Eötvös Loránd Tudományegyetem

Természettudományi Kar

Járványterjedési folyamatok modellezése differenciálegyenletekkel

MSc Diplomamunka

Témavezető:
Faragó István
Alkalmazott Analízis
és Számításmatematikai Tanszék

Szerző:
Szemenyei Adrián László
Alkalmazott matematikus MSc

Budapest, 2023

Köszönetnyilvánítás

Köszönetet szeretnék mondani témavezetőmnek a segítségnyújtásért és iránymutatásért, családomnak a biztatásért.

Contents

Abbreviations

LMM	Linear Multistep method
RK (method)	Runge-Kutta (method)
RK4	A specific Runge-Kutta method
SSP	Strong stability preserving
ODE	Ordinary differential equation
IVP	Initial value problem
PDE	Partial differential equation
PDS	Production-destruction system
MPE (method)	Modified-Patankar Euler (method)
MPRK (method)	Modified-Patankar Runge-Kutta (method)
MPRK22(α)	A family of MPRK methods with parameter α
DFE	Disease-free equilibrium
EE	Endemic Equilibrium

Symbol index

C	SSP coefficient
Δt	step-size of arbitrary numerical scheme
Δt_{FE}	largest step-size for which the SSP property hold for the explicit Euler scheme
R_0	Basic reproduction number
E_0	Disease-free equilibrium
E	Endemic equilibrium
$\mathbb{R}^d_+$	$f(z_1,\dots,z_d) \mid z_i \quad 0, \forall i = 1,\dots,d\}$
$\mathbb{R}^d_{+,0}$	$f(z_1,\dots,z_d) \mid z_i > 0, \forall i = 1,\dots,d\}$
$v^\intercal$	Transpose of a vector v
$C^1(U,V)$	Continuously differentiable functions from U to V

0 Abstract

Just as disease propagation in the real world, compartmental epidemiological models show great diversity in their dynamics. Through the construction of epidemiological models, we can better understand - and possibly predict - the dynamics and qualitative properties of different infectious diseases. In general, one builds an ODE system based on different assumptions from the biological and propagation properties of the disease and studies the qualitative properties of the constructed system. In all of the above cases, the equations cannot be solved analytically (although in some cases special solutions can be found), therefore numerical methods are used when one seeks their approximate solutions (e.g. for forecasting). It is then of interest whether the discretised system possesses the same qualitative behaviour as the continuous model. Since all compartmental systems model the rate of change of different subpopulations or the density of the disease in the environment, a good model should have the property that the solution remains non-negative for all non-negative initial values. Such systems are called positive. It is therefore of interest whether a numerical scheme preserves this property and, if so, for which step sizes. The number of equilibria (and periodic orbits) and their stability properties, the positive invariance of a set, etc. are some other properties that should be preserved by the discretisation.

Considering the classical linear methods, namely the Runge-Kutta and linear multistep methods, it is known that there is no second or higher order method which preserves the positivity for all step-sizes and for all positive ODE systems. Thus, it is of interest to know the largest such step-size for which positivity is preserved for a given method. We introduce the general theory of the above, which is based on the so-called strong stability preservation, namely if we know that the explicit Euler method preserves positivity for step-sizes $0 < \Delta t \quad \Delta t_{FE}$ then the positivity is also preserved for other linear methods under step-sizes $0 < \Delta t \quad C\Delta t_{FE}$ where C depends only on the scheme considered and not on the considered system, and possibly $C = 0$. We introduce the different representations of the Runge-Kutta methods which can be used to find the C coefficients and summarise the known results considering the order barriers, the conditions on the Butcher tableau and possible extensions. We also do this for the linear multistep methods.

To study the sharpness of the C coefficients considering epidemiological compartmental ODE models, we introduce a simple vertical transmission model, which is conservative in the sense that the total population is constant, and a more complex model which is considered to model the COVID-19 disease. For the latter, the environment is also considered as a possible transmission route (i.e. a susceptible individual may acquire the disease through the environment and not directly from susceptible-infectious contacts) and includes a vaccinated subpopulation with imperfect vaccination, which means that vaccinated individuals can also become infected but with different transmission rate. We first study the qualitative properties of these continuous models and show that the second model exhibits backward bifurcation, i.e. for some parameter values a stable disease-free equilibrium coexists with a stable and unstable endemic equilibria. From the conditions, under whitch this occurs, it can be deduced that the existence of such backward bifurcation is independent of the parameters which determine the dynamics of the environment and its disease propagation. For both models, we show that there is a positively invariant region, called the biologically feasible region in the positive quadrant which also implies the positivity of the systems.

Having studied the behaviour of the continuous models, we are ready to investigate how some of its basic qualitative properties change under different discretisations. Since the strong stability preserving methods are based on the qualitative behaviour of the explicit Euler discretisation, we give sufficient conditions for the preservation of the positively invariant region for the non-conservative

model. Considering the conservative model, we give sufficient and necessary conditions considering its positivity and the linear stability of the different equilibria.

We construct numerical simulations to study the sharpness of the C coefficients considering the maximal step-size for which different discretisations preserve positivity (and boundedness for the non-conservative model). We find that the numerical results differ significantly from the theoretical results based on the strong stability preserving theory. Most notably, while the classical RK4 method has $C = 0$, for our specific models, the maximal step-sizes were 1.5 2.5 times larger than for the explicit Euler method. Fortunately, for the conservative system it is also possible in some cases to determine the initial values that lead to the smallest such step sizes for which positivity is lost. These results can be partially explained by letting the internal stages to be negative, which allows the step sizes to be larger due to the special structure of the vector field of the phase-space outside but near to the positive quadrant.

We also introduce a family of (nonlinear) unconditionally positive and conservative schemes, called modified-Patankar-Runge-Kutta schemes and summarise some of the recent developments considering these schemes. While we do not prove global stability under the discretisation for arbitrary step-sizes, the bifurcation diagrams imply it.

In section 1 we introduce the basics of epidemiological modelling and in section 2 we introduce the general theory of autonomous ordinary differential equations and some of the qualitative properties of the solutions, that we will consider later. In section 3, we introduce the Runge-Kutta and linear multistep schemes and theory of their positivity preservation and also some other properties that have been studied in the literature. We also introduce the modified-Patankar-Runge-Kutta schemes, which are unconditionally positive for a class of systems. In section 4 we analyse two continuous epidemiological models and in 5 we study how the systems under different schemes preserve the various properties introduced in 3 and studied for the continuous models in section 4 through partly numerical experiments.

1 Epidemiological modelling

Epidemiological models can be categorized by their mathematical structures: deterministic or stochastic. In deterministic models, one of the most used are the compartmental models, where the dynamics of different compartments are modelled by ordinary differential equations. Different compartments make it possible to 'heterogenise' the population by its relationship to the disease, age, space, vaccination or lack of it, etc. One of the first such model was constructed by A. G. McKendrick and W. O. Kermack in 1927[1]. We will summarise the model in its time-since-infection independent case, to show how one can build different compartmental models - based on different assumptions. In the Kermack-Mckendrick epidemic model, the population is subdivided into subpopulations/classes, namely, the susceptible, infected and recovered classes. Where a person is considered to be in the susceptible class if he or she is not infected but can get infected by contact with an infectious person. In the following model, the infected subpopulation consists of the infected and infectious people, while the recovered subpopulation consists of people who are neither infectious nor susceptible, i.e. can't get infected. The number (or density) of each subpopulation at time t 0 are denoted by $S(t)$, $I(t)$, $R(t)$, respectively.

The dynamics evolve in time based on the following equations:

$$\begin{aligned} \frac{\mathrm{d}S}{\mathrm{d}t} &= \quad \beta IS \\ \frac{\mathrm{d}I}{\mathrm{d}t} &= \beta IS \quad \alpha I \\ \frac{\mathrm{d}R}{\mathrm{d}t} &= \alpha I \end{aligned} \tag{1.1}$$

with non-negative initial conditions $S(0), I(0), R(0)$ and positive real parameters α and β. One can see that the model has the assumption that the total population $N(t) := S(t) + I(t) + R(t)$ remains constant (since $\frac{\mathrm{d}N(t)}{\mathrm{d}t} = 0$). The rate of the change in the susceptible population is given by the number of susceptible individuals infected in a unit of time. If we assume that the contact rate c is proportional to the total population (i.e. cN), then an infected individual makes $cN\frac{S}{N}$ contacts with the susceptible subpopulation per unit of time, of which only a fraction, say p *2* $(0, 1)$ result in disease transmission. Thus, from the entire infected subpopulation $pcSI$ number of individuals become infected in a unit of time. So in the model (1.1) $\beta = pc$ with the unit $\frac{1}{\text{number of people unit of time}}$ and it is called the *transmission rate constant*. The number of people in the infected subpopulation changes by the newly infected and by people who recover from the virus by a constant recovery rate α which has the unit $\frac{1}{\text{unit of time}}$ The infected individuals, who recover leave the infected class and move to the recovered class without any time delay.

In general, most compartmental models are similar to the above model, and one can look at it as its specific modification for the given infection. The most common modifications are

1. Other compartments are added to the model (e.g. exposed subpopulation, vaccinated subpopulation, vector population etc.) with their own dynamics and assumptions.

2. The population is not constant; changes in demographics (births, deaths from disease or other causes, etc.) are embedded in the model. For some viruses, it's also advantageous to subdivide the population by age since the chance of recovery or transmission etc. can depend on age. When age is considered as a continuous variable, the model 'changes' into a partial differential equation (PDE). These models are called age-structured epidemic models[2, Ch. 5].

3. The *force of infection* $\lambda(t)$ is different. The two most commonly used ones are *mass action* $\lambda(t) = \beta S$ (as in the model above) and *standard* $\lambda(t) = \frac{\beta S}{N}$, which is used when it is assumed that the number of contacts cannot increase indefinitely (for example for sexually transmitted diseases)[3]. Note that these two only differ when the population is not constant and there are also other, possibly highly non-linear forces of infections [4, Ch. 3]

4. One can also consider the spatiality of the disease and the population. If the space is considered to be continuous, then one models it with PDEs. In the simplest case, the spatial spread is modelled as diffusion and the PDE is of the reaction-diffusion type[3, Ch. 15].

5. For some diseases, the infectivity changes by the time-since-infection. In these cases, the system is usually modelled as an integro-differential equation. But in general, not only the infectious class, but other classes can be structured by the duration of residence in that class. These models are called *class-age structured epidemic models*[3, Ch. 13].

For all cases, one can define stationary solutions, which does not depend on time. These are called the equilibrium points of the system. The two most common equilibria are the disease-free equilibrium (DFE) and the endemic equilibrium. The first is characterised by that the disease is not present while, for the second, the disease is persistent.

An important question is when - w.r.t the parameters - are these equilibria are (asymptotically) stable in the sense that other, time-dependent solutions whether approach the equilibria or not. In the next subsection, we will introduce a well known method which can be used to analyse the stability of the DFE.

1.1 The Basic reproduction number

In epidemiological modelling, one of the most important measure of a disease model is the *basic reproduction number*, usually denoted by R_0. It denotes the number of secondary infections produced by an infected individual in a completely susceptible population. Therefore, it is a threshold parameter for the invasion of a disease organism into the population and in general coincides with the threshold condition for the stability of the disease-free equilibrium[3][5].

One can compute R_0 for a compartmental ODE system by the next generation approach[5]. First, the system is rewritten as

$$\dot{x}_i = F_i(x, y) \quad V_i(x, y) \quad i = 1, \ldots, k \tag{1.2}$$

$$\dot{y}_j = g_j(x, y) \quad j = 1, \ldots, d \quad k \tag{1.3}$$

where $\dot{x}_i$ is the derivative with respect to time t of the function $x_i(t)$, and $(x_1, \ldots, x_k)$ are the disease compartments, while y_j, $j = 1, d \quad k$ are the non-disease compartments. $F_i(x, y)$ represents the rate of new infection in compartment i, while $V_i(x, y)$ incorporates the remaining transitional terms. There are some epidemiologically meaningful (and not strict) assumptions on the functions $F = fF_i g_{i=1}^k, V = fV_i g_{i=1}^k, g = fg_i g_{i=1}^{d\ k}$, which we won't all state, but can be found in [5, pg. 161]. One important assumption on g is that the disease-free system

$$\dot{y} = g(0, y) \tag{1.4}$$

has a unique equilibrium $E_0 = (0, y_0)$ such that the solutions of this disease-free system approach it as $t \ ! \ 1$. Another assumption is that

$$F_i(0, y) = V_i(0, y) = 0 \quad 8i = 1, \ldots, k, \ 8y \quad 0 \tag{1.5}$$

which means that there is no change in the infectious classes when the infection is not present i.e. all infections are secondary. The Jacobi matrices of the subsystems F and V at the disease free equilibrium E_0 are denoted as F and V. Linearizing the system (1.2)-(1.3) at the DFE gives $\dot{x} = (F \quad V)x$, since the infected compartments x are decoupled from the remaining equations, because for every pair (i, j) by (1.5):

$$\frac{\partial F_i(0, y_0)}{\partial y_j} = \frac{\partial V_i(0, y_0)}{\partial y_j} = 0.$$

Then the *next generation matrix* is $K := FV^{\ 1}$ and its spectral radius denotes the basic reproduction number: $\rho(K) = R_0$.

The question is how R_0 related to the stability of the DFE? In [5], it was proved that the matrix $F \quad V$ has all eigenvalues with negative real part (which implies local stability) if and only if $\rho(FV^{\ 1}) < 1$ and has an eigenvalue with positive real part if and only if $\rho(FV^{\ 1}) > 1$. This gives the correspondence between the local stability of the system (1.2)-(1.3) and R_0 by the assumptions on (1.4).

Later, we will use the above approach to interpret and calculate the basic reproduction number for a non-conservative model. For another model, we will calculate R_0 by the 'Jacobi approach', i.e. directly from the stability of the linearized system, which will be reduced to a single condition, but it can be easily checked that the above approach would give the same results.

2 General theory of ODEs, considered qualitative properties

We first introduce the general theory of ordinary differential equations and some of their properties, namely the stability, positive invariance, positivity and mass-preservation. The following is mostly based on [4],[6] and [7].

For this, consider the following dynamical system defined by the general autonomous ODE:

$$\dot{u} = f(u) \tag{2.1}$$

where U is a region (i.e. open and connected subset) of $\mathbb{R}^d$ and $f : U \to \mathbb{R}^d$ a continuous function. When an initial value $u(t_0) = u_0$ is also given, we call the above an initial value problem (IVP). We call $u(t)$, where $u : I \to U$ a solution of the above IVP if

1. $I \subset \mathbb{R}$ is non-empty open interval containing $t = t_0$
2. u is differentiable with continuous derivative in I
3. $\dot{u}(t) = f(u(t)) \quad (\forall t \in I)$
4. $u(0) = u_0$

Note that if $u(t)$ is a solution with initial value $u(t_0) = u_0$, then $u(t - t_0)$ is a solution with initial value $t_0 = 0$, thus the solutions of (2.1) with initial values $u(0) = u_0$ completely defines the solutions of (2.1) with more general initial values $u(t_0) = u_0$. It is clear that if the above system models a biological process, then it is a basic requirement, that the solution with some initial value $u(0) = u_0$ exists and unique. This holds if f is locally Lipschitz continuous for all $u_0 \in D$, which means that for all $u_0 \in D$ there exist $L, \delta > 0$ constants such that

$$\|f(u_1) - f(u_2)\| \leq L\|u_1 - u_2\|, \quad (\forall u_1, u_2 \in \{u \in U \mid \|u - u_0\| < \delta\})$$

The above conditions from an epidemiological modelling viewpoint is not strict and it can be proved with the fundamental theorem of calculus that if $f \in C^1(D)$, then f is Locally Lipschitz in every point of D. From now on, we will always assume that $f \in C^1(D)$ for (2.1).

Until now, we only considered solutions which are locally defined on some I, but in general, we want to extend - or show that it can be extended - for all $t \in \mathbb{R}$ or for at least $\mathbb{R}^+$. This is not always possible, but it can be shown that the only way it can be violated is if the solution 'blows up' in finite time, i.e. there exists some $T^+ < \infty$ such that $\lim_{t \to T^+} \|u(t)\| = \infty$ or $\exists T < \infty$ such that $\lim_{t \to T} \|u(t)\| = \infty$. This gives us a way to guarantee the global existence of a solution, namely if there exists some region $\Omega \subset D$ such that every solution that start in Ω stays in Ω for both backward and forward in time, then the solution cannot blow up, therefore exist $\forall t \in \mathbb{R}$. Such regions are called invariant (w.r.t. f in (2.1)). The same logic can be used to guarantee a solution 'only' for all $t \in (-a, \infty)$, or in $t \in (-\infty, a)$, $(a > 0)$ when Ω is respectively positively or negatively invariant.

We now turn our interest to the qualitative properties of different solutions of (2.1). We will only talk about such properties which are general interest for epidemiological models. We start with the simplest solutions which does not change in time, therefore they are called *stationary points* or *equilibria* (with the plural form equilibrium). The equilibrium points of the system can be found by solving $f(u\ ) = 0$, then $u(t) \quad u$ is clearly a solution for (2.1) and it is globally defined since it stays bounded. To study the solutions near the equilibrium, the concept of stability is of great importance.

Definition 2.1. *An equilibrium point u 2 U of (2.1) is said to be (locally) stable if for any $\varepsilon > 0$ there exists $\delta > 0$ such that for any initial values u_0 with $ku_0 \quad u\ k < \delta$ the solutions exist for all $t > 0$ and $ku(t) \quad u\ k < \varepsilon\ 8t > 0$. The equilibrium points are called unstable if it is not stable and (locally) asymptotically stable if it is stable and in addition there exist some $\gamma > 0$ such that for all $ku_0 \quad u\ k < \gamma$, holds that $\lim_{t\ !\ 1} ku(t) \quad u\ k = 0$.*

To study the stability of the equilibria, one considers the solutions which are sufficiently close to it (so they can be viewed as a perturbation of the stationary solution). For this, suppose that $u(t)$ is a solution with initial value u_0, then $y(t) = u(t) \quad u$ is a solution and for sufficiently smooth f:

$$\dot{y}(t) = \dot{u}(t) = f(u(t)) = f(u\ ) + f'(u\ )y(t) + r(y(t)) = f'(u\ )y(t) + r(y(t)) \tag{2.2}$$

by Taylor expansion, where $f'(p)$ is the Jacobian of f at p 2 R^d and r is the residue. For small enough y, the linear part dominates, so we obtain the approximate linear system:

$$\dot{y}(t) = f'(u\ )y(t) \tag{2.3}$$

The following theorem tells us what can we infer from the linearized system (3.32) considering the stability of the equilibrium of the nonlinear system (2.1):

> **Theorem 2.1.** *Suppose that $f'(u\ )$ is hyperbolic i.e. it does not have any eigenvalues in the imaginary axis, then*
>
> 1. *If all the eigenvalues of $f'(u\ )$ has negative real parts, then u is locally asymptotically stable equilibrium of (2.1).*
> 2. *If $f'(u\ )$ has an eigenvalue with positive real part, then u is an unstable equilibrium of (2.1).*

Note that we have excluded the cases where $f'(u\ )$ have one or more eigenvalues in the imaginary axis, because in this case additional analysis is required with the higher order terms of the Taylor expansion. We also point out that the above theorem only gives us local results, i.e. asymptotic stability in some region $\tilde{U} \quad U$. Since we use these for biological models, where the solutions on the negative orthants and some other, generally unbounded solutions are not of interest; we call an equilibria globally asymptotically stable (GAS), when $\tilde{U} = \Omega$, where Ω is the so-called *biologically feasible region*.

The above theorem can be proved by constructing a so-called *Lyapunov function*, which can also be used for specific models, where one can also find the asymptotic stability region of the equilibrium. For simplicity, we only define Lyapunov functions for equilibria $u\ = (0, \dots, 0)^{\mathsf{I}}$.

Definition 2.2. *Let $\tilde{U} \quad \mathsf{R}^d$ a region such that 0 2 $\tilde{U}$, V 2 $C^1(\tilde{U}, \mathsf{R})$. Then V, is said to be a Lyapunov function for the equilibria $u\ = 0$ of (2.1), if*

1. $\tilde{U} \quad U$
2. $V(0) = 0$

3. *(positive de niteness)* $V(z) > 0, \quad (8z\ 2\ \tilde{U} \quad f0g)$

4. $\dot{V}(u) := r\,V(u) \quad f(u) \quad 0, \quad (8u\ 2\ \tilde{U} \quad f0g)$

It can be proved, that if a Lyapunov function can be found for (2.1), which is translated such that the equilibrium u is at 0, then u is stable. Moreover, if in (4.), we have strict inequality ('<'), then 0 is asymptotically stable, while if '> 0', then unstable. We don't give formal proof, but geometrically (4.) for the solutions $u(t)$ of (2.1) is:

$$\dot{V}(u(t)) = r\,V(u(t)) \quad f(u(t)) = kf(u(t))kk\,r\,V(u(t))k\,cos(\theta) < 0$$

where θ is the angle between $r\,V(u(t))$ and $f(u(t))$ at given t. So the orbit of $u(t)$ at a given t is crossing the level curve of V from the outside to the inside, since the angle is obtuse. Similarly, it is tangential or cross from the inside to the outside if we have relations $= 0$ and > 0, respectively. It is clear if $\tilde{U}$ is 'large enough', then one can possibly use the Lyapunov function to show GAS of the single equilibrium in Ω.

In some cases, one can show asymptotic stability even when the relation for (4.) is not strict by the LaSalle Invariance Principle

Theorem 2.2 (LaSalle Invariance Principle[8])**.** *Let V be a Lyapunov function of* (2.1) *in* $\tilde{U} \quad U$ *which is continuous on the closure of* $\tilde{U}$*, denoted by* $\bar{U}$*. Let E be the set where the solutions are tangential to the level curves of V, i.e.*

$$E := fz\ 2\ \bar{U}\,j\,\dot{V}(z) = 0g$$

and let M denote the largest invariant subset of the solutions in E. Suppose that any solutions with initial point $u_0\ 2\ \tilde{U}$ *are bounded and remain for all future time in* $\tilde{U}$*. Then every solution starting in* $\tilde{U}$ *approaches M as* $t \ ! \ 1$*.*

It is clear, that if M consist of a single point u , then under the above theorem u is GAS (in $\tilde{U}$).

In epidemiological modelling, when the ODE represents the rate of change of the different subpopulations, the solution represents the number of a given subpopulation or the density evolving in time. In this case, it is clear that the solutions with initial conditions in the positive quadrant should remain there (since negative population or density is biologically incomprehensible):

Definition 2.3 (Positivity of ODE/IVP)**.** *We say that the ODE/IVP* (2.1) *is positive if whenever* $U\ 3\ u_0 \quad 0$*, then* $U\ 3\ u(t) \quad 0, \ 8t \quad 0$ *(where the relation is considered coordinate-wise). We denote the set of positive functions by P.*

Note that altough it is called positivity, we require non-negativity of the solutions and this is a special case of positive invariance. To prove whether $f\ 2\ P$, one has to check whether the solutions are reflected back at the boundary.

Theorem 2.3. *[[9] Thm.7.1.] Suppose that* f *in* (2.1) *is continuous and locally Lipschitz. Then* $f\ 2\ P$ *if and only if* $8v\ 2\ \mathbb{R}^d, \ 8i\ 2\ fi, \ldots, mg$*:* $v \quad 0, v_i = 0$ *implies that* $f_i(v) \quad 0$*.*

Proof. The necessity of the condition follows by considering the solution $u(t)$ with initial value $v_i = 0$, then $\dot{u}_i(0) < 0$, so for the solution $u_i(t) < 0$ for $t\ 2\ (0, t_1)$ for some $t_i\ 2\ \mathbb{R}^+$.

For the sufficiency, not that the conditions imply that for the solutions $u(t) \quad 0$ with $u_i(t) = 0$ it holds that $\dot{u}_i(t) \quad 0$. We need that $\dot{u}_i(t) \quad \varepsilon > 0$, then the solution gets reflected back from the boundary of the positive orthant. Consider the perturbed system

$$\tilde{f}(u) = f(u) + I\varepsilon$$

which is positive i.e. $\tilde{f} \in P$ by the above argument. By the Lipschitz condition, it holds that the unperturbed solution will be arbitrary-well approximated by the perturbed solution if we let $\varepsilon \to 0$. □

In some epidemiological models we can assume that the population does not change over time. This is usually the case for diseases where the dynamics happen fast. Models with this property are called *conservative* or *mass-preserving*.

Definition 2.4. *We call an autonomous system* (2.1) *conservative if for arbitrary initial value* $u(0) \in U \subset \mathbb{R}^d$, $e^\top u(t) = e^\top u(0)$ $\forall t \geq 0$, *where* $e := (1, \ldots, 1) \in \mathbb{R}^d$.

Remark 2.1. *Conservativity is equivalent with* $e^\top f(u) = 0$, $\forall u \in \mathbb{R}^d$, *since it implies that for the solutions we have* $e^\top \dot{u}(t) = 0$, *then integrating both sides we get the condition of conservativity. The other way follows from differentiation.*

3 Considered numerical methods and some of their properties

The discretisation of the different continuous epidemiological models are inevitable if we want to solve them numerically. Then it is of interest whether the discretised model posses the same qualitative behaviour as the continuous model.

In general, numerical k-step methods with fixed step size for autonomous ODEs generate a discrete map

$$\Phi_{f,\Delta t} : (u_n, \ldots, u_{n+k-1}) \mapsto u_{n+k} \tag{3.1}$$

where $u_0, u_1, \ldots, u_k$ initial values are given and u_n approximates $u(t_n) = u(\Delta t n)$, where Δt is the fixed step-size. (3.1) is sometimes called the *numerical flow*. There exist a number of different numerical methods, from which two are the well-known Runge-Kutta (RK) schemes and the linear multistep methods.

Since we are interested in the preservation of different qualitative behaviour under the discretisations, we first define these properties for maps. The positivity of a numerical method can be defined in the logical way

Definition 3.1. *Let there be given a numerical method* (3.1), *a set of functions* $F \subset P$ *and a real number* $0 < H \leq 1$. *We call the method positive on* F *with threshold* H *if the numerical approximation* (3.1) *are non-negative whenever* $f \in F$, $u_0 \in \mathbb{R}^d_+$ *with step size* $0 < \Delta t < H$. *If* $H = 1$, *then we call the method unconditionally positive, otherwise conditionally positive.*

Note that for multistep methods, one can talk about a multistep method being positive with suitable starting procedure or with any starting procedure.

Considering the conservativity preservation of a numerical method, the definition is

Definition 3.2. *We call a numerical method* (3.1) *conservative if for arbitrary conservative system* f *(see 2.4) it holds that for* $(u_{n+1,1}, \ldots, u_{n+1,d}) := \Phi_{f,\Delta t}(u_n, \ldots, u_{n+k-1})$: $\sum_{i=1}^{d} (u_{n+1,i} - u_{n,i}) = 0$ *for arbitrary* $u_n = u(t_n), \ldots, u_{n+k-1} = u(t_{n+k-1})$, *where* $t_n := \Delta t n$.

It is clear that under the condition $e^\top f(u) = 0$, $\forall u \in \mathbb{R}^d$ LMM preserves conservativity using the consistency condition, while RK methods preserves conservativity for the numerical solution and for the stage values also.

The question considering the equilibria and its stability under discretisation by a numerical scheme will be considered in subsection 3.5.

3.1 Runge-Kutta methods

Runge-Kutta methods are special - usually higher order - one-step methods. The general form of a m-stage Runge-Kutta method for an autonomous ODE is:

$$k_i = f\big(u_n + \Delta t \sum_{j=1}^{m} a_{ij} k_j\big), \quad (i = 1, \dots, m) \tag{3.2a}$$

$$u_{n+1} = u_n + \Delta t \sum_{i=1}^{m} b_i k_i \tag{3.2b}$$

where from the consistency conditions we have $b^{\intercal} e = 1$, where $e := (1, \dots, 1)^{\intercal} \in \mathbb{R}^s$. We define $b^{\intercal} := (b_i)_{i=1}^m$, $A := (a_{ij})_{i,j=1}^m$. k_i-s are the approximation of the derivatives at the stages $t_n + \Delta t c_i$, where $(c_i)_{i=1}^m = c := Ae$. If A is a lower triangular matrix with zero diagonal values, we call the method explicit, because the i-th stage can be explicitly calculated from the stages $1, \dots, i-1$. Otherwise the method is called implicit and one has to use some other numerical method to solve the equation at each step. All Runge-Kutta methods are mass-preserving/conservative, which follows directly from remark 2.1. One can write the RK methods in an equivalent but different formulation, when the stage values approximate the solution - not the derivative - at $t_n + \Delta t c_i$, $i = 1 \dots, m$:

$$u^{(i)} = u_n + \Delta t \sum_{j=1}^{s} a_{ij} f\big(u^{(j)}\big), \quad (i = 1, \dots, m) \tag{3.3a}$$

$$u_{n+1} = u_n + \Delta t \sum_{i=1}^{m} b_i f(u^{(i)}) \tag{3.3b}$$

where $u^{(i)}$, $i = 1, \dots, m$ are the stage values. The simplest RK method is the one stage, first order explicit Euler method

$$u_{n+1} = u_n + \Delta t f(u_n)$$

which has the Butcher tableau

$$\begin{array}{c|c} c & A \\ \hline & b^t \end{array} = \begin{array}{c|c} 0 & \\ \hline & 1 \end{array}.$$

One generally used fourth order, explicit four stage method is the *classical RK4* method, which has the Butcher tableau

$$\begin{array}{c|c} c & A \\ \hline & b^t \end{array} = \begin{array}{c|cccc} 0 & & & & \\ 1/2 & 1/2 & & & \\ 1/2 & 0 & 1/2 & & \\ 1 & 0 & 0 & 1 & \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array}$$

It can be seen, that the above scheme is explicit. The RK4 method in the second formulation is:

$$\begin{aligned} u^{(1)} &= u_n \\ u^{(2)} &= u_n + \Delta t \frac{1}{2} f(u^{(1)}) \\ u^{(3)} &= u_n + \Delta t \frac{1}{2} f(u^{(2)}) \\ u^{(4)} &= u_n + \Delta t f(u^{(3)}) \\ u_{n+1} &= u_n + \Delta t \big[\frac{1}{6} f(u^{(1)}) + \frac{1}{3} f(u^{(2)}) + \frac{1}{3} f(u^{(3)}) + \frac{1}{6} f(u^{(4)})\big] \end{aligned} \tag{3.4}$$

3.2 Strong stability preserving (SSP) Runge-Kutta Methods

In this subsection, we introduce the *strong stability preserving Runge-Kutta methods* (SSP RK methods) and their special property. Later, the connection with the classical RK methods and the general questions considering consistency, convergence etc. will be investigated.

To introduce the idea of the SSP RK methods, we define an (explicit) m stage method in the following form:

$$\begin{aligned} u^{(0)} &= u_n \\ u^{(i)} &= \sum_{j=0}^{i-1} \left(\alpha_{ij} u^{(j)} + \Delta t \beta_{ij} f(u^{(j)}) \right), \quad (i = 1, \dots, m) \\ u_{n+1} &= u^{(m)} \end{aligned} \tag{3.5}$$

where $u^{(i)}$, $i = 1, \dots, m$ are the stage values and α_{ij}, β_{ij}; $i, j = 1, \dots, m$ are given constants. For example, a three-stage method is

$$\begin{aligned} u^{(0)} &= u_n \\ u^{(1)} &= u^{(0)} + \Delta t f(u^{(0)}) \\ u^{(2)} &= \frac{3}{4} u^{(0)} + \frac{1}{4} u^{(1)} + \frac{1}{4} \Delta t f(u^{(1)}) \\ u_{n+1} = u^{(3)} &= \frac{1}{3} u^{(0)} + \frac{2}{3} u^{(2)} + \frac{2}{3} \Delta t f(u^{(2)}). \end{aligned}$$

Now, suppose that we require from our numerical solution that

$$\|u_{n+1}\| \quad \|u_n\|, \quad (\forall n = 0, 1, 2, \dots) \tag{3.6}$$

where $\|.\|$ is some not yet specified norm. One can guarantee this, by assuming that we have this property for the explicit Euler method under some step-size condition:

$$\|u + \Delta t f(u)\| \quad \|u\|, \quad (\forall u \in \mathbb{R}^d, \forall \Delta t \quad \Delta t_{FE}). \tag{3.7}$$

Note that (3.7) states that the property (3.6) holds for the explicit Euler method under small enough step-sizes, where $\Delta t_{FE} \in \mathbb{R}^+$ is the largest such step-size for which it holds. Then we can guarantee (3.6) for (3.5) if α_{ij}, β_{ij} are non-negative and $\sum_{j=0}^{i-1} \alpha_{ij} = 1$ under the step-size restriction $\Delta t \max_{ij} \frac{\beta_{ij}}{\alpha_{ij}} \quad \Delta t_{FE}$, i.e. $\Delta t \quad \mathcal{C}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \Delta t_{FE}$, where $\mathcal{C}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \min_{ij} \frac{\alpha_{ij}}{\beta_{ij}}$, because

$$\begin{aligned} \|u^{(i)}\| &= \left\| \sum_{j=0}^{i-1} \alpha_{ij} \left(u^{(j)} + \Delta t \frac{\beta_{ij}}{\alpha_{ij}} f(u^{(j)}) \right) \right\| \\ &\sum_{j=0}^{i-1} \alpha_{ij} \left\| u^{(j)} + \Delta t \frac{\beta_{ij}}{\alpha_{ij}} f(u^{(j)}) \right\| \\ &\sum_{j=0}^{i-1} \alpha_{ij} \|u_n\| = \|u_n\| \end{aligned} \tag{3.8}$$

Specifically, (3.8) for $i = m$ we have $\|u_{n+1}\| \quad \|u_n\|$, which is the desired property (3.6).

These methods are called strong stability preserving Runge-Kutta methods and their coefficients $\mathcal{C}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ are called the apparent SSP coefficients. The strong stability adjective comes from the property (3.6), which implies the absolute stability for linear (stable) problems. The strength of these methods comes from the fact that we have not fixed the norm $\|.\|$, so this preservation holds for arbitrary norms and since we have only used it for (3.8), it is easy to see that it also holds

for arbitrary seminorms and convex functionals. A popular application of these methods is for the semidiscretisation of nonlinear convection PDEs with discontinuous initial values. In this case, the discontinuous initial data can give rise to unwanted oscillations for the numerical solutions; in this case the used seminorm is the *total variation*, which is defined as for a vector $v := (v_i)_{i=1}^n$ as $\|v\| := TV(v) := \sum_{j=2}^n |v_j - v_{j-1}|$. One can avoid spatial oscillations by considering this norm in (3.6) and (3.7)[10][11]. Such methods with the above property are sometimes also called *Total Variation Diminishing*. Two other property preservations we will use and can be considered in the context of strong stability preservation are the positivity and boundedness preservation. For the positivity, instead of (3.6) one considers the largest time-step Δt_{FE} for which positivity is preserved for the explicit Euler method:

$$u + \Delta t f(u) \geq 0 \quad (\forall u \in \mathbb{R}^d_+, \forall \Delta t \leq \Delta t_{FE}) \tag{3.9}$$

and for the preservation of the boundedness property in $\|.\|_1$ norm:

$$\|u + \Delta t f(u)\|_1 \leq M \quad (\forall u \in \Omega \subseteq \mathbb{R}^d, \forall \Delta t \leq \Delta t_{FE}) \tag{3.10}$$

where Ω is some region in $\mathbb{R}^d$. Note that in both cases, we reduce the initial values to a region of vectors. It can be easily seen that the above 'proof' (3.8) works for these too (with different Δt_{FE}s, which we suppress in the notation). It is clear, that one can also consider both (3.9) and (3.10) with the smaller time-step restriction for the explicit Euler method and one can also consider (3.6) and (3.10) for only one coordinate (i.e. if we denote $u = (u^i)_{i=1,\dots,d}$, then $|u_n^i| \leq |u_{n+1}^i|$ and $|u_n^i| \leq M$).

The adjective 'Runge-Kutta' is used because - as we will see in the next subsection - the above-mentioned methods are Runge-Kutta methods but in a different form, where $\sum_{j=1}^{i-1} \alpha_{ij} = 1$, $i = 1, \dots s$ is used for the consistency condition. The SSP coefficient $C(\boldsymbol{\alpha}, \boldsymbol{\beta})$ has the adjective 'apparent', because the representation (3.5) is not unique. For example, one can rewrite

$$\alpha_{21} u^{(1)} = (\alpha_{21} - c) u^{(1)} + c u^{(1)} = (\alpha_{21} - c) u^{(1)} + c(\alpha_{10} u^{(0)} + \Delta t \beta_{10} f(u^{(0)})),$$

where $c \in \mathbb{R}$ arbitrary, so

$$u^{(2)} = (\alpha_{20} + c\alpha_{10}) u^{(0)} + \Delta t(\beta_{20} + c\beta_{10}) f(u^{(0)}) + (\alpha_{21} - c) u^{(1)} + \Delta t \beta_{21} f(u^{(1)}).$$

It is clear that in this case $C(\boldsymbol{\alpha}, \boldsymbol{\beta})$ possibly changes, and/or the new α_{21} could become negative. For these reasons, the important task is to find the largest such coefficient between the different representations which we will denote as

$$C := \max_{(\boldsymbol{\alpha}, \boldsymbol{\beta})} C(\boldsymbol{\alpha}, \boldsymbol{\beta})$$

where the maximum is over the different possible representations. C is called the *SSP coefficient*. To find, the SSP coefficients, we introduce other formulations for the SSP RK methods.

3.2.1 Different representations of SSP RK methods and the SSP coefficient

The general SSP method, which incorporates both explicit and implicit methods written in its *modified Shu-Osher form* is:

$$u^{(i)} = v_i u_n + \sum_{j=1}^{m} \left(\alpha_{ij} + \Delta t \beta_{ij} f(u^{(j)}) \right), \quad (i = 1, \dots, m+1) \tag{3.11a}$$

$$u_{n+1} = u^{(m+1)} \tag{3.11b}$$

with the assumption

$$v_i + \sum_{j=1}^{m} \alpha_{ij} = 1, \quad (i = 1, 2, \ldots, m+1). \tag{3.12}$$

The difference between (3.5) and the above is that the summation goes from 1 to m, so the method can be implicit. Also, the u_n terms are explicitly there and the stages are indexed from one. One can prove (similarly as for the explicit case), that if the property (3.6) is preserved for the explicit Euler method under the step-size restriction Δt Δt_{FE}, then it is preserved by the above method when the step-size satisfies

$$0 \quad \Delta t \quad C(\boldsymbol{\alpha}, \boldsymbol{\beta}) \Delta t_{FE} \tag{3.13}$$

where

$$C(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \begin{cases} \min\limits_{i,j} \frac{\alpha_{ij}}{\beta_{ij}} & \text{if all } \alpha_{ij}, \beta_{ij}, v_i \text{ are non-negative} \\ 0 & \text{otherwise} \end{cases}$$

and if $\beta_{ij} = 0$, for some i, j then $\frac{\alpha_{ij}}{\beta_{ij}} = +\boldsymbol{1}$. To show that these methods are Runge-Kutta methods in different representations, first we have to rewrite (3.11a) in vector notations. For this we define the matrices $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}$ 2 R^{m+1} $^{m+1}$ as $f\hat{\boldsymbol{\alpha}}g_{i,j=1}^{m+1,m} = \alpha_{ij}$ and similarly $f\hat{\boldsymbol{\beta}}g_{i,j=1}^{m+1,m} = \beta ij$ where the yet undefined last columns has zero entries. Because the considered ODE systems are not necessarily one dimensional, we define:

$$\begin{aligned} y &= \left(u_1^{(1)}, u_1^{(2)}, \ldots, u_1^{(m+1)}, u_2^{(1)}, \ldots, u_d^{(m+1)}\right)^{\mathsf{I}} \\ y_{m+1} &= \left(u_1^{(m+1)}, u_2^{(m+1)}, \ldots, u_d^{(m+1)}\right)^{\mathsf{I}} \\ \boldsymbol{v} &= I \quad v \\ \boldsymbol{\alpha} &= I \quad \hat{\boldsymbol{\alpha}} \\ \boldsymbol{\beta} &= I \quad \hat{\boldsymbol{\beta}} \end{aligned}$$

where I is the d dimensional identity matrix and the symbol denotes the Kronecker product:

$$A \quad B := \begin{pmatrix} a_{11}B & a_{12}B & & a_{1n}B \\ a_{21}B & a_{22}B & & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}B & a_{n2}B & & a_{nn}B \end{pmatrix}$$

where A 2 R^{n} ^{n} and the size of B is arbitrary but fixed. Note that y depends on n, but we omit this from the notation. Using the above defined notations, we can rewrite (3.11a)-(3.11b) as

$$\begin{aligned} y &= \boldsymbol{v}u_n + \boldsymbol{\alpha}y + \Delta t \boldsymbol{\beta} f(y) \\ u_{n+1} &= y_{m+1}. \end{aligned} \tag{3.14}$$

If I $\boldsymbol{\alpha}$ is invertible, then we can rewrite (3.14) as

$$\begin{aligned} y &= (I \quad \boldsymbol{\alpha})^{\ 1}\boldsymbol{v}u_n + \Delta t(I \quad \boldsymbol{\alpha})^{\ 1}\boldsymbol{\beta} f(y) \\ &= \boldsymbol{e}u_n + \Delta t(I \quad \boldsymbol{\alpha})^{\ 1}\boldsymbol{\beta} f(y) \end{aligned}$$

where we have used that $(I \quad \boldsymbol{\alpha})^{\ 1}\boldsymbol{v} = \boldsymbol{e}$ (i.e. (3.12)). We will denote

$$\boldsymbol{\beta}_0 := (I \quad \boldsymbol{\alpha})^{\ 1}\boldsymbol{\beta}. \tag{3.15}$$

A Runga-Kutta methods in vector notation with

$$\mathsf{A} = \begin{pmatrix} I & A & 0 \\ I & b^{\mathsf{I}} & 0 \end{pmatrix}$$

is

$$y = \boldsymbol{e}u_n + \Delta t \mathsf{A} f(y)$$

so, in the case if $I \quad \boldsymbol{\alpha}$ is invertible, then the SSP method is equivalent with the Runge Kutta method $\mathsf{A} = \boldsymbol{\beta}_0$. To exclude the cases when $I \quad \boldsymbol{\alpha}$ is singular, we define

Definition 3.3. *A method is said to be zero-well de ned if for the IVP* $\dot{u} = 0$ *with initial value* $u(0) = u_0$ *the method has a unique solution.*

Now, if we assume that the method is zero-well defined, then (3.14) for the IVP $\dot{u} = 0$; $u(0) = u_0$ is $(I \quad \boldsymbol{\alpha})y = \boldsymbol{v}u_n$. Since $(I \quad \boldsymbol{\alpha}) \ ^1\boldsymbol{v} = \boldsymbol{e}$ (i.e. (3.12)), then $I \quad \boldsymbol{\alpha}$ is invertible under the assumption of zero-well definiteness by contradiction. Note that Runge-Kutta methods are all zero-well-defined since for the IVP in the definition we have $y = \boldsymbol{e}u_n$. It should be noted here that in the literature, assumption (3.12) is generally called 'consistency assumption', but despite its name, it is not sufficient for the consistency of its RK representation. This can be easily seen from the Taylor expansion.

It is clear that the Butcher-form of the method can be considered as a unique representation of the method but the SSP coefficient is not apparent. To find the SSP coefficient, one uses a third representation, called *Canonical Shu-Osher form*. For this, consider a particular representation of the method in its modified Shu-Osher for which it holds that

$$\frac{\alpha_{ij}}{\beta_{ij}} = const = r \quad (\forall i, j) \tag{3.16}$$

i.e. $\boldsymbol{\alpha}_r = r\boldsymbol{\beta}_r$, where we use the indices $._r$ to show that it depends on r. In this case, (3.14) is

$$y = \boldsymbol{v}_r u_n + \boldsymbol{\alpha}_r\big(y + \frac{\Delta t}{r} f(y)\big). \tag{3.17}$$

This is the so-called Canonical Shu-Osher form. To see which Runge-Kutta methods can be represented in this form with a given r we have to write down the relationship between $\boldsymbol{\beta}_0$ and $(\boldsymbol{\alpha_r}, \boldsymbol{\beta_r})$:

$$\begin{aligned} (I \quad r\boldsymbol{\beta}_r)\boldsymbol{\beta}_r = \boldsymbol{\beta}_0 \ &\Longrightarrow \ \boldsymbol{\beta}_r = \boldsymbol{\beta}_0(I + r\boldsymbol{\beta}_0) \ ^1 \\ &\Longrightarrow \ \boldsymbol{\alpha}_r = r\boldsymbol{\beta}_0(I + r\boldsymbol{\beta}_0) \ ^1 \\ \boldsymbol{v}_r = (I \quad \boldsymbol{\alpha}_r)\boldsymbol{e} \ &\Longrightarrow \ \boldsymbol{v}_r = (I + r\boldsymbol{\beta}_0) \ ^1\boldsymbol{e} \end{aligned} \tag{3.18}$$

if $(I + r\boldsymbol{\beta}_0)$ is non-singular. We have used (3.15), (3.16) and that the method is consistent. It is clear that in this case we have

$$\mathsf{C}(\boldsymbol{\alpha}_r, \boldsymbol{\beta}_r) = r.$$

In summary, to calculate the SSP coefficient for a given RK method, we have to check whether for a given $r > 0$ it holds that

$$\boldsymbol{\alpha}_r, \boldsymbol{v}_r \quad 0 \text{ and } (I + r\boldsymbol{\beta}_0) \text{ is nonsingular} \tag{3.19}$$

where $\boldsymbol{\beta}_r \quad 0$ is omitted because of (3.16) and $\boldsymbol{\alpha}_r, \boldsymbol{\beta}_r, \boldsymbol{v}_r$ are defined as in (3.18). The relations meant elementwise. It turned out that such r-s for which (3.19) hold are closed sets in the form $[0, r_{max}]$ and the largest such r (r_{max}) is the SSP coefficient for the given RK method[12, Thm. 3.2.]. The importance of the former is that one can use bisection to approximate the value $r_{max} = \mathsf{C}$ for a given RK method.

It should be noted that the conditions (3.19) are written in numerous but equivalent forms in the literature. One of the reason is that instead of using the Shu-Osher form, one can give conditions for the contractivity preservation of different test equations by the positivity of the coefficients in their Taylor expansions at given points (the test problems are the scalar problems $\dot{u} = \lambda u$, $\dot{u} = \lambda(t)u$, and the vectorial linear problems $\dot{u} = L(t)u$, where $L(t)$ is a square matrix). This theory was developed by Kraaijevanger where he defined the radius of absolute monotonicity of an RK method[13]. For detailed comparison of the two theories see [14]. We point out that the SSP methods not only preserve the positivity of the steps, but also of the internal stages (3.3a) of the RK method[14].

3.2.2 Necessity, bounds, sharpness of the SSP coefficient

It is of clear importance whether there exists a p-th order, s stage RK method with positive SSP coefficient. Clearly, for both explicit and implicit methods, one has an upper bound based on the general Runge-Kutta theory. To get an idea what the conditions (3.19) imply on the Butcher coefficients, the following can be stated and proved:

Proposition 3.1 ([12]). *Any Runge-Kutta method with positive SSP coe cient $C > 0$ has A 0, b 0.*

Proof. Given an RK method (A, b) we can rewrite it in its Canonical Shu-Osher form (3.17) for all r 2 $[0, C]$, i.e. the conditions (3.19) holds, from which β_r 0 is

$$\beta_0(I + r\beta_0) \quad ^1 \quad 0,$$

using this for $r = 0$, we get the above conditions since $\beta_0 = A$. □

For DJ-irreducible methods[1] $b > 0$ [12, Pg. 65]. Using this and additional relationship between stage orders and order, one can show that for explicit methods with $C > 0$ we have p 4. It was also shown that for implicit methods, one has the order condition p 6 from A 0. The proofs can be found in [13] (Thm. 8.5., 8.6., 8.7.). Note that we did not give any conditions on the number of stages, so the fourth order 4 stage methods have special interest. It can be shown, that the classical RK4 method is the only (DJ irreducible) method with A 0, b 0[13, Thm. 9.6.]. Unfortunately, this method has $C = 0$. To see this, note that if $C > 0$, then one can choose small enough $r > 0$ such that $r < C$ and the Neumann series of $(I + r\beta_0)$ ^{1} exists (since for the existence of the Neumann series for a matrix, one 'only' needs that its spectral radius is strictly smaller than one and taking the spectral radius and multiplying a matrix by a real scalar commutes). Then:

$$\beta_r = \beta_0(I + r\beta_0) \quad ^1 = \beta_0 \quad r\beta_0^2 + \ldots \quad 0$$

Because β_0 0, one cannot have non-zero elements for β_0^2, where β_0 does have, which clearly implies that the same property is necessary for the RK coefficients A. This does not hold for a classical RK4 method because multiplying a lower triangular matrix with itself "shift down diagonally" the non-zero elements by one.

Another question, which arises, is the maximal SSP coefficient what a given order method can possess. For low order - 1st,2nd - explicit methods, one can construct arbitrary large (integer) C by considering more stages. For first order methods, it is clear that

[1] RK methods, for which the unnecessary stages which does not contribute to / does not in uence explicitly or implicitly the last stage $u^{(m+1)} = u_{n+1}$ are omitted.

$$\hat{\boldsymbol{v}} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}; \hat{\boldsymbol{\alpha}} = \begin{pmatrix} 0 & 0 & & 0 & 0 \\ 1 & 0 & & 0 & 0 \\ 0 & 1 & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & & 1 & 0 \end{pmatrix}; \hat{\boldsymbol{\beta}} = \begin{pmatrix} 0 & 0 & & 0 & 0 \\ \frac{1}{m} & 0 & & 0 & 0 \\ 0 & \frac{1}{m} & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & & \frac{1}{m} & 0 \end{pmatrix}$$

is an m-th stage method with C=m. If One would consider the same scheme with smaller off-diagonal elements in $\hat{\boldsymbol{\beta}}$, then the RK method would lose consistency [15]. For second order methods with s stages, the maximal attainable C is s 1 [15]. To compare two same order methods with different number of stages, one defines the e cient SSP coe cient $C_{eff} = \frac{C}{s}$ (i.e. more function evaluations are penalized). It was shown that for explicit methods C_{eff} 1, while for implicit methods C_{eff} 2 [12]. We also point out that the computational cost of solving nonlinear equations for implicit methods is not 'incorporated' into C_{eff} and is generally more than twice the computation for explicit methods; for this reason, implicit methods are less used.

As we have stated in (3.13) the condition

$$0 \quad \Delta t \quad C\Delta t_{FE} \tag{3.20}$$

for a fixed SSP RK method holds for arbitrary convex functionals k.k and functions f. We have also seen that it is generally strict in the sense that C_{eff} is small, so another important question is the sharpness/necessity of the condition for different function classes and norms. One can construct a function f and show that for some norm the condition (3.20) is necessary [12, Thm. 3.3.], but it is clear that if one considers specific function classes, then the conditions can be weakened. For example, if the considered function class is $ff\ 2\ C^1 jf(u)$ $c > 0; u$ $0g$, then arbitrary explicit RK method is unconditionally positive.

One popular way to expand the family of Runge-Kutta methods with non-zero SSP coefficient is by letting the values of $\boldsymbol{\beta}$ to be negative in (3.5) and in (3.11a). In this case one also requires that for some $\hat{f}$ we have the forward Euler condition (3.7) with 'backward stepping':

$$ku \quad \Delta t\hat{f}(u)k \quad kuk, \quad (8u\ 2\ \mathbb{R}^d, \Delta t \quad \Delta t_{FE}). \tag{3.21}$$

Clearly in this case the derivation of strong stability preservation in (3.8) holds if for the negative β_{ij} values we use the condition (3.21). Then the apparent SSP coefficient is:

$$C(\boldsymbol{\alpha}, \boldsymbol{\beta}\ ) = \begin{cases} \min_{i,j} \frac{\alpha_{ij}}{j\beta_{ij}j} & \text{if all } \alpha_{ij}, v_i \text{ are non-negative} \\ 0 & \text{otherwise} \end{cases}$$

Where the . represents that we let negative β_{ij} values with the function $\hat{f}$. Because we still want our approximation to converge to the solution of the continuous problem, we have to choose $\hat{f}$ in a meaningful way. For hyperbolic PDEs, if we choose $\hat{f}$ similarly as f but with opposite winding (for the semi/spatial-discretisation), then both f and $\hat{f}$ will approximate the original PDE. By allowing this, one can get positive SSP coefficients for (some) methods with negative elements in $\boldsymbol{\beta}$. This is the case for some four stage fourth order methods. For example, the classical RK4 method can be written as

$$\hat{\boldsymbol{v}} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}; \hat{\boldsymbol{\alpha}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{9} & \frac{2}{9} & \frac{2}{3} & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \end{pmatrix}; \hat{\boldsymbol{\beta}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{9} & \frac{1}{3} & 1 & 0 & 0 \\ 0 & \frac{1}{6} & 0 & \frac{1}{6} & 0 \end{pmatrix}.$$

It can be easily checked that $\mathsf{C}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{2}{3}$, but this representation $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is not unique. The above considered theory breaks down; namely, one cannot rewrite the RK4 method in the Canonical Shu-Osher form (3.17). Therefore, the same question arises as before: is the above representation optimal? (i.e. is $\mathsf{C}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ maximal over the different representations?). Fortunately, it turns out that one can study these methods (with negative β_{ij}-s) in a similar way as for positive β_{ij}-s considered above, by studying, the RK method under some perturbation. For the concrete theory, see [14]. The two problems with this is that the computational cost is more, because of the function evaluations of $\hat{f}$ and it is not clear how to construct $\hat{f}$ for ODEs which are not semi-discretisations of hyperbolic PDEs. The former was partially answered in [16] by using additive Runge-Kutta and additive SSP Runge-Kutta methods in an unconventional way. We leave out the details, but the result is that if $\hat{f} = \quad f$, then the method with negative β_{ij} values preserve positivity under a different step-size restriction which depends in a non-linear way on the considered method, and on the maximal step-sizes for which f and $\hat{f}$ preserves its positivity.

3.3 SSP linear multistep methods

The general form of a k-step linear multistep method (LMM) for an autonomous ODE is:

$$u_n + \alpha_1 u_{n\ \ 1} + \quad + \alpha_k u_{n\ \ k} = \Delta t(\beta_0 f_n + \beta_1 f_{n\ \ 1} + \quad + \beta_k f_{n\ \ k}), \quad n = k, k+1, \ldots \tag{3.22}$$

where $f_{n\ \ k} := f(u_{n\ \ k})$, where Δt is the constant step size and u_k is the approximation of $u(\Delta t k)$. Unlike for the RK methods, the consistency conditions of the LMM methods can easily be found from the Taylor expansions, which are $\sum_{j=0}^{k} \alpha_j = 0$ and $\sum_{j=0}^{k} j\alpha_j + \sum_{j=0}^{k} \beta_j = 0$, where $\alpha_0 = 1$. All linear multistep methods are mass-preserving/conservative, which follows directly from remark 2.1.

One can use the idea of considering convex combinations of forward Euler steps -introduced above for Runge-Kutta methods - for linear multistep methods to get positivity preservation: suppose that for a given f under the discretisation of the explicit Euler method is conditionally positive i.e.

$$0 \quad u + \Delta t f(u), \ \ 8u\ 2\ \mathsf{R}_+^n, \ 8\Delta t \quad \Delta t_{FE}. \tag{3.23}$$

Then an explicit LMM can be rewritten as:

$$\begin{aligned} u_n &= \sum_{j=1}^{k} \Big(\quad \alpha_j u_{n\ \ j} + \beta_j \Delta t f(u_{n\ \ j}) \Big) \\ &= \sum_{j=1}^{k} \quad \alpha_j \Big(u_{n\ \ j} + c_j \Delta t f(u_{n\ \ j}) \Big) \end{aligned}$$

where $c_j := \frac{\beta_j}{\alpha_j}$. The positivity holds for arbitrary starting procedures if $\alpha_j \quad 0$, $\beta_j \quad 0$ and $c_j \Delta t \quad \Delta t_{FE}$, $j = 1, \ldots, k$ i.e.

$$\Delta t \quad \mathsf{C} \Delta t_{FE}, \ \ \mathsf{C} := \min_{j=1,\ldots,k} \frac{\alpha_j}{\beta_j}$$

where C is the SSP-coefficient of the LMM. Clearly, the SSP LMM representation is unique ($\alpha_0 = 1$ fixed). Similarly for SSP RK methods, this gives severe restrictions for consistent (zero stable)

methods; it was shown in [12] that the maximal attainable order is strictly smaller than the number of steps:

Theorem 3.1 (Conditions for the order of SSP LMM, [12]). *For p 2, there is no p step, order p SSP explicit linear multistep method - considering arbitrary starting procedure - with all non-negative β_i coe cients.*

This can be seen for second order methods, where from the order conditions we have

$$\alpha_1 = \xi \quad 2, \ \alpha_2 = 1 \quad \xi, \ \beta_1 = \frac{\xi}{2} + 1, \ \beta_2 = \frac{\xi}{2} \quad 1$$

where for zero-stable methods ξ is arbitrary in the interval $(0, 2]$. Under the SSP conditions $\alpha_j \quad 0$, $\beta_j \quad 0$, $j = f1, 2g$, we get that $\xi = 2$, but in this case $C := \min f\frac{1}{2}, 0g = 0$, so there exist no second order 2-step LMM with positive SSP coefficient. In [11] and later in [17] the monotonicity and positivity preservation was analysed in the case if one fixes the starting procedure. They showed that the above theorem does not hold. To see this suppose we use the explicit Euler method as the starting procedure, i.e.

$$u_1 = u_0 + \Delta t f(u_0). \tag{3.24}$$

By introducing a constant θ, and adding and subtracting $\theta u_{n \ 1}$, we get:

$$u_n = \quad (\alpha_1 \quad \theta)u_{n \ 1} + \beta_1 \Delta t f(u_{n \ 1}) + \theta u_{n \ 1} \quad \alpha_2 u_{n \ 2} + \beta_2 \Delta t f(u_{n \ 2}).$$

Using the method to rewrite $\theta u_{n \ 1}$ and adding and subtracting $\theta^2 u_{n \ 2}$ we get:

$$\begin{aligned} u_n = \quad & (\alpha_1 \quad \theta)u_{n \ 1} + \beta_1 \Delta t f(u_{n \ 1}) \quad (\alpha_2 + \theta\alpha_1 + \theta^2)u_{n \ 2} + (\beta_2 + \theta\beta_1)\Delta t f(u_{n \ 2}) + \\ & + \theta^2 u_{n \ 2} + \theta(\quad \alpha_2 u_{n \ 3} + \beta_2 \Delta t f(u_{n \ 3})). \end{aligned}$$

Similarly adding and subtracting $\theta^j u_{n \ j}$ for $j = f3, \ldots, n \quad 3g$ and using the explicit Euler starting procedure (3.24):

$$\begin{aligned} u_n = \quad & (\alpha_1 + \theta)u_{n \ 1} + \beta_1 \Delta t f(u_{n \ 1}) \\ & + \sum_{j=2}^{n \ 1} \theta^{j \ 2}\big(\quad \alpha_2 \quad \theta\alpha_1 \quad \theta^2)u_{n \ j} \quad (\beta_2 + \theta\beta_1)\Delta t f(u_{n \ j})\big) \\ & + \theta^{n \ 2}\big((\theta \quad \alpha_2)u_0 + (\theta + \beta_2)f(u_0)\big). \end{aligned} \tag{3.25}$$

If $\theta > 0$ and all the coefficients are positive in (3.25), then positivity preservations holds under the step-size condition $0 < \Delta t \quad C(\theta)\Delta t_{FE}$, where

$$\begin{aligned} C(\theta) &:= \min\left\{A(\theta), B(\theta), C(\theta)\right\} \\ &:= \min\left\{\frac{\alpha_1 \quad \theta}{\beta_1}, \frac{(1 \quad \theta)(\theta \quad \alpha_2)}{\beta_2 + \theta\beta_1}, \frac{\theta \quad \alpha_2}{\theta + \beta_2}\right\} \end{aligned}$$

where we have used that $\alpha_1 = \quad 1 \quad \alpha_2$. For $0 < \xi \quad 2$, $A(\theta), B(\theta), C(\theta)$ are monotonic decreasing functions of θ. The minimal θ such that the coefficients in (3.25) are positive is $\theta_{min} = \max f\alpha_2, \ \frac{\beta_2}{\beta_1}, \ \beta_2 g = \quad \beta_2$. Then the optimal $C(\theta)$ is

$$\begin{aligned} \max_\theta C(\theta) &= \min\left\{A(\theta_{min}), B(\theta_{min}), C(\theta_{min})\right\} \\ &= \begin{cases} B(\theta_{min}) = \frac{\xi}{2 \ \xi} & \text{if } 0 < \xi \quad \frac{2}{3} \\ A(\theta_{min}) = \frac{2 \ \xi}{2+\xi} & \text{if } \frac{2}{3} \quad \xi \quad 2 \end{cases} \end{aligned}$$

From the considered two-step methods the SSP coefficient is maximal, when $\xi = \frac{2}{3}$ and it is $C = \frac{1}{2}$. In this case the boundedness property (3.10) also holds, since the coefficients in (3.25) add up to 1 ($\theta =$ β_2). In the literature, this method is called extrapolated BDF2 method:

$$u_n = \frac{4}{3}u_{n\ 1} \quad \frac{1}{3}u_{n\ 2} + \Delta t\left(\frac{4}{3}f(u_{n\ 1}) \quad \frac{2}{3}f(u_{n\ 2})\right). \tag{3.26}$$

Considering the barriers of SSP LMM methods, it can be shown that these methods does not suffer from order barriers - unlike RK SSP methods - so there exist an SSP LMM method with $C > 0$. In the case of arbitrary starting procedures there is a barrier considering C_{eff}, namely for explicit methods C_{eff} 1, while for implicit methods C_{eff} 2[12]. We point out that we had the same conditions for SSP RK methods.

3.4 Modified Patankar-Runge-Kutta methods

A popular family of methods which possess unconditional positivity and conservativity for a large class of systems are the Modified-Patankar-Runge-Kutta (MPRK) methods. These methods are based on the so-called *Patankar-trick* introduced in [18], but modified in the way that not only the source term is changed. These methods can be used for positive and fully conservative production-destruction systems.

Definition 3.4. *We call an ODE production-destruction system (PDS) if it can be represented in the following form:*

$$\dot{u}_i = \sum_{j=1}^{d} p_{i,j}(u) \quad \sum_{j=1}^{d} d_{i,j}(u), \quad (i = 1, \dots, d) \tag{3.27}$$

where $d_{i,j}, p_{i,j} : \mathbb{R}_+^d \ ! \ \mathbb{R}_+ \ [\ f0g$ *are functions such that* $d_{i,j}(u) = p_{j,i}(u)$ 0; $8i, j = 1, ..., d$; $8u \ 2$ $\mathbb{R}_+^d$. *The PDS is positive if* $8u_0 := u(0) > 0$ *initial value, the solution positive* $u(t) > 0$, $(8t > 0)$ *and fully conservative if* $p_{i,i} = d_{i,i} = 0$, $i = 1, \dots, d$.

In the case of chemical reactions $u_i(t)$ is the concentration of the i-th constituent, $p_{i,j}(u)$ is the rate at which the j-th constituent transforms into the i-th component, while $d_{ij}(u)$ is the rate at which the i-th constituent transforms into the j-th component. The first order MPRK method - originally introduced in [19] - called the modified-Patankar-Euler scheme has the following form:

$$u_{n+1,i} = u_{n,i} + \Delta t \sum_{j=1}^{d} \left(p_{ij}(u_n)\frac{u_{n+1,j}}{u_{n,j}} \quad d_{ij}(u_n)\frac{u_{n+1,i}}{u_{n,i}} \right), \quad (i = 1, \dots, d) \tag{3.28}$$

One can see that the method is the explicit-Euler method modified by step dependent weights in the form of $\frac{u_{n+1,j}}{u_{n,j}}$. These weights make the method semi-implicit, which means that the method can be rewritten in the matrix-vector product form

$$Au_{n+1} = u_n, \text{ where} \tag{3.29a}$$

$$a_{ii} = 1 + \Delta t \sum_{k+1}^{d} \frac{d_{i,k}(u_n)}{u_{n,i}}, \quad i = 1, \dots, d \tag{3.29b}$$

$$a_{ij} = \quad \Delta t \frac{p_{i,j}(u_n)}{u_{n,j}}, \quad i, j = 1, \dots, d, \ i \ 6= \ j \tag{3.29c}$$

where we have used that the original system is fully conservative (i.e. $p_{i,i} = 0$, $i = 1, \dots, d$). Because the matrix $A := fa_{ij}g_{i,j=1}^{d}$ does not depend on any of the coordinates of u_{n+1}, it can be solved by any system of linear equations solver. To see that the MPE possess the desired properties, we state the following

Theorem 3.2 ([19]). *The modified-Patankar-Euler scheme* (3.28) *used for fully conservative and positive PDS* (3.27) *is unconditionally positive and fully conservative.*

Proof. It is fully conservative since

$$\sum_{i=1}^{d}\left(u_{n+1,i} \quad u_{n,i}\right) = \Delta t \sum_{i=1}^{d}\left(\sum_{j=1}^{d} p_{i,j}(u_n)\frac{u_{n+1,j}}{u_{n,j}} \quad \sum_{j=1}^{d} d_{i,j}(u_n)\frac{u_{n+1,i}}{u_{n,i}}\right)$$
$$= \Delta t\left(\sum_{i,j=1}^{d} p_{i,j}(u_n)\frac{u_{n+1,j}}{u_{n,j}} \quad \sum_{i,j=1}^{d} p_{j,i}(u_n)\frac{u_{n+1,i}}{u_{n,i}}\right)$$

where we have used that $p_{i,j} = d_{j,i}$.

To show the positivity, first note that the off-diagonal elements (3.29c) in the matrix A are negative, while the diagonal elements (3.29b) are strictly positive. It is sufficient to show that A is a nonsingular M matrix, because then $A^{-1} \quad 0$. Then since A^{-1} is also nonsingular and non-negative at least one element in each row has to be positive then for $u_n \quad 0$, $u_n \neq 0$ we have $u_{n+1} \quad 0$, $u_{n+1} \neq 0$. Now we will show that A is a nonsingular M-matrix. Note that A is strictly diagonally dominant since for arbitrary $i \in \{1,\dots,d\}$:

$$|a_{i,i}| = 1 + \Delta t \sum_{k=1}^{d} \frac{d_{i,k}(u_n)}{u_{n,i}} d > \Delta t \sum_{k=1}^{d} \frac{p_{k,i}(u_n)}{u_{n,i}} = \sum_{k=1,k\neq i}^{d} (-a_{k,i}) = \sum_{k=1,k\neq i}^{d} |a_{k,i}|.$$

Since A is strictly diagonally dominant and with $g := e^t = (1,\dots,1)^t \in \mathbb{R}^d$ we have $Ag > 0$ (elementwise), then the matrix A is an M-matrix by definition. □

A second-order modified-Patankar-Runga scheme was introduced in [19], which was later generalized in [20] in the following way:

$$y^{(1)} = u_n$$
$$y_i^{(2)} = u_{n,i} + \Delta t\alpha \sum_{j=1}^{d}\left(p_{ij}(y^{(1)})\frac{y_j^{(2)}}{y_j^{(1)}} \quad d_{ij}(y^{(1)})\frac{y_i^{(2)}}{y_i^{(1)}}\right)$$
$$u_{n+1,i} = u_{n,i} + \Delta t \sum_{j=1}^{d}\left[\left((1 \quad \frac{1}{2\alpha})p_{ij}(y^{(1)}) + \frac{1}{2\alpha}p_{ij}(y^{(2)})\right)\frac{u_{n+1,j}}{(y_j^{(2)})^{\frac{1}{\alpha}}(y_j^{(1)})^{1 \quad \frac{1}{\alpha}}}\right.$$
$$\left.\left((1 \quad \frac{1}{2\alpha})d_{ij}(y^{(1)}) + \frac{1}{2\alpha}d_{ij}(y^{(2)})\right)\frac{u_{n+1,i}}{(y_i^{(2)})^{\frac{1}{\alpha}}(y_i^{(1)})^{1 \quad \frac{1}{\alpha}}}\right]$$

where $i = 1,\dots,d$. The above scheme is called MPRK22(α) method, where $\alpha \quad \frac{1}{2}$. The name comes from the result such that all second order two stage RK methods has the following Butcher-tableau

0		
α	α	
	1-$\frac{1}{2\alpha}$	$\frac{1}{2\alpha}$

Notice that for $\alpha = 1$ there is a simplification to

$$
\begin{aligned}
y^{(1)} &= u_n \\
y_i^{(2)} &= u_{n,i} + \Delta t \sum_{j=1}^{d} \left(p_{ij}(y^{(1)}) \frac{y_j^{(2)}}{y_j^{(1)}} + d_{ij}(y^{(1)}) \frac{y_i^{(2)}}{y_i^{(1)}} \right) \\
u_{n+1,i} &= u_{n,i} + \Delta t \sum_{j=1}^{d} \left[\left(\frac{1}{2} p_{ij}(y^{(1)}) + \frac{1}{2} p_{ij}(y^{(2)}) \right) \frac{u_{n+1,j}}{(y_j^{(2)})} \right. \\
&\quad \left. \left(\frac{1}{2} d_{ij}(y^{(1)}) + \frac{1}{2} d_{ij}(y^{(2)}) \right) \frac{u_{n+1,i}}{(y_i^{(2)})} \right]
\end{aligned}
$$

Similarly, the following holds

Theorem 3.3 ([20]). *The MPRK22(α) scheme used for fully conservative and positive PDS* (3.27) *is second order, unconditionally positive and fully conservative with unconditionally positive and fully conservative stage values.*

We won't reproduce the proof, because the positivity and conservativity is identical with the proof used for the MPE scheme, while we omit the proof of the consistency order by its length. In [21] third order MPRK methods were introduced. These methods are based on the 3 stage third-order explicit RK methods. Interestingly, similarly for SSP theory, to ensure the positivity of the methods, a necessary condition is that A 0 and β 0. The derived MPRK methods are called MPRK43(α, β) schemes, because one needs to solve an additional linear system.

These methods are widely used for ODE models of chemical reactions, where the positivity and conservativity of the constituents hold by the principles of law of conservation of mass. They can be also used for diffusion-convection-reaction systems, where one, after semi-discretisation and splitting uses one of the MPRK method for the reaction part. Such systems can be found for example in geobiochemical marine modelling[19]. Apart from the preservation properties, the scheme can be applied to stiff systems too[19]. Numerous numerical schemes have been modified by the Patankar-trick - the explicit scheme is weighted by some implicit stage - to get unconditionally positive and conservative schemes. In [22] two stage second order RK SSP methods in their Shu-Osher representation (3.5) were modified in such way. The main advantage of this scheme is that one can easily modify it for the semi-discretisation of convection-reaction systems, where the positivity of the convection part is preserved under the assumption that it is preserved for the explicit Euler method (instead of using splitting and solving the two part separately). By this, the positivity of the semi-discretised system is 'only' conditionally preserved, but the reaction can be stiff. In [23] third order SSP MPRK schemes were introduced.

3.4.1 Stability questions, problematic behaviours

The absolute stability of the MPRK methods has been recently studied in [24] and in [25]. The reason is that the usual Dahlquist test equation $\dot{u} = \lambda u$ cannot be used because it is not conservative and even for linear systems, the resulting iteration $u_{n+1} = g(u_n)$ is nonlinear. For these reasons, the following test equation was considered: $\dot{u} = Au$ where the matrix A 2 $\mathbb{R}^{d}$ ^{d} has the following properties:

- The matrix A is a Metzler matrix (i.e. the off-diagonal elements are positive), to ensure the positivity of the continuous system (see theorem 2.3).
- The matrix A possesses at least one linear invariants i.e. there exists $\boldsymbol{n}_1, \ldots, \boldsymbol{n}_k$ 2 $\mathbb{R}^{d}$ $f0g$ such that $\boldsymbol{n}_i^t A = 0$, to ensure that $\boldsymbol{n}_i$ $u(t) = \boldsymbol{n}_i$ $u(0)$. Note that the conservativity is a linear

invariant with $\boldsymbol{n} = e$. Also note that this implies that $\boldsymbol{n}_1, \ldots, \boldsymbol{n}_k$ are (left) eigenvectors of the eigenvalue $\lambda = 0$.

- All non-zero eigenvalues of A has negative real part and the eigenvalues with zero real part have Jordan block size 1, to ensure the (Lyapunov) stability of the system.

We will denote the generated map by $g : \mathsf{R}^d \; ! \; \mathsf{R}^d$ (i.e. $u_{n+1} = g(u_n)$). By the nonlinear nature of the generated map g and since the linearization of g around the steady state is non-hyperbolic, center manifold theory was used. It was shown [25, Thm. 2.9], that despite its non-hyperbolic nature, the stability is determined by the eigenvalues of the linearized map, namely the stability is preserved if the eigenvalues which are not one, are smaller than one. This works, since after affine transformation (one of) the central manifold is locally the zero function, which simplifies the reduced system considerably. It was also shown that since MPRK preserves all linear invariants the stability in the above case is asymptotically stable in the subspace defined by the invariants.

Two problematic behaviour of the MPRK schemes were discussed in the literature. One is that in some cases, MPRK scheme can give rise to spurious oscillatory behaviour even for linear systems[26]. This is not specific to MPRK schemes; it is well known that linear schemes can give rise to spurious oscillatory behaviour also. In [26],[27] that for two dimensional systems the non-oscillatory behaviour can be guaranteed for small enough step-sizes depending on the scheme. The other problematic behaviour is the order reduction for initial values close to 0, which was analysed for linear models in [26].

3.5 Regularity of numerical methods, stability of equilibria

Another question considering the dynamics under the discretisation by different numerical methods is whether the continuous model has the same equilibria as the discrete map (3.1). We will see that this is not the case for many linear methods. We denote the set of the equilibria of (2.1) and of (3.1) as F and F_t, respectively. ($\mathsf{F}_t := \{u \in \mathsf{R}^n : \Phi_{f,t}(u, \ldots, u) = u\}$). The question was first studied in [28].

For LMMs, if we suppose that $u \in \mathsf{F}_t$, then by consistency we have that $\sum a_k = 0$ and $\sum b_k \neq 0$, so $f(u) = 0$ i.e. $u \in \mathsf{F}$. On the other hand, $u_k \equiv u \in \mathsf{F}$ implies $\Phi_{f,t}(u, \ldots, u) = u$ by consistency. In conclusion, for (consistent) linear multistep methods $\mathsf{F} = \mathsf{F}_t$ for all $\Delta t > 0$. It is also clear that an irregular RK method as a starting procedure of an LMM does not alter the equilibria.

For Runge-Kutta methods, if $u \in \mathsf{F}$, then $\Phi_{f,t}(u) = u$ holds with the choice $k_i = 0$ for all $i = 1, \ldots, s$. So $\mathsf{F} \subseteq \mathsf{F}_t$ holds by the supposed uniqueness of the solution. Hairer et al. gave conditions on the RK methods for $\mathsf{F} = \mathsf{F}_t$ and they named these methods *regular* [29]. They showed that for regular (s-stage) methods one can construct an $s - 1$ stage RK method which preserves the regularity (Thm. 3.). From the exact construction, it follows easily that the only explicit regular method is the explicit-Euler. This construction also gives an algorithm to determine the regularity of any RK method. For methods of order $p \geq 2$, they also showed that a necessary condition for regularity is that the trace of the matrix A is $\frac{1}{2}$ (this is also sufficient for $s = 2$) and there is no regular A-stable method with order larger than 4 (Thm. 7.,11.). One advantage of the implicit RK methods, that is the order compared to the number of stages is large, does not hold for regular methods because:

Theorem 3.4 ([29], Barriers of regular RK methods)**.** *The order p of a regular s stage RK method*

satis es

$$p \quad s+2 \quad \textit{if } s \textit{ is even}$$
$$p \quad s+1 \quad \textit{if } s \textit{ is odd}$$

It should be noted that regularity does not imply the non-existence of spurious periodic solutions, a well-known example for the latter is the period-doubling behaviour of the explicit-Euler discretised logistic equation[30]. The full characterization for LMM for the existence of spurious 2-cycles are known, but in general these conditions are strict, so one puts some condition on the function f to get less strict conditions[31]. This shows the well-known fact that one has to choose a preferred numerical method (partly) in a problem-driven way. We also want to point out, that conditional properties are sufficient, since the stability is also conditional for most of the methods.

To obtain similar dynamics for the numerical maps, it is also required that the asymptotic behaviour of the equilibria of (3.1) is the same as that of the equilibria of the continuous model. This holds for the limit $\Delta t \; ! \; 0$ by convergence, but might not hold for arbitrary $\Delta t > 0$.

For discrete maps, one can define the stability of an equilibrium as it was done in the continuous case (def. 2.1) with straightforward alteration. For simplicity we define it for a general scheme in the form

$$u_{n+1} = u_n + \Delta t F_{\ t}(u_n)$$

The explicit Euler method and the RK method is in this form. Note that in this case u is an equilibrium if and only if $F_{\ t}(u\ ) = 0$. To find the stability of an equilibrium, one uses a similar technique as in (2.2)-(3.32) for the continuous system. Namely, we perturb the fixed point u of the discrete map $y_n = u_n \quad u$, which has a Taylor expansion at the equilibrium u

$$y_{n+1} = u_n \quad u \ + \Delta_t F_{\ t}(y_n + u\ ) = y_n + \Delta t F_{\ t}^{\prime}(u\ )y_n + r(y_n)$$

Where we have used that $F_{\ t}(u\ ) = 0$ and $F_{\ t}$ is sufficiently smooth. For small enough y_n, the linear part dominates, so we obtain the approximate linear system:

$$y_{n+1} = \big(1 + \Delta t F^{\prime}{}_{t}(u\ )\big)y_n \tag{3.32}$$

It is clear that for one dimensional systems, the linearised system has an asymptotic stable equilibrium if $j\big(1 + \Delta t F^{\prime}{}_{t}(u\ )\big)j < 1$ and unstable if it is strictly larger than one. Similarly, for d dimensional systems, it is asymptotically stable if all eigenvalues of $1 + \Delta t F^{\prime}{}_{t}(u\ )$ are inside the unit disk, while unstable if there is an eigenvalue has strictly larger modulus than one. Note that, we excluded that case, when an eigenvalue lie on the unit circle. In this case the equilibrium of the nonlinear system is called non-hyperbolic, and we have to consider the larger order terms of the Taylor expansion. It is clear that for the explicit Euler method $F_{\ t} = f$, so if λ is an eigenvalue of $f^{\prime}(u\ )$, then $1 + \Delta t f^{\prime}(u\ )$ is an eigenvalue for the method.

From the absolute stability theory it is clear that the preservation of equilibria is a step-size and problem dependent question. The existence of irregular RK methods motivates and complicates this question, even in the case if the spurious equilibrium is unstable, because it may happen that this unstable equilibrium has an unstable manifold which connects to infinity, so the boundedness property of the solutions of the IVP can get lost[31].

4 Epidemiological models

In the following two subsections, we will introduce and analyse two epidemiological compartmental models. The first is a more simple model, where the total population is constant, while this does not hold for the second model.

4.1 conservative model

Vertical transmission occurs when a newborn (or unborn) is infected via its parents, while horizontal transmission is when a person is infected by physical contacts or through droplets etc. Diseases where vertical transmission can occur are AIDS, Hepatitis B, herpes simplex virus, Keystone virus etc. It should be noted that one can define vertical and horizontal transmission for animal and plant diseases similarly. The following model is an extension of the Kermack-McKendrick model (1.1). To incorporate the possibility of vertical transmission, one has to consider a model with demography/changing population. The population is considered to be asexual or it is only the female subpopulation. The immunity is considered to be non-permanent and a fraction of the newborn population gets vaccinated (or every newborn gets vaccinated but the vaccine is imperfect). It is also assumed that the vaccine does not create immunity in those born of infected parents and there is no death by the disease. Under these assumptions, the model is

$$\frac{dS}{dt} = \quad kSI + (1 \quad m)b(S+R) + pb^{0}I \quad rS + \varphi R \tag{4.1a}$$

$$\frac{dI}{dt} = kSI + qb^{0}I \quad r^{0}I \quad vI \tag{4.1b}$$

$$\frac{dR}{dt} = vI \quad rR + mb(S+R) \quad \varphi R \tag{4.1c}$$

where S, I, R denotes the susceptible, infected and recovered subpopulation, respectively. The above model was introduced in [2] and should be considered as some test model, since there are better - though more complex - models for the above mentioned diseases. A flowchart can be found in 1. The parameters are strictly positive and they denote the following:

Parameters	
b	Birth rate of uninfected individuals
b^{0}	Birth rate of infected individuals
r	Death rate of uninfected individuals
r^{0}	Death rate of infected individuals
v	Recovery rate
φ	Rate of immunity loss
$q \in (0,1)$	Rate of vertical transmission $q + p = 1$
$m \in (0,1)$	Fraction born vaccinated (or vaccine effectiveness)

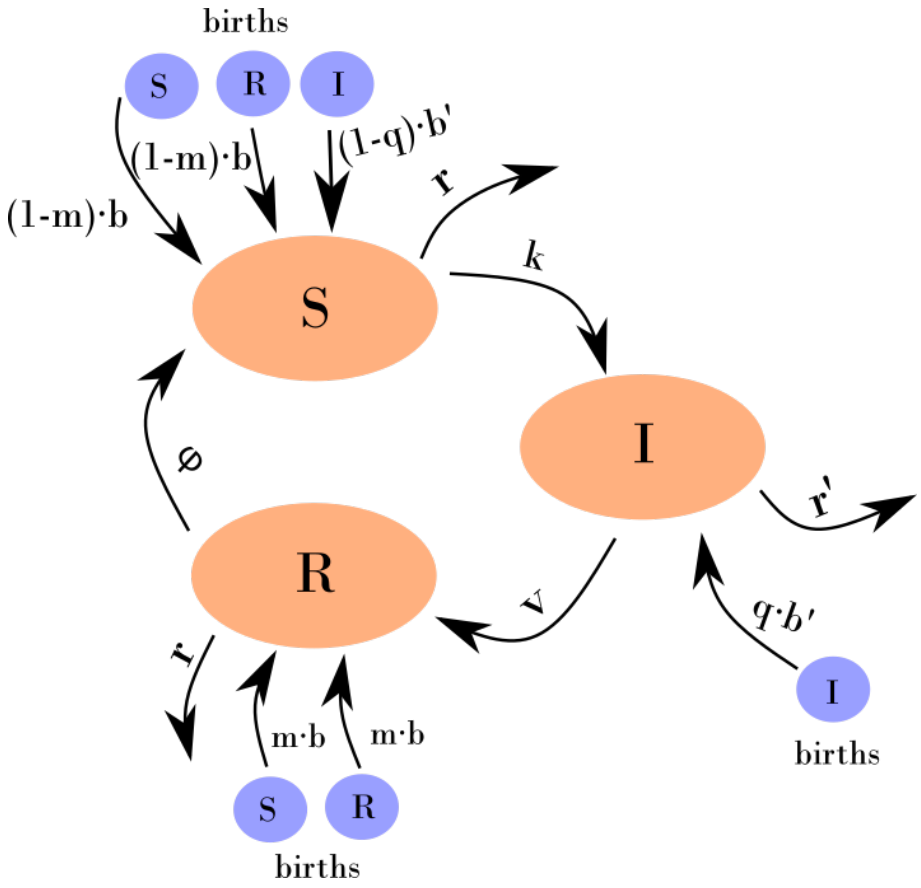

Figure 1: Flowchart of the conservative epidemiological model (4.1a):(4.1c)

We will denote the total population at time $t \quad 0$ as $N(t)$. If $b = b^0$ and $r = r^0$, then the population remains constant, we will assume this with $N \quad N_0 = 1$ from so on, but the calculations can be easily carried out to other or arbitrary N_0. We can rewrite the system in the following reduced form:

$$\frac{dS}{dt} = \quad kSI + (1 \quad m)b(N_0 \quad I) + pb^0 I \quad rS + \varphi(N_0 \quad I \quad S) \tag{4.2a}$$

$$\frac{dI}{dt} = kSI + qb^0 I \quad r^0 I \quad vI \tag{4.2b}$$

We denote the biologically feasible region for the above system as $\Omega \quad = f(s,i) \; 2 \; \mathsf{R}^2 j \, 0 \quad s,i \quad 1g$, which is positively invariant since the positivity follows from theorem 2.3 and the boundedness from the constant population. The solutions with initial values from Ω exists for all $t \quad 0$ by the positive invariance of Ω and by the system (4.2a)-(4.2b) being in C^1 by its polynomial structure. The Disease-free equilibrium can be calculated by letting LHS in (4.2a)-(4.2b) be zero and $I = 0$:

$$E_0 := (S_0, I_0) = \left(\frac{(1 \quad m)b + \varphi}{\varphi + b}, 0 \right)$$

while the endemic equilibrium can be calculated by letting LHS in (4.2a)-(4.2b) be zero:

$$E \; := (S \; , I \;) = \left(\frac{pb^0 + v}{k}, \frac{((1 \quad m)b + \varphi)k \quad (b + \varphi)(pb^0 + v)}{(\varphi + (1 \quad m)b + v)k} \right) \tag{4.3}$$

The Jacobian of (4.2a)-(4.1b) at the disease free equilibrium E_0 is

$$\begin{pmatrix} r & \varphi & kS_0 & (1 \quad m)b + pb^0 & \varphi \\ 0 & & kS_0 & pb^0 & v \end{pmatrix}$$

Thus the system is asymptotically stable if the eigenvalues are negative i.e.

$$r \quad \varphi < 0 \quad and \tag{4.4}$$

$$kS_0 \quad pb^0 \quad v < 0 \tag{4.5}$$

where the first condition always holds and the second condition is equivalent with

$$R_0 := k\frac{(1 \quad m)b + \varphi}{(\varphi + b)(pb^0 + v)} < 1$$

where R_0 is called the *Basic Reproduction Number*. From (4.3), we can see that this is equivalent with that there is no infected subpopulation (since $I \ < 0$), while when $R_0 > 1$, then there is an equilibria in the positive orthant ($I \ > 0$).

The proof of the global stability of E can be found in [4]. We simplified the proof to omit the graph-based arguments (Lemma 2.2.(a)).

Theorem 4.1. *The endemic equilibrium* (4.3) *of the system* (4.1a)-(4.1c) *is GAS (within* Ω *).*

Proof. One can rewrite the system (4.2a)-(4.2b) as:

$$\dot{z} = F(z) := \mathrm{diag}(z)(e + Az) + Bz + c \tag{4.6}$$

where $z = (S, I)$ and

$$A = \begin{pmatrix} 0 & k \\ k & 0 \end{pmatrix}; B = \begin{pmatrix} 0 & (m \quad 1)b + pb^0 & \varphi \\ 0 & 0 \end{pmatrix}; e = \begin{pmatrix} (b & \varphi) \\ (v & pb^0) \end{pmatrix}; c = \begin{pmatrix} (1 & m)b + \varphi \\ 0 \end{pmatrix}$$

and denote $b(z) := Bz + c$. Let $E \ = z \ > 0$ (element-wise), so $F(z \ ) = 0 = \mathrm{diag}(z \ )(e + Az \ ) + Bz \ + c$. After rearranging to e, we get that

$$e = \quad Az \quad \mathrm{diag}(1/z \ )b(z \ )$$

After the substitution of e to the ODE 4.6:

$$\dot{z} = \mathrm{diag}(z)\Big(A + \mathrm{diag}(1/z \ )B\Big)(z \quad z \ ) \quad \mathrm{diag}(z \quad z \ )\mathrm{diag}(1/z \ )b(z). \tag{4.7}$$

We define the Lyapunov function

$$V(z) := \sum_{i=1}^{d} w_i\big(z_i \quad z_i \quad z_i \, ln\frac{z_i}{z_i}\big) \tag{4.8}$$

where $w_i > 0$, $i = 1, \ldots, d$ are real constants. So $V(z \ ) = 0$ and V maps from the strictly positive quadrant to R^+ (see fig 2). Differentiating V alongside the solution of (4.7) we have

$$\dot{V}(z) = (z \quad z \ )^{\mathsf{I}} W\tilde{A}(z \quad z \ ) \quad \sum_{i=1}^{d} w_i \frac{b_i(z)}{z_i z_i}(z_i \quad z_i)^2 \tag{4.9}$$

where $\tilde{A} := A + diag(1/z \ )B$ and $W := \mathrm{diag}(w_1, \ldots, w_d)$ strictly positive matrix. Now, if we can choose W such that $\tilde{A}W$ is antisymmetric, then (4.9) simplifies to

$$\dot{V}(z) = \quad \sum_{i=1}^{d} \frac{w_i b_i(z)}{z_i z_i}(z_i \quad z_i)^2.$$

Since in our case

$$\tilde{A} = \begin{pmatrix} 0 & \frac{k[(m \quad 1)b \quad \varphi \quad v]}{pb^0 + v} \\ k & 0 \end{pmatrix} \tag{4.10}$$

where $\tilde{a}_{12} < 0$ since all parameters are positive and $m \in (0, 1]$. Thus we can choose $W := \text{diag}(1, \quad 1/\tilde{a}_{12}) > 0$ to make $\tilde{A}W$ antisymmetric.

If we assume that $b(z)$ 0 - which holds in our specific case - then $\dot{V}(z)$ 0 i.e. z is stable Ω . To show that the stability is asymptotic in Ω ,thus z is GAS, we use the LaSalle invariance principle (Thm. (2.2)). For this, — denote $E := \{z \in \Omega | \dot{V}(z) = 0\}$, and the largest invariant subset of E by M. We show that M is the singleton $\{z\}$. Since $b_1 > 0$, we must have $z_1 = z_1$ for the solution which are in M, which implies that $\dot{z}_1 = 0$. Using this for the equation (4.7), we get that for the solutions in M, it must hold that $\tilde{a}_{12}(z_2 \quad z_2) = 0$, and since $\tilde{a}_{12} \neq 0$, it implies that we also must have $z_2 = z_2$. In other words $M = \{z\}$, and the global asymptotic stability follows by the LaSalle invariance principle (Thm. (2.2)) since the solutions starting in Ω stay there by the positive invariance of the set, which implies that they are also bounded. □

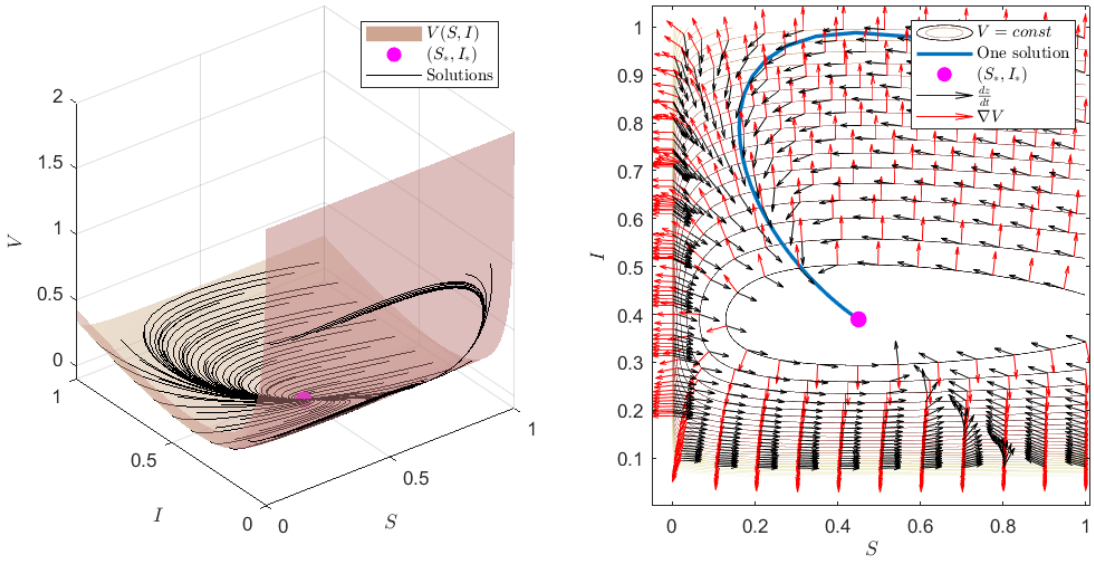

Figure 2: Left: The Lyapunov function (4.8) for some specific values and the phase portrait. Right: The contours of the Lyapunov function, its gradient and the vector field specified by the system (4.1a)-(4.1c). From the figure, it can be seen that the angle between the gradient of V and the vector field f is always obtuse implying $\dot{V}(z(t))$ 0

4.2 Non-conservative model

In 2020, Yang and Wang proposed the following model to investigate the early days of the epidemic of COVID-19 in Wuhan, China, with incorporation of the possibility that the environment is a possible transmission route (besides the infected people)[32]. The reason for including the environmental reservoir as a possible transmission route was that officials received a positive result when they collected samples from the Huanan Seafood Market area. In addition, some studies suggest that the virus can survive on different surfaces such as metal, glass, and plastic for up to 9 days[33][34]. By fitting the outbreak data to the proposed model, they found that the environmental reservoir had a significant contribution to the overall infection risk[32]. We have modified their proposed model to include a class with infected but not infectious subpopulation. We have also included a class with imperfect vaccination, which means that vaccinated people can become infected also. We have made the following assumptions:

1.A1 There is always an infected but non-infectious phase.

1.A2 Vaccination is imperfect w.r.t infectious individuals and the environment, but in general for the vaccinated subpopulation to become infected at a lower rate.

1.A3 The imperfection of the vaccine is the same against infected people and the environment.

1.A4 A vaccinated person can lose immunity.

Our proposed model:

$$\frac{dS}{dt} = \Lambda \quad \beta_I SI \quad \beta_V SV + \Psi C + \delta R \quad (\chi + \mu)S \tag{4.11a}$$

$$\frac{dE}{dt} = \beta_I SI + \beta_V SV + \rho\beta_I CI + \rho\beta_V CV \quad (\alpha + \mu)E \tag{4.11b}$$

$$\frac{dI}{dt} = \alpha E \quad (\gamma + \omega + \mu)I \tag{4.11c}$$

$$\frac{dR}{dt} = \gamma I \quad (\mu + \delta)R \tag{4.11d}$$

$$\frac{dC}{dt} = \chi S \quad \rho\beta_I CI \quad \rho\beta_V CV \quad (\Psi + \mu)C \tag{4.11e}$$

$$\frac{dV}{dt} = \xi I \quad \sigma V \tag{4.11f}$$

where $S(t)$, $E(t)$, $I(t)$, $R(t)$, $C(t)$ are the number of susceptible, exposed (infected but not yet infectious), infected (infectious), recovered, and vaccinated at time instance t, respectively. V represents the environmental reservoir and is integrated to the model to include the possibility that a susceptible individual may acquire the disease through the environment and not directly by susceptible-infectious contacts. Note that there are no space variables, so the virus concentration in the environment is assumed to be homogeneous (e.g. possibly a city). A flowchart can be seen in 5. All the parameters are non-negative and their "meaning" can be seen in the table. By assumption [1.A2] ρ *2* $(0, 1)$.

Parameters

Λ	Population influx
μ	Natural death rate
ω	Disease induced death rate
$1/\alpha$	Mean incubation period
γ	Recovery rate
$1/\delta$	Mean-time spent in the recovered class
β_I	Transmission rate by infected individual
β_V	Transmission rate by the environmental reservoir
1 ρ	Vaccine effectiveness
χ	Vaccination rate of the susceptible class
Ψ	Rate of the vaccination loss
ξ	Rate of the exposed individuals contributing the virus to the environment
σ	Rate of (natural and artificial) removal of the virus from the environment

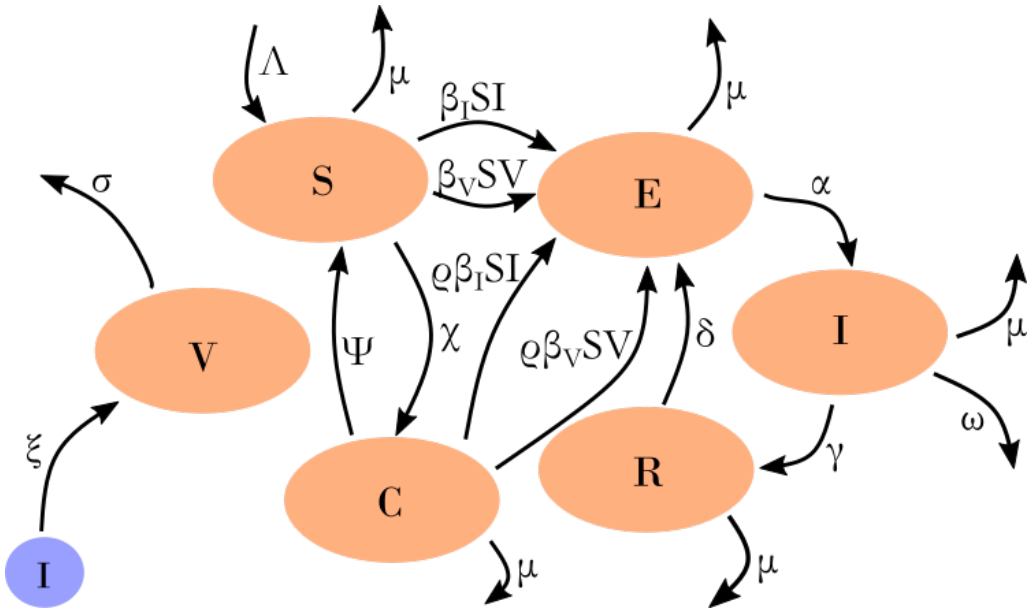

Figure 3: Flowchart of the non-conservative epidemiological model (4.11a)-(4.11f)

The disease free equilibrium (DFE) can be obtained by setting all the derivatives in (4.11a)-(4.11f) to 0 and also E, I, V equal to zero (i.e. no infections in the population):

$$F_0 := (S_0, E_0, I_0, R_0, C_0, V_0) = \left(\frac{\Lambda(\Psi+\mu)}{\mu(\Psi+\chi+\mu)}, 0, 0, 0, \frac{\Lambda\chi}{\mu(\Psi+\chi+\mu)}, 0\right). \tag{4.12}$$

For the endemic equilibrium when $\rho \neq 1$, we get a quadratic function for I, When $\rho = 0$, the function reduces to a linear function.

We will compute R_0 for the system by the next generation approach introduced in 1.1: The infection components for the model (4.11a)-(4.11f) are E, I, V. Rewriting the model as:

$$\begin{aligned} \dot{x}_i &= F_i(x,y) \quad V_i(x,y) \quad i = 1,2,3 \\ \dot{y}_j &= g_j(x,y) \quad j = 1,2 \end{aligned} \tag{4.13}$$

where $(x_1, x_2, x_3) = (E, I, V)$, $(y_1, y_2, y_3) = (S, R, C)$ where

$$F = \begin{pmatrix} \beta_E SI + \beta_V SV + \rho\beta_I SI + \rho\beta_V SV \\ 0 \\ 0 \end{pmatrix} \qquad V = \begin{pmatrix} (\alpha+\mu)E \\ \alpha E + (\gamma+\omega+\mu)I \\ \xi I + \sigma V \end{pmatrix}.$$

The Jacobi matrices of the subsystems F and V at the disease free equilibrium $(0, y_0) = (0, 0, 0, S_0, R_0, C_0)$

$$F := F^0(X_0) = \begin{pmatrix} 0 & \beta_I S_0 + \rho\beta_I C_0 & \beta_V S_0 + \rho\beta_V C_0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad V := V^0(X_0) = \begin{pmatrix} \alpha+\mu & 0 & 0 \\ \alpha & w+\gamma+\mu & 0 \\ 0 & \xi & \sigma \end{pmatrix}.$$

Then, the next generation matrix is $K = FV^{-1}$, which is an upper triangular matrix, so its spectral radius is

$$\begin{aligned} \rho(K) = R_0 &= \frac{\alpha\beta_I S_0}{(\alpha+\mu)(\gamma+\omega+\mu)} + \frac{\alpha\rho\beta_I C_0}{(\alpha+\mu)(\gamma+\omega+\mu)} + \frac{\beta_V S_0 \xi\alpha}{(\alpha+\mu)(\gamma+\omega+\mu)\sigma} + \frac{\rho\beta_V C_0 \xi\alpha}{(\alpha+\mu)(\gamma+\omega+\mu)\sigma} \\ &= R_0^1 + R_0^2 + R_0^3 + R_0^4. \end{aligned} \tag{4.14}$$

It is important to check whether R_0 can indeed be interpreted as some secondary infection. In our case it can be interpreted as the expected number of secondary infections produced in compartment E by an infected individual originally in compartment E:

- R_1 is the number of the secondary infections in the susceptible subpopulation of the initially exposed individual in his/her infectious stage, as the ratio $\frac{\alpha}{\alpha+\mu}$ is the proportion of individuals that progress from E to I and one infectious individual causes $\frac{\beta_I S_0}{w+\gamma+\mu}$ secondary infections in the susceptible subpopulation through his/her infectious stage. Similarly, R_2 is the number of the secondary infections in the vaccinated subpopulation of the initially exposed individual in his/her infectious stage.

- $R_0^3 + R_0^4$ is the secondary infections by the environment from the initially exposed individual. R_0^3 is the fraction of initially exposed individuals that progress to V through I ($\frac{\alpha}{\alpha+\mu}\frac{\xi}{w+\gamma+\mu}$) causing $\beta_V S_0$ number of new infections in $\frac{1}{\sigma}$ time. Similarly, R_0^4 can be interpreted for the vaccinated subpopulation.

Note that by setting $\xi = 0$, the environmental disease-route disappears.
We will show that there exist a positively invariant biologically feasible invariant set:

$$\Omega = \left\{ S, E, I, R, C, V \; 2 \; \mathsf{R}^{+} : S + E + I + R + C \quad \frac{\Lambda}{\mu}; V \quad \frac{\xi}{\omega}\frac{\Lambda}{\mu} \right\} \quad \mathsf{R}_{+}^{6}$$

We will show this through the positivity and boundedness of the solutions.

Theorem 4.2 (The proposed epidemic model positive)**.** *The system* (4.11a)-(4.11f) *is positive in the sense of* (2.3)*.*

Proof. Because the system (4.11a)-(4.11f) is clearly in C^1 since it has a polynomial structure, following theorem 2.3 the positivity is equivalent with the condition that the sign of the derivatives at the boundary points are non-negative (i.e. the solutions are reflected from the boundary), that is: $f_i(u_1, \ldots, u_{i\ 1}, 0, u_{i+1}, \ldots, u_d) \quad 0, \quad 8i \; 2 \; f1, \ldots, dg$. By this, the positivity of (4.11a)-(4.11f) follows since the parameter values are non-negative. For example, for E:

$$\beta_I SI + \beta_V SV + \rho\beta_I CI + \rho\beta_V CV \quad 0 \quad (8 \; S, I, R, C, V \; 2 \; [0, 1 \;))$$

□

Theorem 4.3 (Ω is positively invariant)**.** *The system* (4.11a)-(4.11f) *is positively invariant on* Ω*, that is, with initial conditions in* Ω *the solutions stays in* Ω *for arbitrary* $t \quad 0$*.*

Proof. Let $N(t)$ denote the total population at an arbitrary time instance t: $N(t) := S(t) + E(t) + I(t) + R(t) + C(t)$, which by assumption $N(0) \quad \frac{}{\mu}$ and from the system (4.11a)-(4.11f) $N^0(t) = \Lambda \quad \mu N(t) \quad \omega I(t)$. By multiplying both sides by $e^{\mu t}$, we get that

$$(N(t)e^{\mu t})^0 = (\Lambda \quad \omega I(t))e^{\mu t}$$

After integration from 0 to t:

$$\begin{aligned} N(t) &= N(0)e^{\ \mu t} + e^{\ \mu t}\int_0^t \big(\Lambda \quad \omega I(s)\big)e^{\mu s}ds \\ &= N(0)e^{\ \mu t} + \frac{\Lambda}{\mu}(1 \quad e^{\ \mu t}) \quad \omega\int_0^t I(s)ds \\ &\quad N(0)e^{\ \mu t} + \frac{\Lambda}{\mu}(1 \quad e^{\ \mu t}) = e^{\ \mu t}\big(N(0) \quad \frac{\Lambda}{\mu}\big) + \frac{\Lambda}{\mu} \\ &\quad \frac{\Lambda}{\mu} \end{aligned}$$

where we have used the non-negativity of $I(t)$ and the parameter ω. Similarly, for $V(t)$:

$$V'(t) + \sigma V(t) = \xi I(t)$$
$$(e^{\sigma t} V(t))' = e^{\sigma t} \xi I(t)$$
$$V(t) = V(0)e^{-\sigma t} + e^{-\sigma t}\xi \int_0^t I(s)ds \leq V(0)e^{-\sigma t} + e^{-\sigma t}\frac{\xi}{\sigma}\frac{\Lambda}{\mu}(e^{\sigma t} - 1)$$
$$= e^{-\sigma t}\left(V(0) - \frac{\xi}{\sigma}\frac{\Lambda}{\mu}\right) + \frac{\xi}{\sigma}\frac{\Lambda}{\mu} \leq \frac{\xi}{\sigma}\frac{\Lambda}{\mu}$$

where besides the non-negativity of I, we also used its boundedness property. □

Note that since Ω is positively invariant, the solutions of the system with initial values in Ω exist for all $t \geq 0$ (see section 2).

We also want to obtain stability conditions on the disease free equilibrium and the endemic equilibrium(s). Van den Driessche et al. showed that the endemic equilibrium is asymptotically stable under some assumptions on F, V and g in (4.13)[5]. Most of these assumptions are not strict and follows from the logic of endemic modelling. These conditions hold for our model, except assumption $A4$, but that only used to show that V is an M-matrix, which holds (and can be checked directly by calculating V^{-1}). In conclusion, we can state the following theorem for our model:

Theorem 4.4 (Stability of the DFE). *If $R_0 < 1$, then the DFE E_0 for the system (4.11a)-(4.11f) is locally asymptotically stable, while for $R_0 > 1$ it is unstable.*

To get stability on the endemic equilibria, we use the following theorem from [35]:

Theorem 4.5 (Condition on backward bifurcation[35]). *Consider the system of ODEs with parameter ϕ:*

$$\frac{dx}{dt} = f(x;\phi), \quad f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n, \quad f \in C^2(\mathbb{R}^n \times \mathbb{R}),$$

where 0 is an equilibrium for the system for all ϕ. Assume that

CCS-A1 *Denote $A := D_x f(0,0) = (\frac{\partial f_i}{\partial x_j}(0,0))$. Assume that zero is a simple eigenvalue of A, and all the other eigenvalues have negative real part.*

CCS-A2 *The matrix A for the eigenvalue 0 has a non-negative right eigenvector w and left eigenvector v.*

Let

$$a := \sum_{k,i,j}^{n} v_k w_i w_j \frac{\partial^2 f_k}{\partial x_i \partial x_j}(0,0)$$
$$b := \sum_{k,i}^{n} v_k w_i \frac{\partial^2 f_k}{\partial x_i \partial \phi}(0,0)$$

Then the local dynamics of the system is fully determined by the signs of a and b, specially:

case i. $a > 0$, $b > 0$. *When $\phi < 0$ with $|\phi|$ $\ll 1$, 0 is locally asymptotically stable, and there exists a positive unstable equilibrium; when $0 < \phi \ll 1$, 0 is unstable and there exists a negative and locally asymptotically stable equilibrium;*

case ii . $a < 0$, $b < 0$. *When $\phi < 0$ with $|\phi|$ $\ll 1$, 0 is unstable; when $0 < \phi \ll 1$, 0 is locally asymptotically stable, and there exists a positive unstable equilibrium;*

case iii. $a > 0$, $b < 0$. *When* $\phi < 0$ *with* $|\phi|$ $\ll 1$, 0 *is unstable, and there exists a locally asymptotically stable negative equilibrium; when* $0 < \phi$ $\ll 1$, 0 *is stable, and a positive unstable equilibrium appears;*

case iv. $a < 0$, $b > 0$. *When* ϕ *changes from negative to positive,* 0 *changes its stability from stable to unstable. Correspondingly, a negative unstable equilibrium becomes positive and locally asymptotically stable.*

This theorem is based on center manifold theory. From the assumptions, one can conclude that the center manifold is one-dimensional. After decomposing the center manifold into parts in the center and stable eigenspaces, the 'part' in the center eigenspace $c(t)$ can be approximated locally by $\frac{dc(t)}{dt} = \frac{a}{2}c^2 + b\phi c$. By the above theorem one can check whether forward or backward bifurcation occurs at $R_0 = 1$. For forward bifurcation, the DFE and the endemic equilibrium changes their stability, this is the *case iv*. While in the case of backward bifurcation, there is an interval for R_0, where a stable and unstable endemic equilibria coexist with a stable DFE. In the above theorem, this is *case i*.

Theorem 4.6. *The system* (4.11a)-(4.11f) *exhibits forward bifurcation at* $R_0 = 1$ *if*

$$\frac{\delta\gamma}{(\delta+\mu)} > \frac{(\alpha+\mu)(\gamma+\mu+\omega)(\mu+\Psi+\rho(\chi+2\mu))}{\alpha(\mu+\Psi+\chi\rho)}, \tag{4.15}$$

otherwise it exhibits backward bifurcation at $R_0 = 1$.

Proof. We will use the above theorem for the DFE E_0, with the parameter $\phi := \Lambda$. Λ is the critical value obtained from $R_0 = 1$:

$$\Lambda = \frac{\sigma\mu(\alpha+\mu)(\gamma+\omega+\mu)(\Psi+\chi+\mu)}{\alpha(\Psi+\mu+\rho\chi)(\sigma\beta_I+\xi\beta_V)}.$$

The matrix of the linearized system at (E_0, Λ) is

$$A := \begin{pmatrix} (\chi+\mu) & 0 & \beta_I S_0 & \delta & \Psi & \beta_V S_0 \\ 0 & (\alpha+mu) & \beta_I S_0 + \rho\beta_I C_0 & 0 & 0 & \beta_V S_0 + \rho\beta_V C_0 \\ 0 & \alpha & (\gamma+\omega+\mu) & 0 & 0 & 0 \\ 0 & 0 & \gamma & (\mu+\delta) & 0 & 0 \\ \chi & 0 & \rho\beta_I C_0 & 0 & (\Psi+\mu) & \rho\beta_V C_0 \\ 0 & 0 & \xi & 0 & 0 & \sigma \end{pmatrix}$$

where $S_0 = \frac{(\ +\mu)}{\mu+\chi+\mu}$ and $C_0 = \frac{\chi}{\mu+\chi+\mu}$.

The matrix A has a simple zero eigenvalue, what can be checked directly. The remaining eigenvalues cannot be easily calculated, but we only need to check their signs. This can be done by using the Hurwitz criterion (using the characteristic polynomial of the reduced system, i.e. without the 0 root). To check the signs of the determinant of the minor matrices of the Hurwitz matrix, I wrote a simple (symbolic) MATLAB code. From the results, we can conclude that the other eigenvalues have negative real parts.

One left eigenvector for the 0 eigenvalue is

$$v = \left(0, 1, \frac{\alpha+\mu}{\alpha}, 0, 0, \frac{\beta_v(\alpha+\mu)(\gamma+\omega+\mu)}{\alpha(\beta_V\xi+\beta_I\sigma)}\right),$$

which has non-negative entries. After some algebraic manipulation, we get that one right eigenvector for the 0 eigenvalue is:

$$w = \left(\frac{\rho(\alpha+\mu)(\gamma+\mu+\omega)}{\alpha(\mu+\Psi+\chi\rho)} + \frac{\Psi+\mu}{\mu(\chi+\mu+\Psi)}q, \frac{\gamma+\omega+q\mu}{\alpha}, 1, \frac{\gamma}{\mu+\delta}, \frac{\chi}{\mu(\chi+\mu+\Psi)}q, \frac{\xi}{\omega}\right)^T$$

where

$$q := \frac{(\alpha+\mu)(\gamma+\mu+\omega)(\mu+\Psi+\rho(\chi+\mu))}{\alpha(\mu+\Psi+\chi\rho)} \qquad \frac{\delta\gamma}{\delta+\mu}.$$

This vector has non-negative components that corresponds to zero entries in the DFE, which is sufficient[35].

By taking into account the zero entries of the right eigenvector and the second derivative of f:

$$\begin{aligned} b &= v_2 w_3 \frac{\partial f_2}{\partial I \partial \Lambda}(E_0, \Lambda\ ) + v_2 w_6 \frac{\partial f_2}{\partial V \partial \Lambda}(E_0, \Lambda\ ) \\ &= (v_2 w_3 \beta_V + v_2 w_6 \beta_I) \frac{\Psi+\mu+\rho\chi}{\mu(\chi+\Psi+\mu)} \\ &= (\beta_V + \frac{\xi}{\sigma}\beta_I) \frac{\Psi+\mu+\rho\chi}{\mu(\chi+\Psi+\mu)} \\ &> 0 \end{aligned}$$

and

$$\begin{aligned} a &= 2v_2 \left(w_1 w_3 \frac{\partial f_2}{\partial S \partial I}(E_0, \Lambda\ ) + w_1 w_6 \frac{\partial f_2}{\partial S \partial V}(E_0, \Lambda\ ) + w_5 w_3 \frac{\partial f_2}{\partial C \partial I}(E_0, \Lambda\ ) + w_5 w_6 \frac{\partial f_2}{\partial C \partial V}(E_0, \Lambda\ ) \right) \\ &= 2v_2(w_1+w_5)(w_3\beta_I + w_6\beta_V). \end{aligned}$$

from which we can conclude that backward bifurcation occurs if and only if

$$\frac{\delta\gamma}{\mu(\delta+\mu)} < \frac{(\alpha+\mu)(\gamma+\mu+\omega)(\mu+\Psi+\rho(\chi+2\mu))}{\mu\alpha(\mu+\Psi+\chi\rho)}$$

i.e. $a > 0$. □

Note that from (4.15) we can conclude that the parameters ξ, σ and the transmission rate β_V, which directly determine the dynamics of the environmental reservoir, does not have any influence on the type of the bifurcation.

5 Numerical Experiments

In this section, we investigate how the explicit Euler discretisation alters the qualitative properties, namely the positivity and the stability of the equilibria for the two epidemiological models introduced in section 4. We have performed numerical experiments to see how these results change when we use higher order SSP or modified-Patankar-Runge-Kutta methods.

We used the two stage second order SSPRK(2,2) method:

$$\begin{aligned} u^{(1)} &= u_n + \Delta t f(u_n) \\ u_{n+1} &= \frac{1}{2}u_n + \frac{1}{2}u^{(1)} + \frac{1}{2}\Delta t f(u^{(1)}) \end{aligned}$$

(which has $C = 1$ and $C_{eff} = 1$), the second order extrapolated BDF2 linear multistep method introduced in (3.26), the classical RK4 method introduced in (3.4) (which has $C = 0$, as we have seen). For the conservative model, we have also used the $MPRK22(1)$ method and to analyse the SSP methods, we have also used the explicit Euler method. Note that the SSPRK(2,2) method is the explicit trapezoidal rule with Butcher tableau

$$\begin{array}{c|c} c & A \\ \hline & b^t \end{array} \quad = \quad \begin{array}{c|cc} 0 & & \\ 1 & 1 & \\ \hline & 1/2 & 1/2 \end{array}$$

from Runge-Kutta theory. While at first glance it may sound preferable to also use the SSP MPRK methods for the non-conservative case, this would be problematic since these methods preserve the 'mass' with the convection part [22, Thm. 2.2].

5.1 Conservative model case

Due to the low dimensionality of the conservative model, the explicit Euler discretisation can be extensively analysed. To study the positivity under the explicit Euler discretisation, we rewrite the system (4.1a)-(4.1c) in the so-called *Graph-Laplacian form* [36]

$$\dot{u} = A(u)u \tag{5.1}$$

where $A : \mathsf{R}^d \; ! \; \mathsf{R}^{d \;\; d}$ with the properties

1. $a_{ij}(u) \quad 0$ for $i, j = 1, \ldots, d; i \; 6= j$, while $a_{ii}(u) \quad 0$ for $i = 1, \ldots, d; \; 8u \quad 0$.
2. The column-sums add to zero i.e. $8u \quad 0: \sum_{i=1}^{d} a_{ij}(u) = 0; \; 8j = 1, \ldots, d.$

The system (4.1a)-(4.1c) in the graph Laplacian form with $u = (S, I, R)$ is

$$A := \begin{pmatrix} kI & r + (1 \quad m)b & pb^{\emptyset} & \varphi + (1 \quad m)b \\ & kI & v \quad pb^{\emptyset} & 0 \\ & mb & v & \varphi + b(m \quad 1) \end{pmatrix} \tag{5.2}$$

Now, considering the positivity of the discretisation, one can state the following

Proposition 5.1. *The explicit Euler discretised system is positively invariant in Ω if*

$$\Delta t \quad \min\left\{ \frac{1}{k + bm}, \frac{1}{v + pb^{\emptyset}}, \frac{1}{\varphi + b(1 \quad m)} \right\}$$

Proof. Suppose that $(s_n, i_n, r_n) \; 2 \; \Omega$, we want to prove that $(s_{n+1}, i_{n+1}, r_{n+1}) \; 2 \; \Omega$. It is clear, that the Explicit Euler method is conservative, so one has to prove only the positivity. Using the graph-Laplacian (5.2) form of the system, the explicit Euler discretisation reads

$$u_{n+1} = u_n + \Delta t A(u_n)u_n = (I + \Delta t A(u_n))u_n$$

since $u_n \quad 0$, it is sufficient that all elements of $A(u_n)$ are positive. By the properties of $A(u_n)$, the off-diagonal elements are positive for any step-size, for $a_{11}(u_n)$, we require

$$1 \quad \Delta t(\quad ki_n \quad mb).$$

since in the model r=b. If $\Delta t(\quad ki_n \quad mb) \quad 0$, then it holds for arbitrary Δt. If $\Delta t(\quad ki_n \quad mb) > 0$, then for the positivity we require that

$$\frac{1}{(\quad ki_n \quad mb)} > \Delta t.$$

One can ensure this by choosing Δt such that $\frac{1}{k+mb} \quad \Delta t$ since the explicit Euler method preserves conservativity unconditionally i.e. $i_n \quad 1; \; 8n \; 2 \; \mathsf{N}$. This 'technique' can be used to the other diagonal elements in a straightforward way to get the other conditions. □

Note that the second and third value in 5.1 is sharp, in the sense that if we use any step size for which, the condition does not hold, then there exist $u_0 \; 2 \; \mathsf{R}^3_+$, for which the numerical solutions 'steps out' from R^3_+, i.e. the positive orthant is not positively invariant.

To see this, let us choose Δt such that $(v+pb^0)\Delta t = 1+\varepsilon$, where $\varepsilon > 0$. Then clearly the condition in the theorem does not hold. The first step for the infectious compartment is

$$\begin{aligned} i_1 &= i_0(1 \quad \Delta t(v+pb^0)) + \Delta t k i_0 s_0 \\ &= \left(\quad \varepsilon + \frac{(1+\varepsilon)ks_0}{v+pb^0} \right) i_0. \end{aligned}$$

If we suppose that $i_0 \neq 0$, then i_1 is negative if

$$s_0 < \frac{\varepsilon}{1+\varepsilon}\frac{v+pb^0}{k}. \tag{5.3}$$

Since $\lim_{\varepsilon \to 0} \frac{\varepsilon}{1+\varepsilon} = 0$, it is sharp for $s_0 = 0$ with i_0, r_0 arbitrary.

For the third condition, if we choose the step-size as $\Delta t = \frac{1+\varepsilon}{\varphi + b(1 \quad m)}$ and use the conservativity property $r_0 = 1 \quad s_0 \quad i_0$, we get that r_1 is negative if

$$\left(\varepsilon + \frac{1+\varepsilon}{\varepsilon}\frac{v}{\varphi + b(1 \quad m)}\right) i_0 + \left(\varepsilon + \frac{1+\varepsilon}{\varepsilon}\frac{mb}{\varphi + b(1 \quad m)}\right) s_0 < \varepsilon \tag{5.4}$$

which is sharp for $i_0 = s_0 = 0$, $r_0 = 1 \quad s_0 \quad i_0 = 1$.

Considering its stability, the following can be showed

Proposition 5.2. *The explicit Euler discretised system of* (4.1a)-(4.1c)

1. *has two, and no more equilibria, which coincides with the equilibria (DFE and EE) of the continuous system. The DFE is positive if and only if* $R_0 \quad 1$ *while the EE is always positive.*
2. *The DFE equilibrium is conditionally locally asymptotically stable with step sizes* $\Delta t < H_1$, *if* $R_0 < 1$ *where*
$$H_1 = \min\left\{\frac{2}{r+\varphi}, \frac{2}{(pb^0+v)(1 \quad R_0)}\right\}$$
and unstable if $\Delta t > H_1$ *or* $R_0 > 1$.
3. *The endemic equilibria is locally asymptotically stable with step sizes* $\Delta t < H_2$, *if* $R_0 > 1$ *where*
$$H_1 = \min\left\{\Delta t < \frac{1}{\varphi + (1 \quad m)b + v}, \frac{4(\varphi + (1 \quad m)b + v)}{(\varphi+b)(pb^0+v)(R_0 \quad 1) + (r+\varphi)(\varphi + (1 \quad m)b + v)}\right\}$$
and unstable if $R_0 > 1$.

Proof. As it was shown in 3.5, the explicit Euler method preserves the equilibria of the continuous method and no spurious equilibria emerges independently of the step-size. So we are done with the first half of the proof. To show the second half, following the theory introduces in the subsection 3.5, if λ is an eigenvalue of the linearized system at equilibrium point u , then $1+\Delta t\lambda$ is an eigenvalue for the linearized system of the Euler method $u + \Delta t f(u)$. To have asymptotic stability for the discretised system, it is sufficient that $|1+\Delta t\lambda| < 1$ for all the eigenvalues of $f^0(u \ )$.

The eigenvalues of the considered system at the DFE are (4.4)-(4.5). The first eigenvalue implies that for the asymptotic stability of the DFE, one must have

$$\Delta t \quad \frac{2}{r+\varphi}.$$

Considering the second eigenvalue, it is asymptotically stable if

$$|1 + \Delta t(kS_0 \quad (pb^0+v))| \quad 1$$

which is equivalent with

$$j1 + \Delta t(pb^0 + v)(R_0 \quad 1)j \quad 1.$$

Hence, the other necessary conditions for the asymptotic stability are $R_0 \quad 1$ and

$$\Delta t \quad \frac{2}{(pb^0 + v)(1 \quad R_0)} \tag{5.5}$$

and if $R_0 > 1$ then the linearized system is unstable. Considering the endemic equilibria, the linearized system is

$$\begin{pmatrix} s_{n+1} \\ i_{n+1} \\ r_{n+1} \end{pmatrix} = \begin{pmatrix} 1 & \Delta t\big(\frac{(\varphi+b)(pb^0+v)(R_0 \quad 1)}{\varphi+(1 \quad m)b+v} + r + \varphi\big) & \Delta t(v + (1 \quad m)b + \varphi) \\ & \Delta t\frac{(\varphi+b)(pb^0+v)(R_0 \quad 1)}{\varphi+(1 \quad m)b+v} & 1 \end{pmatrix} \begin{pmatrix} s_n \\ i_n \\ r_n \end{pmatrix}.$$

Since the eigenvalues of the above matrix does not simplify, we use the Schur-Cohn conditions, which states that a second order polynomial $p(\lambda) = \lambda^2 + \lambda a_1 + \lambda a_0$ has its eigenvalues inside the unit disk in the complex plane if and only if $ja_1j < 1 + a_2 < 2$ [37, Thm. 4.5], which for the above matrix is

$$jtrace(A)j < 1 + det(A) < 2.$$

Considering the case $1 \quad det(A) > 0$, it is

$$1 < \frac{(\varphi + b)(pb^0 + v)(R_0 \quad 1)(1 \quad \Delta t(\varphi + (1 \quad m)b + v)}{(r + \varphi)(\varphi + (1 \quad m)b + v)},$$

which holds if $R_0 > 1$ and

$$\Delta t < \frac{1}{\varphi + (1 \quad m)b + v}.$$

The condition $1 \quad tr(A) + det(A) > 0$ is equivalent with $R_0 > 1$, while the condition $1 + tr(A) + det(A) > 0$ is equivalent with

$$p(\Delta t) := 4 \quad 2\Delta t\Big(\frac{(\varphi + b)(pb^0 + v)(R_0 \quad 1)}{\varphi + (1 \quad m)b + v} + r + \varphi\Big) + \Delta t^2(\varphi + b)(pb^0 + v)(R_0 \quad 1) > 0$$

The roots of the above polynomial can be calculated and one can give a sharp condition. A sufficient condition can be given by linear approximation of the polynomial at $\Delta t = 0$, where $p(0) = 4$. The root of the linear approximant will be smaller than the root of the polynomial if $R_0 > 1$ since in that case the polynomial is monotonically decreasing at $\Delta t = 0$ and convex. The condition reads as

$$\Delta t < \frac{4(\varphi + (1 \quad m)b + v)}{(\varphi + b)(pb^0 + v)(R_0 \quad 1) + (r + \varphi)(\varphi + (1 \quad m)b + v)}.$$

□

We point out that the explicit Euler discretisation also preserves the geometric property of the DFE that, it is a stable node (since both eigenvalues are real) with reversed orientation when $det\big(1 + \Delta t f^0(u\ )\big) < 0$. For the definitions of the above, see [7]. We can compare the conditions for the positivity from prop. 5.1 with the conditions for the stability for the DFE from 5.2 since they are sharp. Since

$$\frac{2}{R_0 \quad 1} > 1$$

is equivalent with $R_0 < 3$ when $R_0 < 1$, so it always holds for $R_0 < 1$. This implies that in the case $\frac{2}{r+\varphi} > \frac{1}{(pb^0+v)}$ the positivity breaks down first, then the local stability of the DFE, when we increase the step-size for the explicit Euler method.

We have done extensive numerical calculations to see for which step-sizes the solution of the problem becomes negative. In order to cover the entire biologically feasible region Ω, we solved the problem with the different schemes for initial values in the triangle with resolution 0.01 and at the boundary with resolution 0.001 and at the neighbouring points, where the explicit Euler method first (w.r.t. Δt) 'steps' into one of the negative quadrants with resolution 10^{-14}. Thus we have approximately found the 'system-dependent SSP coefficients C' (i.e. the smallest step size for which the numerical solution for a particular numerical scheme loses its positivity divided by the smallest step size for which the numerical solution for the explicit Euler method loses its positivity) for 4 realisations of (4.1a)-(4.1c). Two systems have $R_0 < 1$, while two have $R_0 > 1$. The results can be seen in figure 4, where the different colours are the different systems (i.e. different specific parameters), while the values are the first time-step instances where the numerical solution was negative weighted by the first instance where the approximation with the explicit Euler became negative. In all cases, the positivity conditions of proposition 5.1 were sharp. We emphasise that different colours represent different specific systems (i.e. different specific parameters for the model). We excluded the MPRK method from the numerical simulation as it is unconditionally positive.

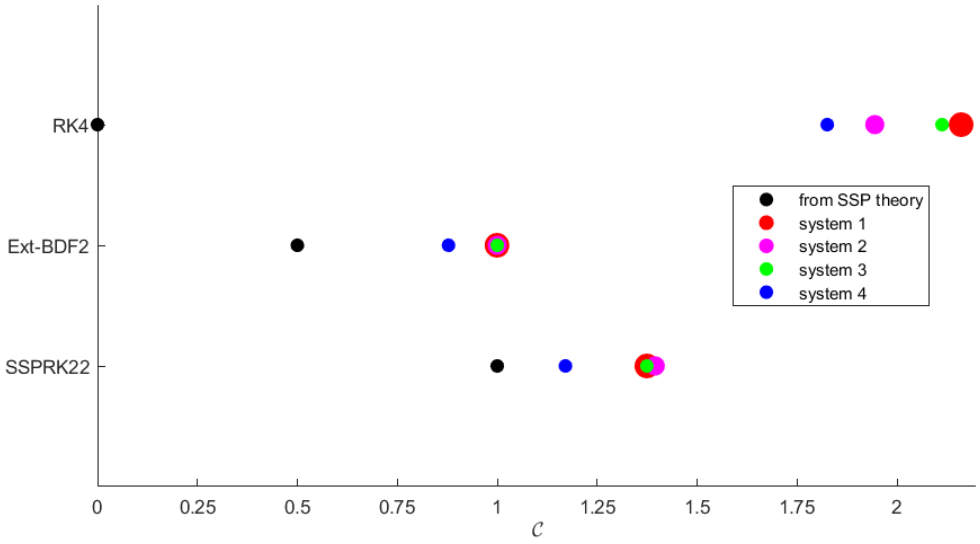

Figure 4: Numerically found 'system dependent SSP coefficients' (i.e. the coefficient C in the figure is the smallest step size for which the numerical solution loses its positivity for a given numerical scheme divided by the smallest step size for which the numerical solution loses its positivity for the explicit Euler method). The four systems are four different realisations of the model (4.1a)-(4.1c), i.e. the parameters are specified, and listed in the table 1.

From figure 4 and table 1, it is clear that the maximal step-sizes, for which positivity is preserved are significantly larger than the values what one would get from the SSP theory, and in all cases the largest is for the classical RK4 method, which has the smallest theoretical SSP coefficient, namely 0. It is also clear, that for the extrapolated-BDF2 method, one could not expect a coefficient greater than 1 for any system, since the explicit Euler starting procedure is used to preserve positivity. The numerical simulations also showed that the positivity conditions of proposition 5.1 are not only sufficient, but also necessary, and the first instances, where the positivity is lost are $s_0 = 0$, i_0 arbitrary when the positivity condition is $\frac{1}{v+pb^0}$ and $s_0 = i_0 = 0$ when the positivity condition is $\frac{1}{\varphi+b(1-m)}$ (see (5.3) and (5.4)). From the table, it also follows, that the condition $\Delta t \frac{1}{k+bm}$ is not

	System 1.	System 2.	System 3.	System 4.
p	0.7	0.4	0.7	0.6
b	1	3	1	3
b'	1	5	1	3
v	2	1	2	2
k	2	2	6	5
m	0.3	0.2	0.3	0.2
φ	5	0.1	5	2.2
R_0	0.7037	0.5376	2.1111	1.1640
$\frac{1}{v+pb^0}$	0.3704	**0.3333**	0.3704	0.2632
$\frac{1}{k+bm}$	0.4348	0.3846	**0.1587**	**0.1786**
$\frac{1}{\varphi+b(1 \quad m)}$	**0.1754**	0.4	0.1754	0.2174
explicit Euler Δt	0.1755	0.3334	0.1754	0.2174
SSPRK22 Δt	0.2412	0.4652	0.2412	0.2545
Ext-BDF2 Δt	0.1755	0.3334	0.1754	0.1909
RK4 Δt	0.3792	0.6483	0.3708	0.3969

Table 1: Specific parameter values and numerically found positivity conditions of the four system which were considered in the numerical simulation. The positivity conditions (see prop. 5.1) for the explicit Euler method are highlighted in bold.

neccesary. Interestingly, for the larger order methods, the first instances where positivity is lost are in some cases different from the explicit Euler case. Namely, for the SSPRK2 method, these first instances are at $s_0 = 0$, $i_0 = 1$ for all the systems. For the Ext-BDF2 method, the fourth system while for the RK4 method the first and the second system differ (with $s_0 = 0$, $i_0 = 1$). The larger coefficients can be partially explained by the fact, that we did not require positivity preservation for the internal stages of the Runge-Kutta methods. The theory introduced in [14] and summarised at the end of the subsection 3.2.2, namely, one can guarantee non-zero step-size restriction for the positivity preservation for the classical RK4 method. This can be done if for the continuous model it holds that its explicit Euler discretisation also preserves positivity for backward steps

$$0 \quad u + \hat{f}(u) = u \quad f(u), \quad 8u \quad 0, 8\Delta t \quad \Delta t_{FE}$$

where Δt_{FE} is the largest step-size for which it holds. This does not explain our results, since at the boundary $i_0 = 0$ c_1 is strictly positive in (4.6), which implies (with the continuity of the vector field) that $\Delta t_{FE} = 0$. Note that this is the general case for epidemiological models of the form (4.6). In [14] the logistic equation $\dot{u} = u(u \quad 1)$ was considered, which has a stable equilibrium at $u = 0$.

If we also require the positivity of the internal stages, then the SSPRK(2,2) has the same values as the explicit Euler, since that is its internal stage. For the classical RK4 method, the smallest such step-sizes when one of the internal stages loses its positivity can be found in the table 2. For all considered systems, the step sizes are still larger than for the explicit Euler method. Note that in this case it cannot be larger than twice the explicit Euler's, since the second stage is $u_n + \Delta t \frac{1}{2} f(u_n)$.

	System 1.	System 2.	System 3.	System 4.
RK4 + internal stages Δt	0.2249	0.4393	0.2081	0.2545

Table 2: Numerically found positivity conditions for the classical RK4 method when we also require the positivity preservation of the internal stages. The parameters of the four systems considered in the numerical simulation can be found in the table 1.

We also checked the trajectories of the numerical solutions with specific initial values for which the positivity first breaks down. In all cases the positivity was lost, but the long term behaviour of the continuous model was preserved, except for the two systems with $R_0 > 1$ for the RK4 discretisation. In these cases, the positivity and the local stability were lost at the same time.

The conservative model (4.1a)-(4.1c) in the PDS formulation (3.27) with $(u_1, u_2, u_3) = (S, I, R)$, which is used for the MPRK method is

$$\begin{aligned} p_{12} &= pb^0 I & p_{23} &= 0 \\ p_{13} &= (\varphi + b)R & p_{31} &= mb(S + R) \\ p_{21} &= kSI & p_{32} &= vI \end{aligned}$$

and $d_{i,j} = p_{j,i}$, i, j 2 f1, 2, 3g. To address the issue of order reduction, we have compared the numerical solutions with a very accurate numerical solution obtained with Matlab's ode45 method - which is based on the Dormand-Prince embedded RK method - with minimal tolerances ($Abstol = 10^{-14}$, $Reltol = 10^{-13}$). Note that embedded RK methods give values at t points where the consistency error is small enough, and not at any t points, therefore their values cannot be compared with the approximate solutions of the other methods in a straightforward way. This problem can be solved using Matlab's 'deval' function, which works by using a third RK method (alongside the embedded method) called the *dense output RK method*. From this, an interpolating polynomial can be constructed based on the end points and some intermediate points, which for the Dormand-Prince method gives a fifth order approximation at any inner points[38]. We have not found order-reduction for our model, considering solutions with initial values near the boundary. A case can be seen in figure 5.

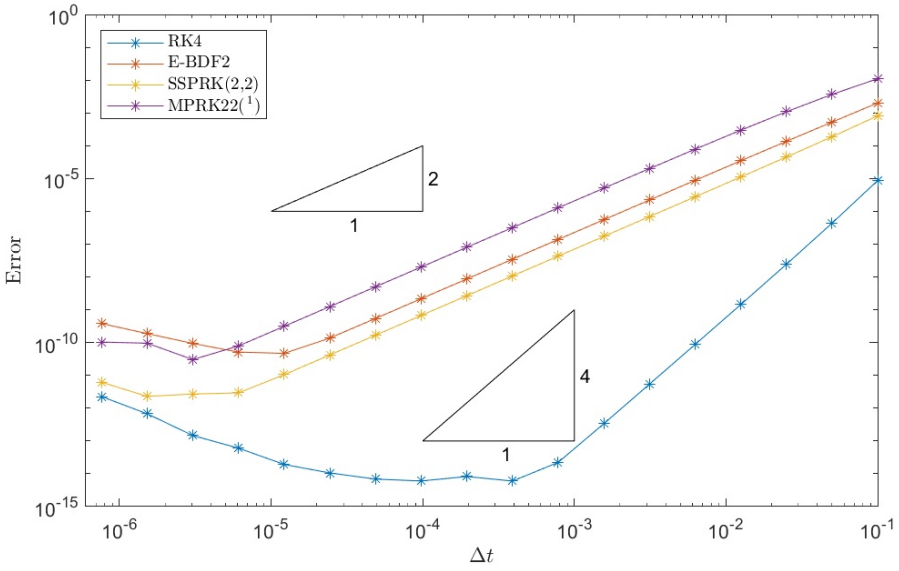

Figure 5: Order of the different method by comparing the numerical solutions with a very accurate method.

It is clear that to have a good numerical approximation of the continuous model, we expect not only local, but global asymptotic stability as in the continuous model. While we have not proved global stability for the discretisations - for sufficiently small step sizes - it can be 'checked' by numerical simulations. To do this, we solved the system numerically using the various methods with numerous initial conditions from the feasible region. More precisely, with resolution 0.01 for large enough times to obtain the long-term behaviour of the solutions. By this procedure, one can find the asymptotically stable solutions - equilibria, periodic orbits - of the system and it also shows whether the system exhibits chaotic behaviour or whether the solutions may become unbounded and diverge. We have done this procedure for different values of R_0 and plotted the long-term numerical solutions of the infectious subpopulation, obtaining so-called bifurcation diagrams. For the explicit methods, for sufficiently small step-sizes, all the solutions converged to the DFE equilibrium for $R_0 < 1$ and to the endemic equilibrium for $R_0 > 1$. For large step-sizes, the numerical solutions diverged to ∞. For the explicit Euler method, the global convergence and divergence coincide with the conditions of proposition 5.2 for the DFE. or the MPRK method, the long term behaviour of the numerical solutions 'mimics' the continuous model for arbitrarily large step sizes. This can be seen on 6 with $\Delta t = 10$. Note that the stability of the equilibria locally follows from the unconditional absolute stability of the method, but not globally.

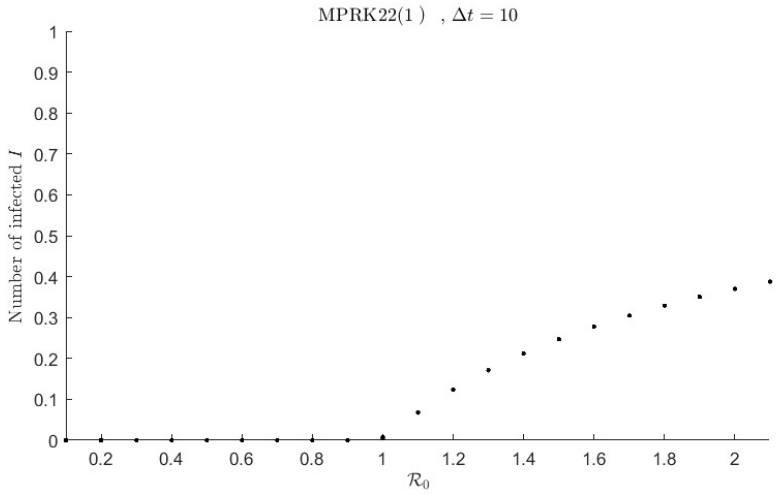

Figure 6: Numerical bifurcation diagram of the MPRK22(1) scheme with the same parameters as in system 1 in table 1, but with varied parameter k.

5.2 Non-conservative system

By discretising the system (4.11a)-(4.11f) by the explicit Euler method, we get the following (discrete) model:

$$s_{n+1} = s_n + \Delta t\big[\Lambda \quad \beta_I s_n i_n \quad \beta_V s_n v_n + \Psi c_n + \delta r_n \quad (\chi+\mu)s_n\big] \tag{5.6a}$$
$$e_{n+1} = e_n + \Delta t\big[\beta_I s_n i_n + \beta_V s_n v_n + \rho\beta_I c_n i_n + \rho\beta_V c_n v_n \quad (\alpha+\mu)e_n\big] \tag{5.6b}$$
$$i_{n+1} = i_n + \Delta t\big[\alpha e_n \quad (\gamma+\omega+\mu)i_n\big] \tag{5.6c}$$
$$r_{n+1} = r_n + \Delta t\big[\gamma i_n \quad (\mu+\delta)r_n\big] \tag{5.6d}$$
$$c_{n+1} = c_n + \Delta t\big[\chi s_n \quad \rho\beta_I c_n i_n \quad \rho\beta_V c_n v_n \quad (\Psi+\mu)c_n\big] \tag{5.6e}$$
$$v_{n+1} = v_n + \Delta t\big[\xi i_n \quad \sigma v_n\big] \tag{5.6f}$$

The positivity and boundedness of the above method can be guaranteed by the following sufficient condition:

Theorem 5.1. *The explicit-Euler discretisation of the system* (4.11a)-(4.11f) *is conditionally positive with step-size* $\Delta t \quad H$*, where*

$$H = \min\left\{\frac{1}{\alpha+\mu}, \frac{1}{\sigma}, \frac{1}{\mu+\delta}, \frac{1}{\gamma+\omega+\mu}, \frac{1}{\chi+\mu+\frac{}{\mu}(\beta_I+\beta_V\frac{\xi}{\sigma})}, \frac{1}{\Psi+\mu+\rho\frac{}{\mu}(\beta_I+\beta_V\frac{\xi}{\sigma})}\right\}$$

Proof. For the positivity, we will need some boundedness, so in general we will show that if $(s_n, e_n, i_n, r_n, c_n, v_n)$ *2* Ω then $(s_{n+1}, e_{n+1}, i_{n+1}, r_{n+1}, c_{n+1}, v_{n+1})$ *2* Ω. Denote $n_n := s_n + e_n + i_n + r_n + c_n$, then $n_{n+1} = n_n + \Delta t(\Lambda \quad \mu n_n \quad \omega i_n) \quad (1 \quad \Delta t\mu)n_n + \Delta t\Lambda$, which is smaller or equal than $\frac{}{\mu}$ if $\Delta t \quad \frac{1}{\mu}$. Similarly, if $\Delta t \quad \frac{1}{\sigma}$, then $v_{n+1} \quad \frac{\xi}{\sigma\mu}$.

For the positivity, we will use the same logic as in [39]. For the first variable, we want to show that s_{n+1} *2* $[0, \frac{\lambda}{\mu}]$. From the explicit-Euler discretisation:

$$s_{n+1} = s_n + \Delta t(\Lambda \quad \beta_I s_n i_n \quad \beta_V s_n v_n + \Psi c_n + \delta r_n \quad (\chi+\mu)s_n)$$

The positivity holds if and only if

$$s_n \qquad \Delta t(\Lambda \quad \beta_I s_n i_n \quad \beta_V s_n v_n + \Psi c_n + \delta r_n \quad (\chi+\mu)s_n). \tag{5.7}$$

If

$$(\Lambda \quad \beta_I s_n i_n \quad \beta_V s_n v_n + \Psi c_n + \delta r_n \quad (\chi+\mu)s_n) \quad 0$$

then the inequality (5.7) holds for any step size. If

$$\Delta t(\Lambda \quad \beta_I s_n i_n \quad \beta_V s_n v_n + \Psi c_n + \delta r_n \quad (\chi+\mu)s_n) > 0$$

then the positivity holds for step sizes

$$\Delta t < \frac{s_n}{\Lambda + \beta_I s_n i_n + \beta_V s_n v_n \quad \Psi c_n \quad \delta r_n + (\chi+\mu)s_n)}. \tag{5.8}$$

From the inequality:

$$\frac{1}{(\chi+\mu)+(\beta_I+\beta_V\frac{\xi}{\sigma})\frac{}{\mu}} = \frac{s_n}{s_n(\chi+\mu)+(\beta_I+\beta_V\frac{\xi}{\sigma})\frac{}{\mu}s_n} \\ \frac{s_n}{\Lambda+\beta_I s_n i_n+\beta_V s_n v_n+(\chi+\mu)s_n \quad \Psi c_n \quad \delta r_n} \tag{5.9}$$

So for any Δt $\frac{1}{(\chi+\mu)+(\beta_I+\beta_V\frac{\xi}{\sigma})\frac{\Lambda}{\mu}}$ the inequality (5.7) holds, i.e. s_{n+1} 0.
For $e_{n+1}, i_{n+1}, r_{n+1}, c_{n+1}, v_{n+1}$ the proof can be carried out similarly, but one gets simpler sufficient conditions for Δt because of the sign of the terms. □

Note that (5.9) is sufficient but not necessary condition, the non-negativity for s_n is fully determined by the condition (5.8) and one can get similar conditions for the other coordinates. Considering the local stability of the equilibria under the explicit Euler discretisation, sufficient and necessary conditions cannot be easily given since for this we have to determine the eigenvalues of 6 6 matrix which is equivalent with finding the roots of a $6th$ degree polynomial. Considering only sufficient conditions, Gershgorin circle theorem can be used in general [40], but this was not done in this thesis.

In order to study the sharpness of the above results and the sharpness of the SSP coefficients from the SSP theory introduced in section 3, we performed a similar numerical procedure as for the conservative model. Namely, we solved the system with numerous initial conditions in the feasible region to find the smallest such step-size Δt for which the positivity or the boundedness of the solution does not hold. Since the system is 6 dimensional, solving it for initial conditions with resolution 0.01 (which are then rescaled to be in the feasible region) would require numerically solving 10^{12} initial value problems, which is not feasible, so we choose 1000 points from the grid with resolution 0.01 from the feasible region, and another 1000 from near the boundary. We have done this for two systems with given parameters[2]. The results can be found in table 3. The results are similar to those for the conservative model, namely, if we calculate the smallest step size for which the numerical solution loses its positivity or boundedness for a specific numerical scheme divided by the smallest step size for which the numerical solution loses its positivity or boundedness for the explicit Euler method, then for all schemes and systems, they are significantly larger than the theoretical SSP coefficients and largest for the classical RK4 method when we require the positive invariance of Ω only for the steps and not for the internal stages. From table 3 it is also clear that the sufficient conditions of proposition 5.1 for the explicit Euler method are not sharp. It should be emphasised that since we have not been able to cover the whole feasible region, and since we have not been able to specify the initial value(s) for which the positivity or the boundedness first breaks down, unlike for the conservative case, these results can be questioned.

[2] For both systems β_I = 4, β_V = 3, μ = 2, α = 1, ω = 1, γ = 0.1, ξ = 1, σ = 1, ρ = 0.9, ψ = 1, χ = 0.1, δ = 3. For the rst system = 1.2487, for the second system = 3.6058.

	System 1.	System 2.
explicit Euler positivity condition from proposition	0.1442	0.0679
explicit Euler Δt	0.210	0.124
SSPRK22 Δt	0.251	0.338
Ext-BDF2 Δt	0.251	0.124
RK4 Δt	0.375	0.356
RK4 + inner stages Δt	0.250	0.161

Table 3: Numerically found positivity conditions of the two systems which were considered in the numerical simulation. The sufficient positivity conditions (see prop. 5.1) for the explicit Euler are also shown in the second row.

6 Summary, Conclusions

Through the construction of epidemiological models, we can better understand - and possibly predict - the dynamics and qualitative properties of different infectious diseases. As these systems cannot generally be solved analytically, various numerical methods are used to obtain the approximate solutions that are of interest, for example, for forecasting. It is well-known that discretisations by different numerical schemes can alter the qualitative properties of the ODE system, therefore it is of interest when and for which schemes these qualitative properties are not changed. One such evident and well-studied property is the positivity (non-negativity) of the compartments/coordinates when the initial conditions are also non-negative.

Considering the classical linear methods, namely Runge-Kutta and linear multistep methods, it is known that there is no second or higher order method that preserves positivity for all step sizes and for all positive ODE systems. It is therefore of interest to find the largest such step-size for which the positivity is preserved for a given method. The positivity preservation of these classical methods can be studied through SSP theory, but this general theory gives strict sufficient step-size conditions, since it preserves not only the positivity, but arbitrary convex functionals, and not for a specific ODE, but for arbitrary ODEs. Hence, it is of interest how sharp these step-sizes are and what are the necessary conditions for preserving positivity under different schemes.

In view of the above, we analysed how sharp the step-sizes are for two different epidemic models. While for we were able to give sufficient, and for one of the models necessary conditions for preserving positivity (and boundedness) under the explicit Euler discretisation, considering higher order methods, we found the smallest such step-sizes through numerical simulations. We found that the theoretical SSP coefficients are significantly smaller than the smallest such step-size for which the positivity is preserved for all considered methods. In particular, while the classical RK4 method has $\mathsf{C} = 0$, for our specific models, the maximal step-sizes were 1.5 2.5 times larger than for the explicit Euler method. These results can be partially explained by the negative internal stages, but even when we required positivity preservation from the internal stages, the maximal step-sizes were larger than for the explicit Euler method. It is a possible future direction to give formal proofs of the results found through numerical simulations. Another possible future direction is to study the influence of the conservativity or the Graph-Laplacian form on the positivity preservation since the proof of proposition 5.1 cannot be carried out for Runge-Kutta methods with more stages, since these discretisations do not preserve the Graph-Laplacian structure.

We have also constructed a new epidemiological method to study the spread of the SARS-CoV-2 virus. This model incorporates the propagation of the virus through the environment and the vaccination of the population with an imperfect vaccine. We showed that the model exhibits backward bifurcation, i.e. for some parameter values a stable disease-free equilibrium coexists with an unstable endemic equilibrium. We also showed that this is independent of the dynamics of the environment and its disease spread.

We also introduced a family of (nonlinear) unconditionally positive and conservative schemes, called modified-Patankar-Runge-Kutta schemes and summarised some of the recent developments considering these schemes. While we did not prove global stability under the discretisation for arbitrary step-sizes, the bifurcation diagrams possibly imply it. Although we have not given the specific running times of the different schemes, it is clear that the MPRK schemes require the most time and the most computational power, since one has to solve two linear algebraic systems in each time step. It depends on the application and the specific circumstances, whether the MPRK method should be used instead of linear explicit methods for non-stiff systems, like the one we had. While we solved them directly, the positivity is also preserved under the Jacobi iterative method[19]. Since these systems are relatively new and have only recently been systematically analysed, there are numerous open questions regarding their dynamics. Some of these are the existence of spurious k-periodic solutions, the existence of spurious equilibria, or the preservation of quasi-monotonic structure/monotonicity, which was analysed for the Runge-Kutta methods in [41].

References

[1] W. O. Kermack and A. G. McKendrick, "Contributions to the mathematical theory of epidemics–i. 1927.," *Bulletin of mathematical biology*, vol. 53, no. 1-2, pp. 33–55, 1991.

[2] S. Busenberg and K. Cooke, *Vertically transmitted diseases: models and dynamics*. Springer Science & Business Media, 2012, vol. 23.

[3] M. Martcheva, *An introduction to mathematical epidemiology*. Springer, 2015, vol. 61.

[4] V. Capasso, *Mathematical structures of epidemic systems*. Springer Science & Business Media, 2008, vol. 97.

[5] P. Van den Driessche and J. Watmough, "Further notes on the basic reproduction number," in *Mathematical epidemiology*, Springer, 2008, pp. 159–178.

[6] J. K. Hale, "Ordinary differential equations," *Wiley-Interscience*, 1969.

[7] J. K. Hale and H. Koçak, *Dynamics and bifurcations*. Springer Science & Business Media, 2012, vol. 3.

[8] J. LaSalle, "Some extensions of liapunov's second method," *IRE Transactions on circuit theory*, vol. 7, no. 4, pp. 520–527, 1960.

[9] W. H. Hundsdorfer, J. G. Verwer, and W. Hundsdorfer, *Numerical solution of time-dependent advection-di usion-reaction equations*. Springer, 2003, vol. 33.

[10] C.-W. Shu and S. Osher, "Efficient implementation of essentially non-oscillatory shock-capturing schemes," *Journal of computational physics*, vol. 77, no. 2, pp. 439–471, 1988.

[11] W. Hundsdorfer, S. J. Ruuth, and R. J. Spiteri, "Monotonicity-preserving linear multistep methods," *SIAM Journal on Numerical Analysis*, vol. 41, no. 2, pp. 605–623, 2003.

[12] S. Gottlieb, D. I. Ketcheson, and C.-W. Shu, *Strong stability preserving Runge-Kutta and multistep time discretizations*. World Scientific, 2011.

[13] J. F. B. M. Kraaijevanger, "Contractivity of runge kutta methods," *BIT Numerical Mathematics*, vol. 31, no. 3, pp. 482–528, 1991.

[14] I. Higueras, "Representations of runge–kutta methods and strong stability preserving methods," *SIAM journal on numerical analysis*, vol. 43, no. 3, pp. 924–948, 2005.

[15] R. J. Spiteri and S. J. Ruuth, "A new class of optimal high-order strong-stability-preserving time discretization methods," *SIAM Journal on Numerical Analysis*, vol. 40, no. 2, pp. 469–491, 2002.

[16] I. H. Sanz, "Positivity properties for the classical fourth order runge-kutta method," *Monograf as de la Real Academia de Ciencias Exactas, F sicas, Qu micas y Naturales de Zaragoza*, no. 33, pp. 125–139, 2010.

[17] N. Pham Thi, W. Hundsdorfer, and B. Sommeijer, "Positivity for explicit two-step methods in linear multistep and one-leg form," *BIT Numerical Mathematics*, vol. 46, no. 4, pp. 875–882, 2006.

[18] B. Baliga and S. Patankar, "A new finite-element formulation for convection-diffusion problems," *Numerical Heat Transfer*, vol. 3, no. 4, pp. 393–409, 1980.

[19] H. Burchard, E. Deleersnijder, and A. Meister, "A high-order conservative patankar-type discretisation for stiff systems of production–destruction equations," *Applied Numerical Mathematics*, vol. 47, no. 1, pp. 1–30, 2003.

[20] S. Kopecz and A. Meister, "On order conditions for modified patankar–runge–kutta schemes," *Applied Numerical Mathematics*, vol. 123, pp. 159–179, 2018.

[21] S. Kopecz and A. Meister, "Unconditionally positive and conservative third order modified patankar–runge–kutta discretizations of production–destruction systems," *BIT Numerical Mathematics*, vol. 58, pp. 691–728, 2018.

[22] J. Huang and C.-W. Shu, "Positivity-preserving time discretizations for production–destruction equations with applications to non-equilibrium flows," *Journal of Scienti c Computing*, vol. 78, pp. 1811–1839, 2019.

[23] J. Huang, W. Zhao, and C.-W. Shu, "A third-order unconditionally positivity-preserving scheme for production–destruction equations with applications to non-equilibrium flows," *Journal of Scienti c Computing*, vol. 79, pp. 1015–1056, 2019.

[24] T. Izgin, S. Kopecz, and A. Meister, "On lyapunov stability of positive and conservative time integrators and application to second order modified patankar–runge–kutta schemes," *ESAIM: Mathematical Modelling and Numerical Analysis*, vol. 56, no. 3, pp. 1053–1080, 2022.

[25] T. Izgin, S. Kopecz, and A. Meister, "On the stability of unconditionally positive and linear invariants preserving time integration schemes," *SIAM Journal on Numerical Analysis*, vol. 60, no. 6, pp. 3029–3051, 2022.

[26] D. Torlo, P. Öffner, and H. Ranocha, "Issues with positivity-preserving patankar-type schemes," *Applied Numerical Mathematics*, vol. 182, pp. 117–147, 2022.

[27] T. Izgin, P. Öffner, and D. Torlo, "A necessary condition for non oscillatory and positivity preserving time-integration schemes," *arXiv preprint arXiv:2211.08905*, 2022.

[28] A. Iserles, "Stability and dynamics of numerical methods for nonlinear ordinary differential equations," *IMA journal of numerical analysis*, vol. 10, no. 1, pp. 1–30, 1990.

[29] E. Hairer, A. Iserles, and J. M. Sanz-Serna, "Equilibria of runge-kutta methods," *Numerische Mathematik*, vol. 58, no. 1, pp. 243–254, 1990.

[30] D. Griffiths, P. Sweby, and H. C. Yee, "On spurious asymptotic numerical solutions of explicit runge-kutta methods," *IMA Journal of numerical analysis*, vol. 12, no. 3, pp. 319–338, 1992.

[31] A. Stuart and A. R. Humphries, *Dynamical systems and numerical analysis*. Cambridge University Press, 1998, vol. 2.

[32] C. Yang and J. Wang, "A mathematical model for the novel coronavirus epidemic in wuhan, china," *Mathematical biosciences and engineering: MBE*, vol. 17, no. 3, p. 2708, 2020.

[33] C. Geller, M. Varbanov, and R. E. Duval, "Human coronaviruses: Insights into environmental resistance and its influence on the development of new antiseptic strategies," *Viruses*, vol. 4, no. 11, pp. 3044–3068, 2012.

[34] A. R. Sahin, A. Erdogan, P. M. Agaoglu, *et al.*, "2019 novel coronavirus (covid-19) outbreak: A review of the current literature," *EJMO*, vol. 4, no. 1, pp. 1–7, 2020.

[35] C. Castillo-Chavez and B. Song, "Dynamical models of tuberculosis and their applications," *Mathematical Biosciences & Engineering*, vol. 1, no. 2, p. 361, 2004.

[36] S. Blanes, A. Iserles, and S. Macnamara, "Positivity-preserving methods for population models," *arXiv preprint arXiv:2102.08242*, 2021.

[37] S. Elaydi, *An Introduction to Di erence Equations*. Springer-Verlag, 2005, ISBN: 978-1-4419-2001-0.

[38] J. R. Dormand and P. J. Prince, "Runge-kutta triples," *Computers & Mathematics with Applications*, vol. 12, no. 9, pp. 1007–1017, 1986.

[39] I. Faragó, M. E. Mincsovics, and R. Mosleh, "Reliable numerical modelling of malaria propagation," *Applications of Mathematics*, vol. 63, no. 3, pp. 259–271, 2018.

[40] R. S. Varga, *Gersgorin and his circles*. Springer Science & Business Media, 2010, vol. 36.

[41] P. E. Kloeden and J. Schropp, "Runge–kutta methods for monotone differential and delay equations," *BIT Numerical Mathematics*, vol. 43, pp. 571–586, 2003.