space-dependent epidemic model

Petra Csomós *, Bálint Takács

Institute of Mathematics, Eötvös Loránd University and MTA–ELTE Numerical Analysis and Large Networks Research Group, Pázmány Péter sétány 1/C, H-1117 Budapest, Hungary

A R T I C L E I N F O

Article history:
Received 3 December 2019
Received in revised form 15 September 2020
Accepted 16 September 2020
Available online 21 September 2020

Keywords:
Space-dependent epidemic model
Operator splitting procedures
Qualitative properties
Non-negativity preservation

A B S T R A C T

We present and analyse numerical methods with operator splitting procedures, applied to an epidemic model which takes into account the space-dependence of the infection. We derive conditions on the time step, under which the numerical methods preserve the non-negativity and monotonicity properties of the exact solution. Our results are illustrated by numerical experiments.

1. Introduction

Epidemic modelling plays an increasingly important role not only in applied mathematics but also in medicine and public health. There is, for instance, a high demand on planning the right place and time of vaccination. The more complex these models are, the less hope we have in obtaining their analytical solution. Thus, the derivation and analysis of biologically adequate numerical methods means a vital challenge.

Epidemic models originate from the seminal work of Kermack and McKendrick [9] published in 1927, who constructed a compartment model to study the process of epidemic propagation. The population is split into three classes: healthy but susceptible individuals, infected people who can infect other individuals, and already recovered or otherwise immune individuals. The first attempts describe two ways the individuals can "change" classes: (i) susceptible individuals get infected with some possibility, and (ii) infected people recover with some other rate of change. There are several directions the original model can be generalised: by considering birth and death processes, by adding more classes of individuals, by considering a latent period, or, as in the present paper, by taking into account the effect of vaccination. In the present work we analyse an epidemic model which also treats the space-dependency of the effect of the infection, that is, the distance between the susceptible and infected individuals.

The novelty of our work is to apply operator splitting procedures when discretising in time. They allow us to split the model into two sub-problems, and solve them one after the other. With the help of operator splitting, the difficulty caused by the space-dependency can be handled separately. Moreover, the exact solution of the remaining part can be computed leading to a more stable and accurate numerical solution.

Since epidemic models describe real-life phenomena, it is vital to study whether they reflect the properties expected from the biological point of view. Therefore, we put an effort to investigate under which conditions the model's numerical

* Corresponding author.
E-mail address: csomos@cs.elte.hu (P. Csomós).

https://doi.org/10.1016/j.apnum.2020.09.010

solution owns the non-negativity and monotonicity properties. We are also after the cases when our method gives higher bound on the time step than the one already presented in the literature. We illustrate our theoretical results by numerical experiments.

Section 2 gives an overview on basic epidemic models and shows how we treat the space-dependency of infection. In Section 3 we introduce the space and time discretisation methods which are used later. In Section 4 we define the qualitative properties to be investigated in the rest of the paper. Section 5 contains some necessary technical tools. Sections 6, 7, and 8 are devoted to the analysis of the sequential, weighted, and Strang splittings, respectively. In Section 9 we present our numerical experiments illustrating the theoretical results. Section 10 briefly summarises our results.

2. Space-dependent SIR model

Most of the currently used and analysed models are derived from the idea of Kermack and McKendrick [9], who constructed the compartment model, introduced above, to study the process of epidemic propagation. Let $S, I, R\colon \mathbb{R}_0^+ \to \mathbb{R}_0^+$ denote the density of susceptible, infected, and recovered individuals among the total population, respectively, and let the constant parameters $a, b > 0$ describe the rate of infection and recovery, respectively. Let $S_0, I_0, R_0 \geq 0$ be given numbers. Then the susceptible–infected–recovered (SIR) epidemic model has the form

$$\begin{cases} S'(t) = -aS(t)I(t), \\ I'(t) = aS(t)I(t) - bI(t), \\ R'(t) = bI(t) \end{cases} \tag{1}$$

for all $t > 0$ with the initial condition

$$S(0) = S_0, \quad I(0) = I_0, \quad R(0) = R_0. \tag{2}$$

The SIR model (1)–(2) is an initial value problem being a system of three ordinary differential equations. Although model (1) already describes several important features real epidemics posses, it does not take into account the spatial distribution of the different species, and supposes a homogeneous distribution on the domain. A model which considers the aforementioned properties was introduced by Kendall [8] in the following way.

For an arbitrary $m \in \mathbb{N}$, we consider a bounded domain $\Omega \subset \mathbb{R}^m$ and the open ball $B_\delta(\mathbf{x})$ around the point $\mathbf{x} \in \Omega$ with radius $\delta > 0$. Let $|B_\delta(\mathbf{x})|$ denote its Lebesgue measure (or volume), and $X_{B_\delta(\mathbf{x})}(t)$ the number of individuals in this ball for each $X \in \{S, I, R\}$ at time $t \geq 0$. Then the density of class X at point $\mathbf{x} \in \Omega$ and at time $t > 0$ is defined as

$$\widetilde{X}(t, \mathbf{x}) := \lim_{\delta \to 0} \frac{1}{|B_\delta(\mathbf{x})|} X_{B_\delta(\mathbf{x})}(t).$$

To ease the notation, we will omit the tilde, and denote the density by $X(t, \mathbf{x})$ for each $X \in \{S, I, R\}$. The consideration above leads to a space-dependent SIR model which is, however, at this point not so beneficial because the density functions behave independently at each point $\mathbf{x} \in \Omega$. Since the infection takes place pointwise, it cannot spread in space, being however the main goal of the generalization. Thus, it is more natural to suppose that the infected individuals have an influence on the susceptibles in a certain distance around themselves in such a way that they less likely infect healthy individuals further away from themselves. That is, a susceptible can get infected only in a predefined domain, e.g., a circle. We note that the radius $\delta > 0$ of the infectious domain can vary depending on the disease considered. We further suppose that the disease process is the same at every point $\mathbf{x} \in \Omega$.

Since it is the most common way, we also formulate our model in two dimensions and suppose a rectangular domain $\Omega = [0, A] \times [0, B] \subset \mathbb{R}^2$ with some $A, B > 0$ arbitrary numbers (although the results of this paper can be extended to more general domains). Around the point $\mathbf{x} = (x, y) \in \Omega$, we denote the infectious domain by $B_\delta(x, y)$, being the circle with origin (x, y) and radius $\delta > 0$. To this end, let $r \geq 0$ denote an arbitrary point's distance from (x, y) and $\vartheta \in [0, 2\pi)$ its angle. The main idea is to replace $I(t)$ in the terms $\pm S(t)I(t)$ in the space-dependent version of model (1) by the following weighted integral on the ball $B_\delta(x, y)$:

$$\mathcal{I}(t, x, y) := \int_0^\infty \int_0^{2\pi} G(x, y, r, \vartheta) I(t, x + r\cos(\vartheta), y + r\sin(\vartheta))\, r \,\mathrm{d}\vartheta \,\mathrm{d}r \tag{3}$$

where function $G\colon \Omega \times \mathbb{R}_0^+ \times [0, 2\pi) \to \mathbb{R}_0^+$ describes the disease process in Ω. Since we aim at imitating the effect of an infected individual at the point $(x, y) \in \Omega$ on its δ-radius neighbourhood $B_\delta(x, y)$, we want function G to represent the combined effect of (i) the non-negative and monotonically decreasing function $g_1\colon [0, \delta] \to \mathbb{R}_0^+$ which describes the dependence on the radius r, and (ii) the non-negative function $g_2\colon [0, 2\pi) \to \mathbb{R}_0^+$ describing the dependence on the angle. For the sake of simplicity, we build the infectious rate $a > 0$ into function g_1. We remark that the case of constant function g_2 is widely studied in [5] and [6]. A non-constant g_2 may be useful for modelling the spread of diseases when there is a constant wind blowing in one direction which was described in [14]. We also suppose that the function g_2 is periodic in the way $g_2(0) = \lim_{\vartheta \to 2\pi} g_2(\vartheta)$. Since it is a natural assumption, we take the function G being separable in r and ϑ:

$$G(x, y, r, \vartheta) = \begin{cases} g_1(r)g_2(\vartheta), & \text{if } \big(x + r\cos(\vartheta), y + r\sin(\vartheta)\big) \in B_\delta(x, y), \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Then the term $\mathcal{I}$ in relation (3) has the following form:

$$\mathcal{I}(t, x, y) = \int_0^\delta \int_0^{2\pi} g_1(r)g_2(\vartheta)I(t, x + r\cos(\vartheta), y + r\sin(\vartheta))\, r\, \mathrm{d}\vartheta\, \mathrm{d}r \tag{5}$$

for all $t \geq 0$, $(x, y) \in \Omega$.

To consider a more realistic model than (1), we take into account the effect of vaccination as well. Let $c > 0$ denote the rate related to the vaccinated population getting immune. Then for the new unknown functions $S, I, R\colon \mathbb{R}_0^+ \times \Omega \to \mathbb{R}_0^+$, we get the following system of integro-differential equations:

$$\begin{cases} \partial_t S(t, x, y) = -S(t, x, y)\mathcal{I}(t, x, y) - cS(t, x, y), \\ \partial_t I(t, x, y) = S(t, x, y)\mathcal{I}(t, x, y) - bI(t, x, y), \\ \partial_t R(t, x, y) = cS(t, x, y) + bI(t, x, y) \end{cases} \tag{6}$$

for all $t \geq 0$, $(x, y) \in \Omega$, and with the initial condition

$$S(0, x, y) = S_0(x, y), \quad I(0, x, y) = I_0(x, y), \quad R(0, x, y) = R_0(x, y), \tag{7}$$

where $S_0, I_0, R_0\colon \Omega \to \mathbb{R}_0^+$ are given continuous functions such that

$$S_0(x, y) + I_0(x, y) + R_0(x, y) \neq 0 \text{ holds for all } (x, y) \in \Omega. \tag{8}$$

As already mentioned, one cannot hope to find an analytical solution to system (6), although it was proved in [13] that such a solution exists, which is also unique. Therefore, we are going to use numerical methods to solve these equations. In the next section we introduce the space and time discretisation methods to be used in the present study.

3. Discretisation methods

The present section aims at introducing the space and time discretisation methods of the space-dependent SIR model (6) as well as the operator splitting procedures. Since it is the most challenging part of the numerical method being constructed, first we show how we approximate the integral appearing in (5).

3.1. Approximating the integral

The key point of the numerical solution of problem (6) is the approximation of the double integral in (5), which can be done in different ways. One approach is to approximate the function $I(t, x+r\cos(\vartheta), y+r\sin(\vartheta))$ by a Taylor expansion: the obtained method is studied in [5] and [6]. We note that this process is not efficient in the case of non-constant function g_2 as shown in [14]. The other approach is to use a combination of interpolation and numerical integration (by using cubature formulas). For the present study we implement the second approach.

We consider a two dimensional cubature formula on the disc B_δ with positive coefficients. For index set $\mathcal{J} \subset \mathbb{N}^2$ and for all $(i, j) \in \mathcal{J}$, let $r_i \in [0, \delta]$ denote the (i, j)th cubature points' distance from the centre point $(x, y) \in \Omega$, and $\vartheta_j \in [0, 2\pi)$ its angle. Then $\mathcal{Q}(x, y)$ denotes the set of cubature points in the disk $B_\delta(x, y)$ parametrized by polar coordinates (see [14] or [13]):

$$\mathcal{Q}(x, y) = \Big\{\big(x + r_i\cos(\vartheta_j), y + r_i\sin(\vartheta_j)\big) \in \operatorname{Int} B_\delta(x, y),\ (i, j) \in \mathcal{J}\Big\},$$

where Int denotes the interior of the set. Numerical integration leads then to the following approximation of the term $\mathcal{I}(t, x, y)$ in (6):

$$T(t, \mathcal{Q}(x, y)) = \sum_{(i,j)\in\mathcal{J}} w_{i,j}\, g_1(r_i)g_2(\vartheta_j)I\big(t, x + r_i\cos(\vartheta_j), y + r_i\sin(\vartheta_j)\big) \tag{9}$$

with some weights $w_{i,j} \geq 0$. For the infected individuals being closer to the boundary of the domain Ω as the radius δ, the approximation of the integral in (5) needs values of I lying outside Ω: for these, we are going to use zero values. After these considerations we get the following system, being still continuous in $t \geq 0$ and $(x, y) \in \Omega$:

$$\begin{cases} \partial_t S(t,x,y) = -S(t,x,y)T(t,\mathcal{Q}(x,y)) - cS(t,x,y), \\ \partial_t I(t,x,y) = S(t,x,y)T(t,\mathcal{Q}(x,y)) - bI(t,x,y), \\ \partial_t R(t,x,y) = cS(t,x,y) + bI(t,x,y) \end{cases} \tag{10}$$

with the original initial condition (7). We note that there are several possibilities how to choose the quadratures. One can use a direct method which results in a uniform cubature, or transform the ball onto a square, and use generalised Gaussian quadratures on it. The results in [13] show that for less quadrature points, the uniform ones result in a smaller error, while for a denser quadrature, the non-uniform ones perform better. One can also compute the convolution in (5) by using the Fast Fourier Transform, which can be the direction of further research.

3.2. Spatial discretisation

In order to discretise problem (10) in space, we need a spatial grid $\mathcal{G}$ on the domain $\Omega = [0, A] \times [0, B]$. To this end we choose the arbitrary numbers $K, L \in \mathbb{Z}^+$ and define the grid resolutions $h_x := A/(K-1) > 0$ and $h_y := B/(L-1) > 0$ in directions x and y, respectively. Then the grid itself is the following set:

$$\mathcal{G} := \big\{(x_k, y_\ell) \in \Omega \mid x_k = (k-1)h_x,\ y_\ell = (\ell-1)h_y,\ k = 1, \dots, K,\ \ell =, \dots, L\big\}.$$

For all $t \geq 0$, $(x_k, y_\ell) \in \mathcal{G}$, and $X \in \{S, I, R\}$, we consider the approximate numbers

$$X_{k,\ell}(t) \approx X(t, x_k, y_\ell) \quad \text{and} \quad T_{k,\ell}(t) \approx T(t, \mathcal{Q}(x_k, y_\ell)). \tag{11}$$

In order to determine the form of $T_{k,\ell}(t)$, we first project $I(t,x,y)$ to the grid $\mathcal{G}$. Note that the points $(x_k + r_i\cos(\Theta_j), y_\ell + r_i\sin(\Theta_j))$ might not be part of the grid $\mathcal{G}$, so we cannot assign any $I_{k,\ell}$ values to them. Therefore, we approximate them by a bilinear interpolating method using the nearest known $I_{k,\ell}$ values and positive coefficients, resulting in the notation $\widetilde{I}$. Then we have

$$T_{k,\ell}(t) := \sum_{(i,j)\in\mathcal{J}} w_{i,j} f_1(r_i) f_2(\Theta_j) \widetilde{I}\big(t, x_k + r_i\cos(\Theta_j), y_\ell + r_i\sin(\Theta_j)\big). \tag{12}$$

It is worth mentioning that higher order interpolations, like cubic and spline, can be also used. Although they do not preserve non-negativity, for a sufficiently small spatial grid resolution they behave as expected. It is also possible to use other, high order, non-negativity preserving methods, see [13]. We note here that if $\widetilde{I} \geq 0$ holds, then $T_{k,\ell}(t)$ is non-negative too, for all $t \geq 0$ and $(x_k, y_\ell) \in \mathcal{G}$. In order to ease the notation, we will leave the tilde throughout our computations.

3.3. Time discretisation

The main novelty of the paper is that (besides the traditional time discretisation) we use another time discretisation-like method: operator splitting. As one can see, the right-hand side of problem (10) can be written as a sum of two terms: one containing the integral and one with the remaining terms. The idea of operator splitting is to “split” the problem into two sub-problems with the corresponding terms alone, and solve them separately by using an appropriate initial condition to link their solutions together. In the present paper we will introduce and study the sequential, the sequential weighted, and the Strang splitting schemes.

As already mentioned, it is natural to split the space-discretised SIR model (10) into the sub-problems with and without the integral term $\mathcal{I}$ specifying the space-dependency of the infection process:

$$\begin{cases} \partial_t S^{[1]}(t,x,y) = -cS^{[1]}(t,x,y), \\ \partial_t I^{[1]}(t,x,y) = -bI^{[1]}(t,x,y), \\ \partial_t R^{[1]}(t,x,y) = bI^{[1]}(t,x,y) + cS^{[1]}(t,x,y) \end{cases} \tag{Sub.1}$$

and

$$\begin{cases} \partial_t S^{[2]}(t,x,y) = -S^{[2]}(t,x,y)\mathcal{I}^{[2]}(t,x,y), \\ \partial_t I^{[2]}(t,x,y) = S^{[2]}(t,x,y)\mathcal{I}^{[2]}(t,x,y), \\ \partial_t R^{[2]}(t,x,y) = 0 \end{cases} \tag{Sub.2}$$

for all $t \geq 0$ and $(x, y) \in \Omega$. The link between the sub-problems is the initial condition, as will be shown in the next sections.

For the later use we remark that sub-problem (Sub.1) can be solved exactly:

$$\begin{cases} S^{[1]}(t+\tau,x,y) = \mathrm{e}^{-c\tau} S^{[1]}(t,x,y), \\ I^{[1]}(t+\tau,x,y) = \mathrm{e}^{-b\tau} I^{[1]}(t,x,y), \\ R^{[1]}(t+\tau,x,y) = R^{[1]}(t,x,y) + (1-\mathrm{e}^{-c\tau})S^{[1]}(t,x,y) + (1-\mathrm{e}^{-b\tau})I^{[1]}(t,x,y) \end{cases} \tag{13}$$

for all $t \geq 0$ and $(x, y) \in \Omega$, where $\tau \geq 0$ is an arbitrary time difference.

On the other hand, sub-problem (Sub.2) cannot be solved exactly. Its approximate solution can be obtained by another time discretisation method. For instance, the use of the first-order explicit Euler method with time step $\tau > 0$ leads to

$$\begin{cases} S^{[2]}((n+1)\tau, x, y) = S^{[2]}(n\tau, x, y) - \tau S^{[2]}(n\tau, x, y)\mathcal{I}^{[2]}(n\tau, x, y), \\ I^{[2]}((n+1)\tau, x, y) = I^{[2]}(n\tau, x, y) + \tau S^{[2]}(n\tau, x, y)\mathcal{I}^{[2]}(n\tau, x, y), \\ R^{[2]}((n+1)\tau, x, y) = R^{[2]}(n\tau, x, y) \end{cases} \tag{14}$$

for all $n \in \mathbb{N}$ with $X^{[2]}(0, x, y) = X_0(x, y)$ for each $X \in \{S, I, R\}$. We note that we take $0 \in \mathbb{N}$.

The use of the second-order Heun's method in Shu–Osher form (which preserves the strong stability, see [7]) with time step $\tau > 0$ results in the following steps:

$$\begin{cases} \widehat{S}^{[2]}((n+1)\tau, x, y) = S^{[2]}(n\tau, x, y) - \tau S^{[2]}(n\tau, x, y)\mathcal{I}^{[2]}(n\tau, x, y), \\ \widehat{I}^{[2]}((n+1)\tau, x, y) = I^{[2]}(n\tau, x, y) + \tau S^{[2]}(n\tau, x, y)\mathcal{I}^{[2]}(n\tau, x, y), \\ \widehat{R}^{[2]}((n+1)\tau, x, y) = R^{[2]}(n\tau, x, y), \end{cases} \tag{15}$$

$$\begin{cases} S^{[2]}((n+1)\tau, x, y) = \frac{1}{2}S^{[2]}(n\tau, x, y) \\ \qquad\qquad + \frac{1}{2}\big(\widehat{S}^{[2]}((n+1)\tau, x, y) - \tau\widehat{S}^{[2]}((n+1)\tau, x, y)\widehat{\mathcal{I}}^{[2]}(n+1)n\tau, x, y)\big), \\ I^{[2]}((n+1)\tau, x, y) = \frac{1}{2}I^{[2]}(n\tau, x, y) \\ \qquad\qquad + \frac{1}{2}\big(\widehat{I}^{[2]}((n+1)\tau, x, y) + \tau S^{[2]}((n+1)\tau, x, y)\mathcal{I}^{[2]}((n+1)\tau, x, y)\big), \\ R^{[2]}((n+1)\tau, x, y) = \frac{1}{2}R^{[2]}(n\tau, x, y) + \frac{1}{2}\widehat{R}^{[2]}((n+1)\tau, x, y) = R^{[2]}(n\tau, x, y). \end{cases} \tag{16}$$

We do not plug formulae (15) into (16), because the method will be more suitable for analysis in its present form.

We note that the use of an additional time discretisation inside one time step might also lead to a non-negativity preserving method. This technique may be applied only for one sub-problem. Then the time step could be chosen independently of the constraints but related to the accuracy of the scheme. For more details on such kind of adaptive time stepping we refer to [4].

4. Qualitative properties

When combining the space and time discretisation methods presented in the previous section, we obtain a numerical method represented by a system of algebraic equations. By denoting the approximation of $X(n\tau, x_k, y_\ell)$ by $X^n_{k,\ell}$ for all $X \in \{S, I, R\}$, the unknown values $X^{n+1}_{k,\ell}$ of these algebraic equations are computed with the help of $X^n_{k,\ell}$ for all $n \in \mathbb{N}$, $(x_k, y_\ell) \in \mathcal{G}$ and $X \in \{S, I, R\}$.

In what follows we list several important properties which reflect real-life expectations. In the present work we will study whether the numerical solution possesses them.

1. By adding the equations of system (10), one obtains that the total size of the population remains constant in time at each space position:

$$\begin{aligned} &\partial_t S(t, x, y) + \partial_t I(t, x, y) + \partial_t R(t, x, y) = 0 \\ &\qquad S(t, x, y) + I(t, x, y) + R(t, x, y) =: C(x, y) \text{ for all } t \geq 0 \text{ and } (x, y) \in \Omega. \end{aligned} \tag{17}$$

   This we expect from the numerical solution as well, i.e., that there exist numbers $N_{k,\ell}$ such that:

$$S^n_{k,\ell} + I^n_{k,\ell} + R^n_{k,\ell} = N_{k,\ell} \text{ for all } n \in \mathbb{N},\ (x_k, y_\ell) \in \mathcal{G}. \tag{P1}$$

2. Since functions S, I, R denote densities, their values should remain non-negative:

$$X(t, x, y) \geq 0 \text{ for all } t \geq 0,\ (x, y) \in \Omega \text{ and } X \in \{S, I, R\}. \tag{18}$$

   We expect the same from the numerical values as well:

$$X^n_{k,\ell} \geq 0 \text{ for all } n \in \mathbb{N},\ (x_k, y_\ell) \in \mathcal{G} \text{ and } X \in \{S, I, R\}. \tag{P2}$$

3. Since infected or recovered individuals cannot be susceptible again, the function S is non-increasing in time

$$S(t, x, y) \geq S(t + \tau, x, y) \text{ for all } t, \tau \geq 0 \text{ and } (x, y) \in \Omega. \tag{19}$$

   The same should hold for the numerical values as well:

$$S^n_{k,\ell} \geq S^{n+1}_{k,\ell} \text{ for all } n \in \mathbb{N},\ (x_k, y_\ell) \in \mathcal{G}. \tag{P3}$$

4. Similarly, the density of recovered individuals cannot decrease in time:

$$R(t,x,y) \le R(t+\tau,x,y) \text{ for all } t,\tau \ge 0 \text{ and } (x,y) \in \Omega. \tag{20}$$

Which means for the numerical values that

$$R^n_{k,\ell} \le R^{n+1}_{k,\ell} \text{ for all } n \in \mathbb{N},\ (x_k,y_\ell) \in \mathcal{G}. \tag{P4}$$

In [13] it was shown that the properties (17)–(20) hold for the systems (6) and (10). So our aim is to construct such numerical methods which preserve their discrete versions (P1)–(P4).

5. Technical tools

Before the derivation and analysis of the methods, we collect some notations and technical tools we will use later on.

Notation 5.1.

(i) For each $X \in \{S, I, R\}$ we introduce the notation

$$X^n := ((X^n_{k,\ell})_{k=1,\dots,K,\ell=1,\dots,L}) \in \mathbb{R}^{KL\times KL}.$$

(ii) Let $\mathcal{M}\colon \mathbb{R}^{KL\times KL} \to \mathbb{R}^{KL\times KL}$ denote the bounded linear operator (represented by a matrix in applications) that maps I^n to T^n by the rule $T^n = \mathcal{M}(I^n)$. Furthermore, let

$$M := \|\mathcal{M}\|_\infty \cdot \|S^0 + I^0 + R^0\|_\infty$$

in which $\|.\|_\infty$ means the maximum matrix norm taken element-wise. We note that condition (8) implies $M > 0$.

(iii) Let $W_{-1}\colon [-1/\mathrm{e}, 0) \to (-\infty, -1]$ and $W_0\colon (-1/\mathrm{e}, +\infty) \to (-1, +\infty)$ denote the two branches of the Lambert-W function, that is, the inverse of the map $x \mapsto x\mathrm{e}^x$.

(iv) For arbitrary $p, q > 0$, we define the set

$$\mathbb{T}_{p,q} := \big[0, -\tfrac{1}{p} W_0\big(-\tfrac{p}{q}\big)\big] \cup \big[-\tfrac{1}{p} W_{-1}\big(-\tfrac{p}{q}\big), +\infty\big) \subset \mathbb{R}.$$

Furthermore, we define

$$\mathbb{T}_{0,q} := [0, \tfrac{1}{q}) \subset \mathbb{R}.$$

The latter notation makes sense because of the following consideration.

Lemma 5.2. *With Notation 5.1, the limit* $-\frac{1}{p} W_0\big(-\frac{p}{q}\big) \xrightarrow{p\to 0} \frac{1}{q}$ *holds for arbitrary* $q > 0$.

Proof. It suffices to show that $W_0(x)/x \xrightarrow{x\to 0} 1$ for $x = -p/q < 0$. The L'Hospital rule, the derivative of the inverse function, and the identity $W_0(0) = 0$ imply that

$$\lim_{x\to 0} \frac{W_0(x)}{x} = \lim_{x\to 0} W_0'(x) = \lim_{x\to 0} \frac{1}{\mathrm{e}^{W_0(x)} + W_0(x)\mathrm{e}^{W_0(x)}} = 1. \quad \square$$

Remark 5.3. Since we will use it several times throughout the paper, we analyse the solution $x < 0$ to equation

$$x\mathrm{e}^x = \mu \tag{21}$$

for some parameter $\mu < 0$.

(i) For $\mu < -1/\mathrm{e}$, there is no solution to equation (21).
(ii) For $\mu = -1/\mathrm{e}$, there is one solution: $x_1 = -1$.
(iii) For $\mu > -1/\mathrm{e}$, there are two solutions: $x_{-1} = W_{-1}(\mu)$ and $x_0 = W_0(\mu)$.

We also know that $x_{-1} \le x_1 = -1 < x_0$. Hence, for the inequality

$$x\mathrm{e}^x \ge \mu \tag{22}$$

we have the following cases.

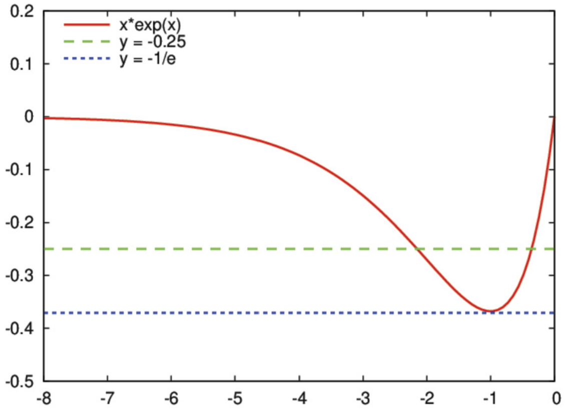

Fig. 1. Graph of function $x \mapsto xe^x$. The horizontal lines indicate the μ-values -0.25 and $-1/e$.

(i) For $\mu < -1/e$, the inequality (22) holds for every $x < 0$.
(ii) For $\mu = -1/e$, the inequality (22) holds for every $x < 0$ (we have $xe^x = \mu$ for $x = -1$).
(iii) For $\mu > -1/e$, we have: $x < x_{-1} = W_{-1}(\mu)$ or $x > x_0 = W_{-1}(\mu)$.

The graph of function $x \mapsto xe^x$ is depicted on Fig. 1.

In the next sections we will present the condition on the time step τ under which the qualitative properties (P1)–(P4) hold for the various operator splitting schemes. We are especially interested in the cases when the application of operator splitting leads to less severe condition than the one obtained without splitting. In [14] the authors applied the same space discretisation as showed in Section 3.2 and the explicit Euler method for the whole system (6) without taking into account the vaccination ($c = 0$). They found that property (P1) was automatically satisfied, and properties (P2)–(P4) held true for time steps τ satisfying

$$\tau \leq \min\left\{\frac{1}{M}, \frac{1}{b}\right\}.$$

The case $c > 0$ was studied in [13], and resulted in a similar bound, namely

$$\tau \leq \min\left\{\frac{1}{M+c}, \frac{1}{b}\right\}. \tag{23}$$

From now on, the upper bound (23) will be considered as a reference value, and we will study the conditions under which the application of operator splitting procedures leads to a higher one.

6. Sequential splitting

Operator splitting is based on the idea to simplify the problem by splitting it into two or more sub-problems which are easier to solve or treat numerically. Since the sub-problems need to be solved separately, we should derive a way to connect their solutions. Depending on these rules, we distinguish several splitting methods. The most basic one is the sequential splitting (initiated first in [1]) when the sub-problems are solved one after the other on a time interval of length $\tau > 0$, always taking the solution of the previous sub-problem as initial condition for the actual one. As we will see, the properties of the sequential splitting depend on the order of the sub-problems, therefore, we will treat the two cases separately.

Another splitting procedure is derived when the solutions of the two types of sequential splittings are weighted by a parameter $\Theta \in [0, 1]$. This kind of method is called weighted sequential splitting, see in [3], and will be discussed in Section 7. The third operator splitting to be discussed in Section 8 is the Strang splitting (derived in [12] and [11]) solving three problems in a single time step: one with the first sub-problem over a time interval of length $\tau/2$, then with the second sub-problem on an interval of length τ, and finally with the first sub-problem again on a $\tau/2$ interval.

In what follows we analyse the splitting procedures in the light of whether they preserve the qualitative properties introduced in Section 4.

6.1. Sequential splitting 1–2

First we treat the sequential splitting in the case when the sub-problems are taken in the order (Sub.1)–(Sub.2). Then the application of the sequential splitting means that in a single time step we first solve sub-problem (Sub.1) whose solution

(13) serves as the initial condition to sub-problem (Sub.2). More precisely, we consider the following iteration steps for all $n \in \mathbb{N}$ and $(x, y) \in \Omega$:

$$\begin{aligned}
&\begin{cases} \text{(Sub.1) for all } t \in (n\tau, (n+1)\tau] \\ \text{with initial condition } X^{[1]}(n\tau, x, y) = X^{\mathrm{spl}}(n\tau, x, y) \end{cases} \\
&\begin{cases} \text{(Sub.2) for all } t \in (n\tau, (n+1)\tau] \\ \text{with initial condition } X^{[2]}(n\tau, x, y) = X^{[1]}((n+1)\tau, x, y) \end{cases} \\
&X^{\mathrm{spl}}((n+1)\tau, x, y) := X^{[2]}((n+1)\tau, x, y)
\end{aligned} \tag{24}$$

where $X^{\mathrm{spl}}(0, x, y) = X_0(x, y)$ is the original initial value in (7) for each $X \in \{S, I, R\}$. After discretising sub-problem (Sub.2) by the explicit Euler method, and discretising in space sub-problems (13) and (14), we get the following two sub-problems:

$$\begin{cases} S_{k,\ell}^{[1],n+1} = \mathrm{e}^{-c\tau} S_{k,\ell}^{[1],n}, \\ I_{k,\ell}^{[1],n+1} = \mathrm{e}^{-b\tau} I_{k,\ell}^{[1],n}, \\ R_{k,\ell}^{[1],n+1} = R_{k,\ell}^{[1],n} + (1-\mathrm{e}^{-c\tau}) S_{k,\ell}^{[1],n} + (1-\mathrm{e}^{-b\tau}) I_{k,\ell}^{[1],n} \end{cases} \tag{25}$$

and

$$\begin{cases} S_{k,\ell}^{[2],n+1} = S_{k,\ell}^{[2],n} - \tau S_{k,\ell}^{[2],n} T_{k,\ell}^{[2],n}, \\ I_{k,\ell}^{[2],n+1} = I_{k,\ell}^{[2],n} + \tau S_{k,\ell}^{[2],n} T_{k,\ell}^{[2],n}, \\ R_{k,\ell}^{[2],n+1} = R_{k,\ell}^{[2],n}. \end{cases} \tag{26}$$

By taking into account the initial conditions as stated in (24), the sub-problems have the following form for all $n \in \mathbb{N}$ and given $S_{k,\ell}^n, I_{k,\ell}^n, R_{k,\ell}^n$:

$$\begin{cases} S_{k,\ell}^{[1],n+1} = \mathrm{e}^{-c\tau} S_{k,\ell}^{n}, \\ I_{k,\ell}^{[1],n+1} = \mathrm{e}^{-b\tau} I_{k,\ell}^{n}, \\ R_{k,\ell}^{[1],n+1} = R_{k,\ell}^{n} + (1-\mathrm{e}^{-c\tau}) S_{k,\ell}^{n} + (1-\mathrm{e}^{-b\tau}) I_{k,\ell}^{n}, \end{cases} \tag{27}$$

$$\begin{cases} S_{k,\ell}^{n+1} = S_{k,\ell}^{[1],n+1} - \tau S_{k,\ell}^{[1],n+1} T_{k,\ell}^{[1],n+1}, \\ I_{k,\ell}^{n+1} = I_{k,\ell}^{[1],n+1} + \tau S_{k,\ell}^{[1],n+1} T_{k,\ell}^{[1],n+1}, \\ R_{k,\ell}^{n+1} = R_{k,\ell}^{[1],n+1}. \end{cases} \tag{28}$$

Notation 5.1(ii) and the linearity of operator $\mathcal{M}$ imply the following relation:

$$T_{k,\ell}^{[1],n+1} = \mathcal{M}(I_{k,\ell}^{[2],n}) = \mathcal{M}(I_{k,\ell}^{[1],n+1}) = \mathcal{M}(\mathrm{e}^{-b\tau} I_{k,\ell}^{n}) = \mathrm{e}^{-b\tau} \mathcal{M}(I_{k,\ell}^{n}) = \mathrm{e}^{-b\tau} T_{k,\ell}^{n}. \tag{29}$$

By combining the sub-problems (27)–(28), and the relation (29), we arrive at the numerical scheme:

$$\begin{cases} S_{k,\ell}^{n+1} = \mathrm{e}^{-c\tau} S_{k,\ell}^{n} (1 - \tau \mathrm{e}^{-b\tau} T_{k,\ell}^{n}), \\ I_{k,\ell}^{n+1} = \mathrm{e}^{-b\tau} (I_{k,\ell}^{n} + \tau \mathrm{e}^{-c\tau} S_{k,\ell}^{n} T_{k,\ell}^{n}), \\ R_{k,\ell}^{n+1} = R_{k,\ell}^{n} + (1-\mathrm{e}^{-c\tau}) S_{k,\ell}^{n} + (1-\mathrm{e}^{-b\tau}) I_{k,\ell}^{n}. \end{cases} \tag{30}$$

In what follows we show the connection between properties (P1)–(P4), and investigate the conditions under which they are fulfilled.

Proposition 6.1. *We have the following assertions.*

(a) *Property* (P1) *holds for the numerical method* (30) *without any restriction.*
(b) *Property* (P3) *and* (P4) *are consequences of property* (P2).

Proof. (a) Property (P1) follows by adding up the equations of system (30).
(b) Since $T_{k,\ell}^n \geq 0$ holds if $I_{k,\ell}^n \geq 0$, and $\mathrm{e}^{-b\tau} > 0$ in the first and the third equations of system (30), we get that properties (P3) and (P4) also hold.

This concludes the proof. □

Due to Proposition 6.1, the monotonicity properties (P3) and (P4) follows from the non-negativity property (P2). Thus, we do not need to treat them separately. Hence, as a next step we study the conditions under which the non-negativity property (P2) holds.

Proposition 6.2. *With Notation 5.1, we have the following assertions.*

(a) *For $M < b\mathrm{e}$, the non-negativity property (P2) is satisfied for all values of time step $\tau > 0$.*
(b) *For $M \geq b\mathrm{e}$, the non-negativity property (P2) holds if $\tau \in \mathbb{T}_{b,M}$.*

Proof. Since the initial values are non-negative, and all steps of the method have the same formulae, it is enough to show the assertion for an arbitrary step. Thus, we suppose that the values $X^n_{k,\ell}$ are non-negative for an arbitrary $n \in \mathbb{N}$, and show the non-negativity of $X^{n+1}_{k,\ell}$, for all $n \in \mathbb{N}$, $(x_k, y_\ell) \in \mathcal{G}$, and $X \in \{S, I, R\}$. The non-negativity of $I^{n+1}_{k,\ell}$ and $R^{n+1}_{k,\ell}$ is trivially satisfied, because all additive terms are non-negative in the second and third equations in (30). In particular, $T^n_{k,\ell} \geq 0$ holds due to its definition (29) for $I^n_{k,\ell} \geq 0$.

Thus, we only have to treat the first equation in problem (30). Condition $S^{n+1}_{k,\ell} \geq 0$ holds if the time step τ fulfils the relation

$$1 - \tau \mathrm{e}^{-b\tau} T^n_{k,\ell} \geq 0. \tag{31}$$

For $T^n_{k,\ell} = 0$, we trivially have $S^{n+1}_{k,\ell} = S^n_{k,\ell} \geq 0$. For $T^n_{k,\ell} > 0$, inequality (31) leads to the condition

$$-\tau \mathrm{e}^{-b\tau} \geq -\frac{1}{T^n_{k,\ell}}.$$

Property (P1) implies $T^n_{k,\ell} \leq M$ for all $n \in \mathbb{N}$ and $(x_k, y_\ell) \in \mathcal{G}$. Hence, we obtain the sufficient condition

$$\begin{aligned} -\tau \mathrm{e}^{-b\tau} &\geq -\frac{1}{M} \\ -b\tau \mathrm{e}^{-b\tau} &\geq -\frac{b}{M}. \end{aligned} \tag{32}$$

With the notations $x := -b\tau < 0$ and $\mu := -b/M < 0$, the inequality (32) has the form $x\mathrm{e}^x \geq \mu$. Due to Remark 5.3, we have now three cases.

(i) For $\mu < -1/\mathrm{e}$ (which means $M < b\mathrm{e}$), all $\tau > 0$ satisfies (32). This proves assertion (a).
(ii) For $\mu = -1/\mathrm{e}$ (which means $M = b\mathrm{e}$), we have $-b\tau = x_1 = -1$, that is, $\tau \neq 1/b$.
(iii) For $\mu > -1/\mathrm{e}$ (which means $M > b\mathrm{e}$), we have

$$x < x_{-1} = W_{-1}(\mu) \quad \text{or} \quad x > x_0 = W_0(\mu).$$

From the notations $x = -b\tau$ and $\mu := -b/M$, we get condition $\tau \in \mathbb{T}_{b,M}$. It remains to show that this bound is well defined. Since W_{-1} is strictly decreasing on $(-1/\mathrm{e}, 0]$ and W_0 is strictly increasing on $(-1/\mathrm{e}, +\infty)$, we have the estimates

$$\begin{aligned} -\tfrac{1}{b} W_0\big(-\tfrac{b}{d}\big) &\leq -\tfrac{1}{b} W_0\big(-\tfrac{b}{T^n_{k,\ell}}\big), \\ -\tfrac{1}{b} W_{-1}\big(-\tfrac{b}{d}\big) &\geq -\tfrac{1}{b} W_{-1}\big(-\tfrac{b}{T^n_{k,\ell}}\big). \end{aligned}$$

Then the cases (ii) and (iii) together prove assertion (b). Note that in case (ii) we got $\mathbb{T}_{b,M} = \mathbb{R}^+ \setminus \{1/b\}$.

With this consideration we proved the non-negativity of $S^{n+1}_{k,\ell}$, and completed the proof. □

Interestingly, the condition $\tau \in \mathbb{T}_{b,M}$ in Proposition 6.2 means that there is a “forbidden interval”

$$\big(-\tfrac{1}{b} W_0(-\tfrac{b}{M}), -\tfrac{1}{b} W_{-1}(-\tfrac{b}{M})\big) \subset \mathbb{R}$$

where τ leads to negative S, I, R values. It is worth mentioning, however, that Proposition 6.2 gives a necessary condition only, so the forbidden interval can be shorter in real applications. The correspondence between the “exact” and the necessary bounds will be investigated in Section 9.

It is important to compare the bounds obtained for the time step in Proposition 6.2 with the similar result obtained for a numerical method without using operator splitting, cf. bound (23).

Proposition 6.3. *With Notation 5.1, we have the following assertions.*

(i) *The estimate* $-\frac{1}{b}W_0\big(-\frac{b}{M}\big) > \frac{1}{M+c}$ *holds for all* $M, b, c > 0$ *with* $M > be$.

(ii) *For an arbitrary* $M > 0$, *we have the limit* $-\frac{1}{b}W_0\big(-\frac{b}{M}\big) \xrightarrow{b\to 0} \frac{1}{M}$.

Proof. (i) The relation $W_0(y) < y$ for all $y < 0$, the strictly increasing of W_0, and the assumption $M > be$ imply the assertion.

(ii) Follows from Lemma 5.2 with $p = b$ and $q = M$. □

Proposition 6.3 means that in the case $M > be$ our method (30) gives a larger upper bound for the time step τ as the application of explicit Euler method without operator splitting. Namely, in this case $M + c > M > be > b$ holds, which leads to $\min\{1/b, 1/(M+c)\} = 1/(M+c)$. Moreover, for the case $M \le be$, our method (30) satisfies the properties (P2)–(P4) without any restriction on the time step τ. Hence, method (30) is more convenient to use than the method proposed in [13].

We note here, that although Proposition 6.2 allows large values for the time step, the use of these is not advised, since it leads to considerable higher error in the numerical solution.

Remark 6.4. Another possible way to perform the time step analysis is to check the non-negativity preservation for each sub-problem separately, and then take the most severe constraint on the time step. In case of the sequential splitting 1–2 (30), however, we obtain a weaker result than the one presented in Proposition 6.2, namely, $\tau \le 1/M$. This can be seen from the following consideration. Sub-problem (27) preserves the non-negativity for all $\tau > 0$, while sub-problem (28) introduces the constraint $\tau \le 1/M$. This bound is always smaller than the one obtained in Proposition 6.2, which can be seen from the proof of Proposition 6.3(i). Thus, the point in analysing the combined method (30) is that it might lead to sharper constrains on the time step, as it does in this case.

6.2. Sequential splitting 2–1

We study now the sequential splitting with the other order of the sub-problems. In a single time step we first solve (Sub.2) and then (Sub.1) with the appropriate initial conditions. Similarly to (24), for all $n \in \mathbb{N}$ and $(x, y) \in \Omega$ we consider the iteration steps

$$\begin{aligned}
&\begin{cases} \text{(Sub.2) for all } t \in (n\tau, (n+1)\tau] \\ \text{with initial condition } X^{[2]}(n\tau, x, y) = X^{\text{spl}}(n\tau, x, y) \end{cases}\\
&\begin{cases} \text{(Sub.1) for all } t \in (n\tau, (n+1)\tau] \\ \text{with initial condition } X^{[1]}(n\tau, x, y) = X^{[2]}((n+1)\tau, x, y) \end{cases}\\
&X^{\text{spl}}((n+1)\tau, x, y) := X^{[1]}((n+1)\tau, x, y)
\end{aligned} \tag{33}$$

where $X^{\text{spl}}(0, x, y) = X_0(x, y)$ is the original initial value in (7) for each $X \in \{S, I, R\}$. Thus, we consider first the space discretised sub-problem (26) and then (25). Then the numerical method takes the form

$$\begin{cases} S^{[1],n+1}_{k,\ell} = S^n_{k,\ell} - \tau S^n_{k,\ell} T^n_{k,\ell}, \\ I^{[1],n+1}_{k,\ell} = I^n_{k,\ell} + \tau S^n_{k,\ell} T^n_{k,\ell}, \\ R^{[1],n+1}_{k,\ell} = R^n_{k,\ell}, \end{cases} \tag{34}$$

$$\begin{cases} S^{n+1}_{k,\ell} = \mathrm{e}^{-c\tau} S^{[1],n+1}_{k,\ell}, \\ I^{n+1}_{k,\ell} = \mathrm{e}^{-b\tau} I^{[1],n+1}_{k,\ell}, \\ R^{n+1}_{k,\ell} = R^{[1],n+1}_{k,\ell} + (1 - \mathrm{e}^{-c\tau}) S^{[1],n+1}_{k,\ell} + (1 - \mathrm{e}^{-b\tau}) I^{[1],n+1}_{k,\ell} \end{cases} \tag{35}$$

for all $n \in \mathbb{N}$ and $(x_k, y_\ell) \in \mathcal{G}$. Combination of sub-problems (34) and (35) yields the method

$$\begin{cases} S^{n+1}_{k,\ell} = \mathrm{e}^{-c\tau} S^n_{k,\ell}(1 - \tau T^n_{k,\ell}), \\ I^{n+1}_{k,\ell} = \mathrm{e}^{-b\tau}(I^n_{k,\ell} + \tau S^n_{k,\ell} T^n_{k,\ell}), \\ R^{n+1}_{k,\ell} = R^n_{k,\ell} + (1 - \mathrm{e}^{-c\tau}) S^n_{k,\ell}(1 - \tau T^n_{k,\ell}) + (1 - \mathrm{e}^{-b\tau})(I^n_{k,\ell} + \tau S^n_{k,\ell} T^n_{k,\ell}). \end{cases} \tag{36}$$

We can state the same result as before.

Proposition 6.5. *Proposition 6.1 holds for the method (36).*

Proof. First, we add up the equations in (36) to obtain property (P1). To prove the next assertion, we consider an arbitrary step again. Property (P2) implies that $T^n_{k,\ell}$ is non-negative. This and $\mathrm{e}^{-c\tau} < 1$ imply that $S^{n+1}_{k,\ell} \le S^n_{k,\ell}$. Moreover, the non-negativity of $S^{n+1}_{k,\ell}$ implies that $1 - \tau T^n_{k,\ell} \ge 0$, therefore, $R^{n+1}_{k,\ell} \ge R^n_{k,\ell}$ holds as well. □

According to Proposition 6.5, it suffices to show the non-negativity property (P2) to obtain the monotonicity properties (P3) and (P4).

Proposition 6.6. *The non-negativity property (P2) holds true for the method (36) if the time step τ satisfies the condition*

$$\tau \le \frac{1}{M} \tag{37}$$

where M is defined in Notation 5.1.

Proof. Since the initial values are non-negative, we assume that $X^n_{k,\ell} \ge 0$ and show that $X^{n+1}_{k,\ell} \ge 0$ for all $n \in \mathbb{N}$, $(x_k, y_\ell) \in \mathcal{G}$, and $X \in \{S, I, R\}$. Since the assumption $\tau \le 1/M$ implies $1 - \tau T^n_{k,\ell} \ge 0$, the non-negativity of $S^{n+1}_{k,\ell}$ is fulfilled. Furthermore, since all additive terms in the second and third equations of (36) are non-negative, we have $I^{n+1}_{k,\ell} \ge 0$ as well as $R^{n+1}_{k,\ell} \ge 0$. □

Hence, for $M + c > b$ we get a better bound for the time step τ than for the explicit Euler method without splitting, cf. (23). If $M + c < b$, it might happen that the bound of the non-split method is better. Since the application of operator splitting usually needs more CPU time than the explicit Euler method itself, it is not advised to use method (36) in this case but the first one (30).

We note that in case of the sequential splitting 2–1 we obtain the same bound (37) on the time step τ both when analysing the combined method (36) or the separate sub-problems (34) and (35).

7. Weighted sequential splitting

Especially on parallel computers, it is a good idea to combine the solution to sequential splittings 1–2 and 2–1 with some $\Theta \in [0, 1]$ parameter as follows:

$$X = \Theta \cdot X_{(30)} + (1 - \Theta) \cdot X_{(36)}$$

where $X_{(30)}$, $X_{(36)}$ denotes the approximate solutions obtained by numerical methods (30) and (36), respectively, for each $X \in \{S, I, R\}$. We note that the choice $\Theta = 0$ results in the method (36), while $\Theta = 1$ gives (30). In this way we get the following numerical method:

$$\begin{cases} S^{n+1}_{k,\ell} = \mathrm{e}^{-c\tau} S^n_{k,\ell}\big(1 - \tau(\Theta \mathrm{e}^{-b\tau} + (1-\Theta))T^n_{k,\ell}\big), \\ I^{n+1}_{k,\ell} = \mathrm{e}^{-b\tau}\big(I^n_{k,\ell} + \tau(\Theta \mathrm{e}^{-c\tau} + (1-\Theta))S^n_{k,\ell}T^n_{k,\ell}\big), \\ R^{n+1}_{k,\ell} = R^n_{k,\ell} + (1 - \mathrm{e}^{-c\tau})S^n_{k,\ell}(1 - (1-\Theta)\tau T^n_{k,\ell}) + (1 - \mathrm{e}^{-b\tau})(I^n_{k,\ell} + (1-\Theta)\tau S^n_{k,\ell}T^n_{k,\ell}). \end{cases} \tag{38}$$

As before, we investigate the validity of properties (P1)–(P4).

Proposition 7.1. *Proposition 6.1 is valid for method (38).*

Proof. The conservation of the size of the population is obtained again by adding up the equations in (38). Since $\Theta, \mathrm{e}^{-b\tau}, \mathrm{e}^{-c\tau} \in (0, 1)$ and $T^n_{k,\ell} \ge 0$ in the first equation of (38), we have $S^{n+1}_{k,\ell} \le S^n_{k,\ell}$. Due to property (P2), all terms in the third equation of (38) are non-negative, therefore, $R^{n+1}_{k,\ell} \ge R^n_{k,\ell}$ holds true. □

In order to study the non-negativity preservation (P2), we need the following notation.

Notation 7.2. For the parameter $\Theta \in [0, 1]$, we define

$$\Theta^* := \frac{\mathrm{e}^2}{\mathrm{e}^2 + 1} \approx 0.8808.$$

It will turn out that we get remarkably different bounds for Θ being under or above Θ^*.

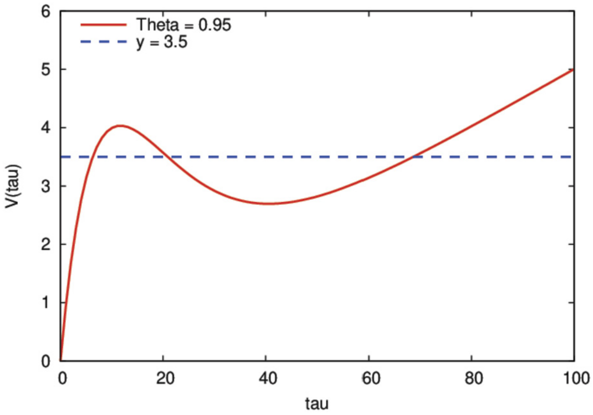

Fig. 2. Graph of function $V_{\Theta,b}(\tau) = \tau(1 - \Theta(1 - \mathrm{e}^{-b\tau}))$ for $\Theta = 0.95$ and $b = 0.1$. The horizontal line indicates the value 3.5.

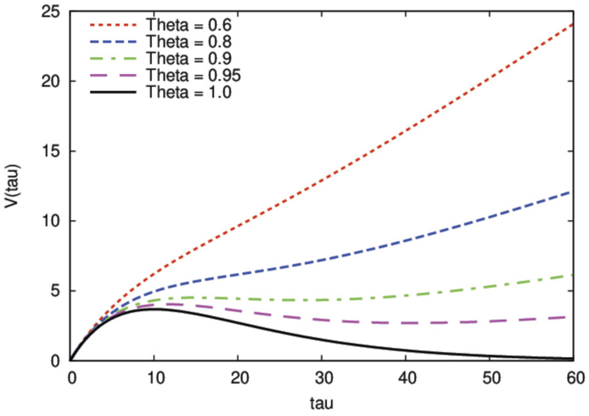

Fig. 3. Graph of function $V_{\Theta,b}(\tau) = \tau(1 - \Theta(1 - \mathrm{e}^{-b\tau}))$ for $\Theta = 0.6$, 0.8, 0.9, 0.95, 1 and $b = 0.1$.

Notation 7.3.

(a) For $\Theta \in [0, 1]$ and $b > 0$, we define the function $V_{\Theta,b}\colon \mathbb{R}^+ \to (0, +\infty)$ as

$$V_{\Theta,b}(\tau) = \tau(1 - \Theta(1 - \mathrm{e}^{-b\tau})).$$

(b) We introduce the values $0 < \tau_0 < \tau_{-1}$ as

$$\tau_{-1} := \tfrac{1}{b}\big(1 - W_{-1}\big(\tfrac{\mathrm{e}(\Theta-1)}{\Theta}\big)\big) \quad \text{for} \quad \Theta \in [\Theta^*, 1),$$
$$\tau_0 := \tfrac{1}{b}\big(1 - W_0\big(\tfrac{\mathrm{e}(\Theta-1)}{\Theta}\big)\big) \quad \text{for} \quad \Theta \in [\Theta^*, 1].$$

On Fig. 2 the graph of function $V_{\Theta,b}$ is shown for $\Theta = 0.95$ and $b = 0.1$. In order to illustrate its dependence on Θ, we present the graph of function $V_{\Theta,b}$ for various values of Θ and $b = 0.1$ on Fig. 3.

7.1. Case of "small" Θ

We take now $\Theta \in [0, \Theta^*)$, and examine first whether the inverse of $V_{\Theta,b}$ exists.

Lemma 7.4. *For $\Theta \in [0, \Theta^*)$, function $V_{\Theta,b}$ is strictly increasing, thus, $V_{\Theta,b}^{-1}$ exists, and is strictly increasing on $(0, +\infty)$.*

Proof. To show that function $V_{\Theta,b}$ is monotone, we calculate its derivative with respect to τ:

$$V'_{\Theta,b}(\tau) = \tfrac{\mathrm{d}}{\mathrm{d}\tau}\big(\tau(1 - \Theta(1 - \mathrm{e}^{-b\tau}))\big) = (1 - \Theta) + \Theta\mathrm{e}^{-b\tau}(1 - b\tau).$$

We now determine its zeros:

$$\begin{aligned}V'_{\Theta,b}(\tau) &= 0\\ (1-\Theta)+\Theta \mathrm{e}^{-b\tau}(1-b\tau) &= 0\\ \mathrm{e}^{-b\tau}(1-b\tau) &= \frac{\Theta-1}{\Theta}\\ \mathrm{e}^{1-b\tau}(1-b\tau) &= \mathrm{e}\frac{\Theta-1}{\Theta}.\end{aligned}$$

With the notations $x := 1-b\tau$ and $\mu := \mathrm{e}(\Theta-1)/\Theta < 0$, we need to examine the solutions x to the equation (21). The relation $\Theta < \Theta^*$ implies $\mu < -1/\mathrm{e}$, hence, there is no solution x to equation (21) according to Remark 5.3. Thus, there are no zeros of function $V'_{\Theta,b}$, therefore, $V_{\Theta,b}$ is monotone. Furthermore, $V'_{\Theta,b}(1/b) = 1-\Theta > 0$ implies that function $V_{\Theta,b}$ is increasing on $(0,+\infty)$. Hence, its inverse $V^{-1}_{\Theta,b}$ exists and is strictly increasing on $(0,+\infty)$. □

We state now the result for the non-negativity preservation.

Proposition 7.5. *For $\Theta \in [0,\Theta^*)$, the non-negativity property (P2) holds for the method (38) if the time step τ satisfies the following criterion:*

$$\tau \le V^{-1}_{\Theta,b}\big(\tfrac{1}{M}\big). \tag{39}$$

Proof. Since the initial values are non-negative, we treat an arbitrary step. We assume $X^n_{k,\ell} \ge 0$ and show $X^{n+1}_{k,\ell} \ge 0$ for all $n \in \mathbb{N}$, $(x_k, y_\ell) \in \mathcal{G}$, and $X \in \{S, I, R\}$. The non-negativity of $I^{n+1}_{k,\ell}$ and $R^{n+1}_{k,\ell}$ follows immediately, because all additive terms are non-negative in the second and third equations of system (38).

From the first equation in (38), $S^{n+1}_{k,\ell}$ is non-negative if

$$\begin{aligned}1-\tau T^n_{k,\ell}\big(1-\Theta(1-\mathrm{e}^{-b\tau})\big) &\ge 0\\ \tau\big(1-\Theta(1-\mathrm{e}^{-b\tau})\big) &\le \frac{1}{T^n_{k,\ell}}\\ V_{\Theta,b}(\tau) &\le \frac{1}{T^n_{k,\ell}}.\end{aligned}$$

Due to relation $\Theta < \Theta^* = \mathrm{e}^2/(\mathrm{e}^2+1)$ and Lemma 7.4, function $V_{\Theta,b}$ is strictly increasing and its inverse is well-defined on $(0,+\infty)$. Thus, we have the following bound for the time step τ:

$$\tau \le V^{-1}_{\Theta,b}\Big(\tfrac{1}{T^n_{k,\ell}}\Big). \tag{40}$$

Property (P1) implies $T^n_{k,\ell} \le M$ for all $n \in \mathbb{N}$ and $(x_k, y_\ell) \in \mathcal{G}$. Hence, the inequality (40) is fulfilled due to Assumption (39) and since $V^{-1}_{\Theta,b}$ is strictly decreasing on $(0,+\infty)$. □

7.2. Case of "large" Θ

We take now $\Theta \in [\Theta^*, 1]$, and examine the behaviour of function $V_{\Theta,b}$.

Lemma 7.6. *With Notations 5.1 and 7.3, we have the following assertions.*

(a) *For $\Theta \in [\Theta^*, 1)$, we have the following strictly monotonicity segments of function $V_{\Theta,b}$:*
   (i) *on $(0, \tau_0)$ the function $V_{\Theta,b}$ is strictly increasing, therefore, its inverse V_1^{-1} exists and is strictly increasing,*
   (ii) *on (τ_0, τ_{-1}) the function $V_{\Theta,b}$ is strictly decreasing, therefore, its inverse V_2^{-1} exists and is strictly decreasing,*
   (iii) *on $(\tau_{-1}, +\infty)$ the function $V_{\Theta,b}$ is strictly increasing, therefore, its inverse V_3^{-1} exists and is strictly increasing*

(b) *For $\Theta = 1$, the inverse of function $V_{1,b} = \tau\mathrm{e}^{-b\tau}$ is strictly increasing on $[0, 1/b)$ and decreasing on $(1/b, +\infty)$.*

Proof. Similarly to the proof of Lemma 7.4, we need to examine the function

$$V'_{\Theta,b}(\tau) = \tfrac{\mathrm{d}}{\mathrm{d}\tau}\big(\tau(1-\Theta(1-\mathrm{e}^{-b\tau}))\big) = (1-\Theta)+\Theta\mathrm{e}^{-b\tau}(1-b\tau)$$

for $\Theta \in [\Theta^*, 1)$ and determine its zeros:

$$\begin{aligned}\tfrac{\mathrm{d}}{\mathrm{d}\tau}V_{\Theta,b}(\tau) &= 0\\ \mathrm{e}^{-b\tau}(1-b\tau) &= \frac{\Theta-1}{\Theta}\\ \mathrm{e}^{1-b\tau}(1-b\tau) &= \mathrm{e}\frac{\Theta-1}{\Theta}.\end{aligned}$$

With the notations $x := 1 - b\tau$ and $\mu := \mathrm{e}(\Theta - 1)/\Theta < 0$, we need to examine the solutions x to the equation (21). Since $\Theta \geq \Theta^* = \mathrm{e}^2/(\mathrm{e}^2 + 1)$, we have $\mu \geq -1/\mathrm{e}$ in Remark 5.3. Thus, we have the following three cases for $\tau = (1 - x)/b$.

(i) On $(0, \tau_0)$ we examine the sign of

$$\lim_{\tau\to 0} V'_{\Theta,b}(\tau) = \lim_{\tau\to 0}\big((1-\Theta) + \Theta\mathrm{e}^{-b\tau}(1-b\tau)\big) = 1 - \Theta + \Theta = 1 > 0,$$

so function $V_{\Theta,b}$ is strictly increasing, and its inverse V_1^{-1} exists and is strictly increasing.

(ii) Since $\tau_1 := (1 - x_1)/b = 2/b$ and $x_{-1} < x_1 = -1 \leq x_0$ implies $\tau_0 \leq \tau_1 = \frac{2}{b} < \tau_{-1}$, on the interval (τ_0, τ_{-1}) we examine the sign of

$$V'_{\Theta,b}(\tfrac{2}{b}) = 1 - \mathrm{e}^2 < 0,$$

so function $V_{\Theta,b}$ is strictly decreasing here, and its inverse V_2^{-1} exists and is strictly decreasing.

(iii) On $(\tau_{-1}, +\infty)$ we examine the sign of

$$\lim_{\tau\to+\infty} V'_{\Theta,b}(\tau) = \lim_{\tau\to\infty}\big((1-\Theta) + \Theta\mathrm{e}^{-b\tau}(1-b\tau)\big) = 1 - \Theta + \lim_{x\to-\infty} x\mathrm{e}^x = 1 - \Theta > 0,$$

so function $V_{\Theta,b}$ is strictly increasing, and its inverse V_3^{-1} exists and is strictly increasing.

This proves assertion (a). Since the case $\Theta = 1$ corresponds to the sequential splitting (30), assertion (b) follows from Remark 5.3 and the considerations in the proof of Proposition 6.2. □

We have then the following result for the non-negativity property in this case.

Proposition 7.7. *For $\Theta \in [\Theta^*, 1]$, the non-negativity property (P2) is fulfilled for the method (38) in the following cases:*

(i) *for $\frac{1}{M} \in (0, V_{\Theta,b}(\tau_{-1})]$: if $\tau \leq V_1^{-1}(\frac{1}{M})$,*
(ii) *for $\frac{1}{M} \in (V_{\Theta,b}(\tau_{-1}), V_{\Theta,b}(\tau_0)]$: if $\tau \in \big(0, V_1^{-1}(\frac{1}{M})\big]$ or $\tau \in \big[V_2^{-1}(\frac{1}{M}), V_3^{-1}(\frac{1}{M})\big)$,*
(iii) *for $\frac{1}{M} > V_{\Theta,b}(\tau_0)$: if $\tau < V_3^{-1}(\frac{1}{M})$*

with Notations 5.1 and 7.3.

Proof. The non-negativity of $I^{n+1}_{k,\ell}$ and $R^{n+1}_{k,\ell}$ follows from the non-negativity of $S^{n+1}_{k,\ell}$. So we prove only the latter one. Similarly as in the proof of Proposition 7.5, we need to determine the intervals where

$$V_{\Theta,b}(\tau) := \tau(1 - \Theta(1 - \mathrm{e}^{-b\tau})) \leq \frac{1}{T^n_{k,\ell}}.$$

To do so, we need an inverse of function $V_{\Theta,b}$ which has the three branches presented in Lemma 7.6. Property (P1) implies the estimate $T^n_{k,\ell} \leq M$ which provides the assertions. □

Remark 7.8. As we have already pointed out, the cases $\Theta = 0$ and $\Theta = 1$ correspond to the sequential splitting methods (36) and (30), respectively. Since $V_{1,b}(\tau) = \tau$, its inverse $V_{1,b}^{-1}$ is the identity in (39), cf. Proposition 6.6. Furthermore, $V_{0,b}(\tau) = \tau\mathrm{e}^{-b\tau}$ implies $V_{0,b}^{-1}(y) = -W(-by)/b$ having the two branches $W = W_{-1}$ and $W = W_0$ as in Proposition 6.2. Hence, as expected, the corresponding results in Propositions 7.5 and 7.7 meet the conditions in Propositions 6.2 and 6.6.

Remark 7.9. As in the case of the sequential splitting 1–2 (cf. Remark 6.4), Propositions 7.5 and 7.7 yield sharper conditions of the time step than the bound is when analysing the non-negativity preservation of the sub-problems separately. This is true due to the following consideration. The weighted splitting consists of the two sequential splittings 1–2 (30) and 2–1 (36). In Proposition 6.3 we showed that the bound obtained for sequential splitting 1–2 is sharper than $1/M$. Furthermore, for the sequential splitting 2–1 we have the bound $1/M$. Hence, the separate treatment of the sub-problems leads to the constraint $\tau \leq 1/M$. The bounds obtained in the present section, however, are always sharper. In the case of "small" Θ, the bound $V^{-1}_{\Theta,b}(\frac{1}{M})$ is larger than $1/M$, since $V^{-1}_{\Theta,b}(\tau)$ is a strictly increasing function in τ by Lemma 7.4, and $V^{-1}_{\Theta,b}(\tau) = \tau$ holds for $\Theta = 0$ (so the bound here is simply $1/M$), furthermore, $V^{-1}_{\Theta,b}(\tau)$ is monotonically decreasing as Θ increases. For "large" values of Θ, we refer to Fig. 3 to see that we can have three cases with respect to the location of the graph of function $V_{\Theta,b}(\tau)$ relative to the horizontal line $1/M$. In all cases we have $\partial V_{\Theta,b}(\tau)/\partial\Theta < 0$, that is, the left point of intersection moves to the right when the value of Θ is increasing (the movement might not be continuous but monotone). This means that the left bound is increasing as well.

8. Strang splitting

In contrast to the sequential splittings presented in Sections 6.1 and 6.2, Strang splitting needs three steps with the two sub-problems (Sub.1)–(Sub.2): the first step uses (Sub.1) with time step $\tau/2$, the second uses (Sub.2) with time step τ, and the third uses (Sub.1) again with time step $\tau/2$, always using the previous solution as an initial condition. Moreover, while the sequential splitting is of first order, the Strang is a second-order method. Therefore, by [2], we need to use a second-order time discretisation method to avoid order reduction. Hence, sub-problem (Sub.2) will be solved by Heun's method as presented in (15)–(16).

We note that the choice of (Sub.2) being the middle step is explained by the fact that it needs more computational effort and time than sub-problem (Sub.1). Hence, computing it only once at each step is more efficient than using the approach having (Sub.1) in the middle, since in that case (Sub.2) should be evaluated twice.

The corresponding steps to be solved one after another, have then the following form with given $S^n_{k,\ell}$, $I^n_{k,\ell}$, $R^n_{k,\ell}$ values:

$$\begin{cases} S^{[1],n+1}_{k,\ell} = \mathrm{e}^{-c\frac{\tau}{2}} S^n_{k,\ell}, \\ I^{[1],n+1}_{k,\ell} = \mathrm{e}^{-b\frac{\tau}{2}} I^n_{k,\ell}, \\ R^{[1],n+1}_{k,\ell} = R^n_{k,\ell} + (1-\mathrm{e}^{-c\frac{\tau}{2}})S^n_{k,\ell} + (1-\mathrm{e}^{-b\frac{\tau}{2}})I^n_{k,\ell}, \end{cases} \tag{41}$$

$$\begin{cases} \widehat{S}^{[2],n+1}_{k,\ell} = S^{[1],n+1}_{k,\ell}\big(1-\tau T^{[1],n+1}_{k,\ell}\big), \\ \widehat{I}^{[2],n+1}_{k,\ell} = I^{[1],n+1}_{k,\ell} + \tau S^{[1],n+1}_{k,\ell} T^{[1],n+1}_{k,\ell}, \\ \widehat{R}^{[2],n+1}_{k,\ell} = R^{[1],n+1}_{k,\ell}, \end{cases} \tag{42}$$

$$\begin{cases} S^{[2],n+1}_{k,\ell} = \frac{1}{2} S^{[1],n+1}_{k,\ell} + \frac{1}{2}\widehat{S}^{[2],n+1}_{k,\ell}\big(1-\tau \widehat{T}^{[2],n+1}_{k,\ell}\big), \\ I^{[2],n+1}_{k,\ell} = \frac{1}{2} I^{[1],n+1}_{k,\ell} + \frac{1}{2}\big(\widehat{I}^{[2],n+1}_{k,\ell} + \tau \widehat{S}^{[2],n+1}_{k,\ell}\widehat{T}^{[2],n+1}_{k,\ell}\big), \\ R^{[2],n+1}_{k,\ell} = R^{[1],n+1}_{k,\ell}, \end{cases} \tag{43}$$

$$\begin{cases} S^{n+1}_{k,\ell} = \mathrm{e}^{-c\frac{\tau}{2}} S^{[2],n+1}_{k,\ell}, \\ I^{n+1}_{k,\ell} = \mathrm{e}^{-b\frac{\tau}{2}} I^{[2],n+1}_{k,\ell}, \\ R^{n+1}_{k,\ell} = R^{[2],n+1}_{k,\ell} + (1-\mathrm{e}^{-c\frac{\tau}{2}})S^{[2],n+1}_{k,\ell} + (1-\mathrm{e}^{-b\frac{\tau}{2}})I^{[2],n+1}_{k,\ell} \end{cases} \tag{44}$$

for all $n \in \mathbb{N}$ and $(x_k, y_\ell) \in \mathcal{G}$. Since the form of the combined method would be too complex, we leave the steps individually written, and will study them separately. In this case we get constraints which are more transparent and easier to verify. As before, we are going to show the validity of the properties (P1)–(P4).

Proposition 8.1. *Proposition 6.1 holds for the method* (41)–(44).

Proof. It suffices to show the assertions for each steps (41)–(44), by taking into account that they are constitutive steps, that is, their solution serves as an initial value for the next step.

The conservation of the size of the total population can be shown by adding up the equations in each step (41)–(44). Since it is conserved in each step, it remains the same for the whole method as well.

By assuming the non-negativity property (P2) and using $\mathrm{e}^{-b\frac{\tau}{2}}, \mathrm{e}^{-c\frac{\tau}{2}} \in (0,1)$, we have that $S^{[1],n+1}_{k,\ell} \le S^n_{k,\ell}$ and $R^{[1],n+1}_{k,\ell} \ge R^n_{k,\ell}$ in (41). Moreover, we have $T^{[1],n+1}_{k,\ell} \ge 0$ which, together with property (P2) for (42), implies $\widehat{S}^{[2],n+1}_{k,\ell} \le S^{[1],n+1}_{k,\ell}$ and $\widehat{R}^{[2],n+1}_{k,\ell} \ge R^{[1],n+1}_{k,\ell}$ in (42). Again, the non-negativity of $I^{[2],n+1}_{k,\ell}$ implies $T^{[2],n+1}_{k,\ell}$, therefore, property (P2) holds for step (43), too. Since $\mathrm{e}^{-b\frac{\tau}{2}}, \mathrm{e}^{-c\frac{\tau}{2}} \in (0,1)$, properties (P3) and (P4) follow for step (44) as well. □

Hence, as before, it suffices to analyse the conditions under which the non-negativity holds. The only difference from the previous sections is that in this case we will perform the analysis separately for each step (41)–(44). We will take into account, however, that they are constitutive steps of the method.

Proposition 8.2. *The non-negativity property* (P2) *holds for the method* (41)–(44) *if*

(i) $\tau \in \mathbb{T}_{c/2,M}$ *for* $2M < b\mathrm{e}$,
(ii) $\tau \in \mathbb{T}_{p,M}$ *with* $p = \min\{b, c\}/2$ *for* $2M \ge b\mathrm{e}$.

Proof. It suffices to show the non-negativity property (P2) step by step for (41)–(44).

Step (41)*:* By assuming $S^n_{k,\ell}, I^n_{k,\ell}, R^n_{k,\ell} \ge 0$, we immediately have $T^n_{k,\ell} \ge 0$. Therefore,

$$S^{[1],n+1}_{k,\ell}, I^{[1],n+1}_{k,\ell}, R^{[1],n+1}_{k,\ell} \ge 0$$

holds in (41).

Step (42)*:* The previous step and the relation

$$T^{[1],n+1}_{k,\ell} = \mathcal{M}(I^{[1],n+1}_{k,\ell}) = \mathcal{M}(\mathrm{e}^{-b\frac{\tau}{2}} I^n_{k,\ell}) = \mathrm{e}^{-b\frac{\tau}{2}} \mathcal{M}(I^n_{k,\ell}) = \mathrm{e}^{-b\frac{\tau}{2}} T^n_{k,\ell} \tag{45}$$

implies $T^{[1],n+1}_{k,\ell} \ge 0$, too. Therefore, $\widehat{I}^{[2],n+1}_{k,\ell}, \widehat{R}^{[2],n+1}_{k,\ell} \ge 0$ are satisfied in (42). However, the non-negativity of $\widehat{S}^{[2],n+1}_{k,\ell}$ only holds if

$$\begin{aligned} 1 - \tau T^{[1],n+1}_{k,\ell} &\ge 0 \\ 1 - \tau \mathrm{e}^{-b\frac{\tau}{2}} T^n_{k,\ell} &\ge 0 \\ -\tau \mathrm{e}^{-b\frac{\tau}{2}} &\ge -\tfrac{1}{T^n_{k,\ell}} \end{aligned}$$

where we used relation (45). Property (P1) implies $T^n_{k,\ell} \le M$ for all $n \in \mathbb{N}$ and $(x_k, y_\ell) \in \mathcal{G}$. Hence, we obtain the necessary condition

$$-\frac{b}{2}\tau \mathrm{e}^{-b\frac{\tau}{2}} \ge -\frac{b}{2M}. \tag{46}$$

With the notation $x := -b\tau/2 < 0$ and $\mu := -b/2M < 0$, we have to analyse the inequality (22). By Remark 5.3 and the proof of Proposition 6.2, we have the following cases:
(a) For $2M < b\mathrm{e}$, inequality (46) holds for all $\tau > 0$.
(b) For $2M \ge b\mathrm{e}$, inequality (46) holds if $\tau \in \mathbb{T}_{b/2,M}$.

Step (43)*:* The non-negativity of $I^{[2],n+1}_{k,\ell}$ and $R^{[2],n+1}_{k,\ell}$ follows immediately, however, $S^{[2],n+1}_{k,\ell} \ge 0$ holds in (43) only if

$$1 - \tau \widehat{T}^{[2],n+1}_{k,\ell} \ge 0 \tag{47}$$

is satisfied. We observe that

$$\begin{aligned} \widehat{T}^{[2],n+1}_{k,\ell} &= \mathcal{M}(\widehat{I}^{[2],n+1}_{k,\ell}) = \mathcal{M}\big(I^{[1],n+1}_{k,\ell} + \tau S^{[1],n+1}_{k,\ell} T^{[1],n+1}_{k,\ell}\big) \\ &= \mathcal{M}\big(\mathrm{e}^{-b\frac{\tau}{2}} I^n_{k,\ell} + \tau \mathrm{e}^{-c\frac{\tau}{2}} S^n_{k,\ell} \mathrm{e}^{-b\frac{\tau}{2}} T^n_{k,\ell}\big) = \mathrm{e}^{-b\frac{\tau}{2}} \mathcal{M}\big(\underbrace{I^n_{k,\ell} + \tau \mathrm{e}^{-c\frac{\tau}{2}} S^n_{k,\ell} T^n_{k,\ell}}_{I^*_{k,\ell}}\big). \end{aligned} \tag{48}$$

The value $I^*_{k,\ell}$ corresponds to a sequential splitting step (30) with the choice $b = c/2$. Hence, we have the following observations.

(i) Proposition 6.1 implies that the total size of the population is conserved, therefore, $I^*_{k,\ell} \le N_{k,\ell}$, so we have the estimate $\mathcal{M}(I^*_{k,\ell}) \le M$ for all $(x_k, y_\ell) \in \mathcal{G}$.
(ii) According to Proposition 6.2, the non-negativity of $I^*_{k,\ell}$ is guaranteed for $\tau \in \mathbb{T}_{c/2,M}$.

By taking into account (48) and (i), the inequality (47) holds if $1 - \tau \mathrm{e}^{-b\frac{\tau}{2}} M \ge 0$. Remark 5.3 implies then the condition $\tau \in \mathbb{T}_{b/2,M}$. Together with (ii) we have the condition $\tau \in \mathbb{T}_{b/2,M} \cap \mathbb{T}_{c/2,M}$.

Step (44)*:* Since all additive terms are non-negative, property (P2) is satisfied for all values of $\tau > 0$.

The strict increase of W_0 and decreasing of W_{-1} imply the relations

$$\begin{aligned} -\tfrac{2}{b} W_0(-\tfrac{b}{2M}) &> -\tfrac{2}{c} W_0(-\tfrac{c}{2M}), \\ -\tfrac{2}{b} W_{-1}(-\tfrac{b}{2M}) &< -\tfrac{2}{c} W_{-1}(-\tfrac{c}{2M}) \end{aligned}$$

for $b > c$, and conversely for $b < c$. Since the condition $\tau \in \mathbb{T}_{c/2,M}$ is necessary in both cases $2M < b\mathrm{e}$ and $2M \ge b\mathrm{e}$, and $\tau \in \mathbb{T}_{b/2,M}$ is needed only for $2M > b\mathrm{e}$, we proved the assertions. □

We remark that if the effect of the vaccination is not taken into account ($c = 0$), we have the condition $\tau < 1/M$, according to Lemma 5.2 with $q = M$. This means that in this case we cannot guarantee a better sufficient condition on the time step than the one without applying operator splitting procedure, cf. (23).

9. Numerical experiments

The present section is devoted to the numerical illustration of our previously obtained theoretical results regarding (i) the preservation of the total density, (ii) the non-negativity of S, I, R, and (iii) the monotonicity of S, R.

There are issues already mentioned earlier which become really important at this point. Since the rectangular domain Ω is bounded, a special attention should be given to the boundary. As pointed out in Section 3.2, we assume that there is no susceptible population outside Ω, thus, we assign zero values there. Using either a uniform or a non-uniform cubature, the cubature points usually do not belong to the spatial grid $\mathcal{G}$. To implement the cubature points at the boundary and in the corners as well, we define ghost cells outside the domain Ω having zero values. This enables the correct calculation of the values which correspond to the cubature points lying outside the domain.

For the numerical experiments, we choose the following functions in (5):

$$g_1(r) = a(-r + \delta),$$
$$g_2(\vartheta) = \beta \sin(\vartheta + \alpha) + \beta,$$

where $a > 0$ is the infection rate. We use the parameter values $\alpha = 0$ and $\beta = 1$ describing a northern wind on the domain. In our numerical experiments we take $a = 100$, $b = 0.1$, and $\delta = 0.05$.

As mentioned before, we can use different quadratures to approximate the integrals in (5). First, we transform the disk-like infectious domain with radius δ to the rectangle $[0, \delta] \times [0, 2\pi)$ in the (r, ϑ) plane. Next, we transform this rectangle to the $[0, 1] \times [0, 1)$ square on the (ξ, η) plane by using the linear transformation $r = \delta\xi$ and $\vartheta = 2\pi\eta$ with Jacobian equals $2\pi\delta$. Using the transformations above, the integral in (5) has the form

$$\int_0^1 \int_0^1 f_{(x,y)}\big(\delta\xi \cos(2\pi\eta), \delta\xi \sin(2\pi\eta)\big)\delta\xi \; 2\pi\delta \, \mathrm{d}\eta \, \mathrm{d}\xi$$

with the notation

$$f_{(x,y)}(\bar{x}, \bar{y}) := g_1(r) g_2(\vartheta) I(t, x + r\cos(\vartheta), y + r\sin(\vartheta)) r,$$

where $r = \sqrt{(x - \bar{x})^2 + (y - \bar{y})^2}$ and $\vartheta = \arctan(\frac{y - \bar{y}}{x - \bar{x}})$. For the integration over the interior of the aforementioned square, we take the generalised Gaussian quadrature rules described in [10]. For $N_w \in \mathbb{N}$, we choose weights w_i, $i = 1, \dots, N_w$, and denote the position of the ith point in the one-dimensional Gaussian quadrature by (ξ_i, η_i). The quadrature has then the form

$$Q(f) = \sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j 2\pi\delta^2 \xi_i f_{(x,y)}\big(\delta\xi_i \cos(2\pi\eta_j), \delta\xi_i \sin(2\pi\eta_j)\big) = \sum_{m=1}^{N^2} \widetilde{w}_m f_{(x,y)}(x_m, y_m)$$

with $x_m = \delta\xi_i \cos(2\pi\eta_j)$, $y_m = \delta\xi_i \sin(2\pi\eta_j)$, and $\widetilde{w}_m = w_i w_j 2\pi\delta^2\xi_i$.

As mentioned before, it is also possible to use symmetric, uniform quadratures on the disc. For further details, see [13].

Regarding the initial conditions, we assume that there are no recovered individuals at the beginning, that is, $R^0_{k,\ell} = 0$ for all $(x_k, y_\ell) \in \mathcal{G}$. For the infected individuals, we use a Gaussian distribution concentrated at the middle of the domain $(A/2, B/2)$ which has a standard deviation $s = \min(A, B)/10$:

$$I^0_{k,\ell} = \frac{1}{2\pi s^2} \exp\left(-\frac{1}{2}\left[\left(\frac{h_x(k-1) - \frac{A}{2}}{s} \right)^2 + \left(\frac{h_y(\ell-1) - \frac{B}{2}}{s} \right)^2 \right] \right),$$

where $A = (K - 1)h_x$ and $B = (L - 1)h_y$ as introduced in Section 3.2. We set here $A = B = 1$. Due to property (P1), the sum $N_{k,\ell}$ of all individuals is constant in time at each point $(x_k, y_\ell) \in \mathcal{G}$. Thus, the initial distribution of the susceptibles is $S^0_{k,\ell} = N_{k,\ell} - I^0_{k,\ell}$. For our tests, we choose $N_{k,\ell} = 20$ for all $(x_k, y_\ell) \in \mathcal{G}$.

In Fig. 4 the numerical solution is plotted for different time levels ($S_{k,\ell}$ is plotted in the left column, $I_{k,\ell}$ in the middle, and $R_{k,\ell}$ on the right). One can see that the number of susceptibles decrease, and the number of infected moves towards the boundaries forming a wave. Both of them tend to the zero function, while the number $R_{k,\ell}$ of recovered tends to $N_{k,\ell} = 20$ at each grid points $(x_k, y_\ell) \in \mathcal{G}$.

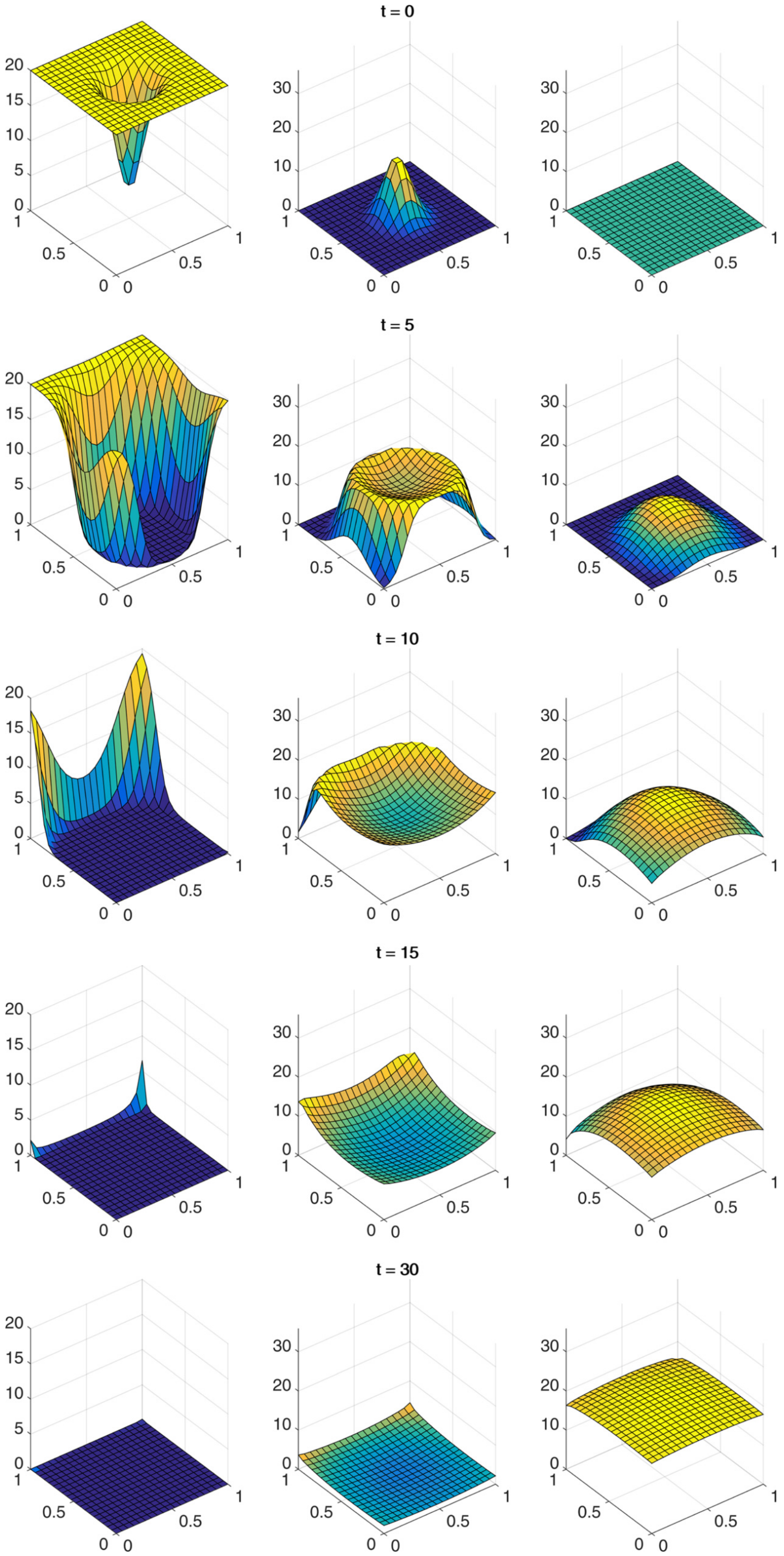

Fig. 4. The numerical solutions $S_{k,\ell}^n$, $I_{k,\ell}^n$, $R_{k,\ell}^n$ shown in columns, respectively, at time levels $t = 0$, $t = 5$, $t = 10$, $t = 15$, $t = 30$, for the sequential splitting (30).

Table 1
Numerical results for sequential splitting 1–2 (30) for various time steps τ. The deficiency is computed at final time $t = 50$ for the upper, and at final time $t = 400$ for the lower table.

Time step τ	0.48	0.50	0.52	0.54	0.56	0.58	0.60
Ratio $\tau/\widehat{\tau}$	0.95	0.99	1.03	1.07	1.11	1.15	1.19
Property (P2)	yes	yes	yes	yes	yes	no	no
Deficiency	0	0	0	0	0	4.17e-3	2.42e-2

Time step τ	37	40	43	46
Ratio $\tau/\widetilde{\tau}$	0.81	0.88	0.94	1.01
Property (P2)	no	no	yes	yes
Deficiency	9.98e-1	8.91e-2	0	0

Table 2
Numerical results for sequential splitting 2–1 (36) for various time steps τ. The deficiency is computed at final time $t = 50$.

Time step τ	0.47	0.49	0.51	0.53	0.55	0.57	0.59
Ratio $\tau/\widehat{\tau}_{21}$	0.98	1.03	1.07	1.11	1.15	1.19	1.24
Property (P2)	yes	yes	yes	yes	no	no	no
Deficiency	0	0	0	0	1.75e-3	5.01e-3	3.99e-2

9.1. *Testing the time step bounds*

The natural question arises how strict the time step bounds derived in Section 4 are. Since Proposition 6.1 holds for all schemes presented, it suffices to analyse the constraints on the non-negativity preservation. For numerical examples which use explicit Euler method and violate the non-negativity preservation, we refer to Fig. 2 in [14] and Fig. 4 in [13]. We note that the aforementioned choice of the parameters yields $M \approx 2.0893$.

Sequential splitting 1–2 (30) Due to Proposition 6.2, the sufficient upper bound (the lower end of the forbidden interval) is

$$\widehat{\tau} = -\tfrac{1}{b}W_0(-\tfrac{b}{M}) = -\tfrac{1}{0.1}W_0(-\tfrac{0.1}{2.0893}) \approx 0.5033, \tag{49}$$

while the sufficient lower bound (the upper end of the forbidden interval) is

$$\widetilde{\tau} = -\tfrac{1}{b}W_1(-\tfrac{b}{M}) = -\tfrac{1}{0.1}W_1(-\tfrac{0.1}{2.0893}) \approx 45.5583.$$

In Table 1 we present our results on the time steps where the non-negativity property (P2) is preserved by the sequential splitting (30). In the second row we indicate the ratios $\tau/\widehat{\tau}$ (for small τ) and $\tau/\widetilde{\tau}$ (for large τ). The deficiency means the maximum of the absolute values of the negative values appeared in the solution at the final time level.

One can see that the necessary bound $\widehat{\tau}$ is relatively close to the numerically obtained "exact" bound. Moreover, there appear certain errors when the time step is further increased, i.e., the solution becomes negative. It is also evident that after increasing the time-step close enough to the other bound $\widetilde{\tau}$, the non-negativity property is satisfied again.

Sequential splitting 2–1 (36) In Proposition 6.6 we have the bound

$$\widehat{\tau}_{21} = \frac{1}{M} \approx \frac{1}{2.0893} \approx 0.4763. \tag{50}$$

Table 2 shows whether the non-negativity (P2) is preserved. The numerical experiments show that the behaviour of this method is similar to the previous one, although it produces slightly bigger errors. Also, it does not become stable for any bigger values of τ, as expected from Proposition 6.6.

Weighted sequential splitting (38) We study first the behaviour of the method for $\Theta = 0.5 < \Theta^*$. Proposition 7.5 leads to the bound

$$\widehat{\tau}_{w1} = V^{-1}_{\Theta,b}\left(\frac{1}{M}\right) \approx V^{-1}_{0.5,0.1}\left(\frac{1}{2.0893}\right) \approx 0.4809, \tag{51}$$

which is between the two previously obtained values (49) and (50). The corresponding errors are also between the errors of the two previous methods, which can be seen in Table 3.

We study next the case $\Theta = 0.9 > \Theta^*$. Then we get the bounds from Proposition 7.7 as

$$\begin{aligned} \tau_{-1} &= \tfrac{1}{b}\big(1 - W_{-1}\big(\tfrac{\mathrm{e}(\Theta-1)}{\Theta}\big)\big) = \tfrac{1}{0.1}\big(1 - W_{-1}\big(\tfrac{\mathrm{e}(0.9-1)}{0.9}\big)\big) \approx 27.6587, \\ \tau_{0} &= \tfrac{1}{b}\big(1 - W_{0}\big(\tfrac{\mathrm{e}(\Theta-1)}{\Theta}\big)\big) = \tfrac{1}{0.1}\big(1 - W_{0}\big(\tfrac{\mathrm{e}(0.9-1)}{0.9}\big)\big) \approx 14.9596. \end{aligned}$$

Table 3
Numerical results for the weighted sequential splitting (38) with $\Theta = 0.5$ for various time steps τ. The deficiency is computed at final time $t = 50$.

Time step τ	0.47	0.49	0.51	0.53	0.55	0.57	0.59
Ratio $\tau/\widehat{\tau}_{w1}$	0.98	1.02	1.06	1.10	1.14	1.18	1.23
Property (P2)	yes	yes	yes	yes	no	no	no
Deficiency	0	0	0	0	9.37e-4	5.00e-3	3.70e-2

Table 4
Numerical results for the sequential weighted splitting (38) with $\Theta = 0.9$ for various time steps τ. The deficiency is computed at final time $t = 50$.

Time step τ	0.47	0.49	0.51	0.53	0.55	0.57	0.59
Ratio $\tau/\widehat{\tau}_{w2}$	0.93	0.98	1.02	1.06	1.10	1.14	1.18
Property (P2)	yes	yes	yes	yes	yes	no	no
Deficiency	0	0	0	0	0	1.46e-3	5.63e-3

Table 5
Numerical results for the Strang splitting (41)–(44) for various time steps τ. The deficiency is computed at final time $t = 50$.

Time step τ	0.47	0.49	0.51	0.53	0.55	0.57	0.59
Ratio $\tau/\widehat{\tau}_S$	0.98	1.02	1.06	1.10	1.15	1.19	1.23
Property (P2)	yes	yes	yes	yes	no	no	no
Deficiency	0	0	0	0	5.39e-4	7.30e-3	2.71e-2

Since we have $1/M \approx 0.4763$, being smaller than both of the above values, we have case (i) in Proposition 7.7. Therefore, we need to compute the following bound:

$$\widehat{\tau}_{w2} = V_1^{-1}(\tfrac{1}{M}) \approx V_1^{-1}(\tfrac{1}{2.0893}) \approx 0.5006,$$

which is closer to the bound (49) than to (50). The corresponding results are listed in Table 4.

Strang splitting (41)–(44) By the choice of parameters, we have

$$2M = 4.1786 > 0.2718 = be.$$

Hence, we consider case (ii) of Proposition 8.2. Moreover, relation $c = 0.01 < 0.1 = b$ leads to the bounds

$$\widehat{\tau}_S = -\tfrac{2}{c} W_0(-\tfrac{c}{2M}) \approx -\tfrac{2}{0.01} W_0(-\tfrac{0.01}{4.1786}) \approx 0.4798, \tag{52}$$

$$\widetilde{\tau}_S = -\tfrac{2}{c} W_1(-\tfrac{c}{2M}) \approx -\tfrac{2}{0.01} W_1(-\tfrac{0.01}{4.1786}) \approx 1626. \tag{53}$$

As we can see, bound (52) is similar to the previously observed bounds (49), (50), and (51). Due to our choice of parameters, any recognizable dynamics of S, I, R is already over before time level $t = 1626$. Therefore, $\widetilde{\tau}_S$ in (53) is far too large to be considered as a suitable time step. Hence, we omit the numerical experiments using it. The numerical results are shown in Table 5.

9.2. Accuracy analysis

Besides the preservation of the qualitative properties, we also studied the accuracy of the presented methods. Since the exact solution to system (6) is not known, we considered a reference solution instead which was computed with a small time step. The time step was first chosen to be the bound acquired in the previous sections, and then by halving it six times, we got seven solutions. The last one was chosen to be the reference solution. We had a spatial mesh of 20 × 20 points and a bilinear interpolation with a 5 × 5 quadrature, and the parameters $a = 100$, $b = 0.1$, $c = 0.01$, and $\delta = 0.1$. We chose the final time $T = 20$. We define the relative global error at time level $T = N\tau$ as

$$\varepsilon(\tau) := \frac{\|\underline{X}^N - X^N\|}{\|\underline{X}^N\|}$$

with $X \in \{S, I, R\}$, where X^N means the matrix with elements $X^N_{k,\ell}$ for $(x_k, y_\ell) \in \mathcal{G}$, and the underlying refers to the reference solution. We took the relative global error by using the maximum norm and the discrete L^1 and L^2 norms.

In Fig. 5 the order plot can be seen for the relative errors $\varepsilon(\tau)$ of the four presented splitting schemes, where the colours represent the methods and the line styles the various norm types: the solid line states for the maximum norm, the dashed

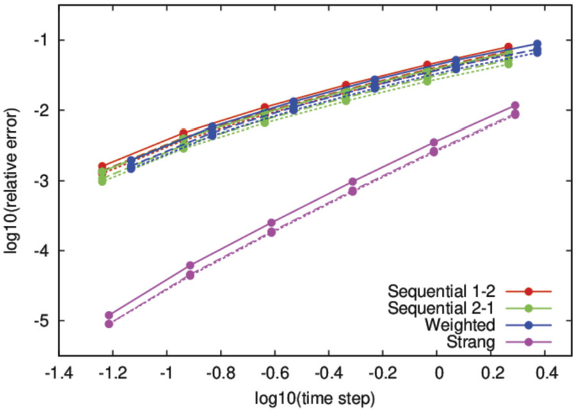

Fig. 5. Order plot for the relative global error of the four splitting schemes identified with colours. Solid line states for the maximum norm, dashed line for the discrete L^1 norm, and dotted line for the discrete L^2 norm. The lines separated from the others correspond to the Strang splitting. (For interpretation of the colours in the figure, the reader is referred to the web version of this article.)

Table 6
Slopes of the lines fit to the curves in Fig. 5 which approximate the orders of the methods in the three norms considered.

Splitting	Order in max. norm	Order in L^1 norm	Order in L^2 norm
Sequential 1–2 (30)	1.11046	1.11216	1.11263
Sequential 2–1 (36)	1.10598	1.10145	1.09238
Weighted (38) with $\Theta = 0.9$	1.04794	1.04965	1.05635
Strang (41)–(44)	1.97442	1.98131	1.97148

line for the discrete L^1 norm, and the dotted line for the discrete L^2 norm. Since the slope of the curve on the log-log plot corresponds to the order of the method, one can see that the sequential and the weighted splittings are of first order, while the Strang splitting is of second order convergent. We fit a line to the data and obtained the values given in Table 6. The fairly good approximations to the theoretical orders can be clearly read from the data.

10. Conclusion

Application of operator splitting leads to sub-problems being easier to solve or possessing advantageous numerical properties. In the case of the space-dependent epidemic SIR model with vaccination, the use of operator splitting resulted in numerical methods which preserve the total size of the population. Furthermore, they yield non-negative population densities and proper monotonicity properties under some requirements on the method's time step.

We showed that in case of "rapid" recovery (i.e., $b > M/\mathrm{e}$ holds for fixed initial values) the sequential splitting 1–2 needs no restriction on the time step to yield non-negative population densities. Hence, it behaves qualitatively better than the method which does not use operator splitting. Moreover, sequential splitting requires time step from a broader interval as the method without splitting also for "slow" recovery ($b \le M/\mathrm{e}$). The same behaviour was observed in the case of weighted and Strang splittings, too. Namely, we obtained a larger upper bound for the time step than the reference one.

With the help of the numerical experiments we could illustrate how sharp the necessary conditions on the time step were. We could see that in all cases the difference in the ratio of the "exact" and necessary bound was about 15%, and, as expected, it decayed as the recovery rate b decreased (this is an immediate consequence of Lemma 5.2). Moreover, numerical experiments show that each method possesses the theoretical convergence order.

Acknowledgement

The work was supported by the National Research, Development and Innovation Office under the grants PD–121117 and SNN–125119. P.Cs. also acknowledges the Bolyai János Research Scholarship of the Hungarian Academy of Sciences. The authors thank the anonymous referees and the Assoc. Editor for their valuable suggestions and the careful reading of the manuscript.

References

[1] K.A. Bagrinovskii, S.K. Godunov, Difference schemes for multidimensional problems, Dokl. Akad. Nauk USSR 115 (1957) 431–433.
[2] P. Csomós, I. Faragó, Error analysis of the numerical solution of split differential equations, Math. Comput. Model. 48 (2008) 1090–1106.
[3] P. Csomós, I. Faragó, Á. Havasi, Weighted sequential splittings and their analysis, Comput. Math. Appl. 50 (2005) 1017–1031.

[4] S. Descombes, M. Duarte, M. Massot, Operator splitting methods with error estimator and adaptive time-stepping. Application to the simulation of combustion phenomena, in: R. Glowinski, S. Osher, W. Yin (Eds.), Splitting Methods in Communication, Imaging, Science, and Engineering, Springer International Publishing, 2015, pp. 1–13.

[5] I. Faragó, R. Horváth, On some qualitatively adequate discrete space-time models of epidemic propagation, J. Comput. Appl. Math. 293 (2016) 45–54.

[6] I. Faragó, R. Horváth, Qualitative properties of some discrete models of disease propagation, J. Comput. Appl. Math. 340 (2018) 486–500.

[7] S. Gottlieb, C-W. Shu, Total variation diminishing Runge–Kutta schemes, Math. Comput. 67 (1998) 73–85.

[8] D.G. Kendall, Mathematical models of the spread of infection, in: Mathematics and Computer Science in Biology and Medicine, H.M.S.O, London, 1965, pp. 213–225.

[9] W.O. Kermack, A.G. McKendrick, A contribution to the mathematical theory of epidemics, Proc. R. Soc. A, Math. Phys. Eng. Sci. 115 (1927) 235–240.

[10] J. Ma, V. Rokhlin, S. Wandzura, Generalized Gaussian quadrature rules for systems of arbitrary functions, SIAM Numer. Anal. 33 (1996) 971–996.

[11] G.I. Marchuk, Some application of splitting-up methods to the solution of mathematical physics problems, Appl. Math. 12 (1968) 103–132.

[12] G. Strang, On the construction and comparison of difference schemes, SIAM J. Numer. Anal. 5 (1968) 506–517.

[13] B. Takács, Y. Hadjimichael, High order discretisation methods for spatial dependent SIR models, https://arxiv.org/abs/1909.01330.

[14] B. Takács, R. Horváth, I. Faragó, Space dependent models for studying the spread of some diseases, Comput. Math. Appl. 80 (2) (2020) 395–404.