Dynamically consistent nonstandard finite difference schemes for epidemiological models

R. Anguelov[†], Y. Dumont[*], J. M.-S Lubuma [†] and M. Shillor[+]

[†] Department of Mathematics and Applied Mathematics

University of Pretoria, Pretoria, South Africa

roumen.anguelov@up.ac.za, jean.lubuma@up.ac.za

[*] CIRAD, Umr AMAP, 34000 Montpellier, France

yves.dumont@cirad.fr

[+] Department of Mathematics and Statistics

Oakland University, Rochester, Michigan, USA

shillor@oakland.edu

March 19, 2013

Abstract

This work is the numerical analysis and computational companion of the paper by Kamgang and Sallet (*Math. Biosc.* **213** (2008), pp. 1–12) where threshold conditions for epidemiological models and the global stability of the disease-free equilibrium (DFE) are studied. We establish a discrete counterpart of the main continuous result that guarantees the global asymptotic stability (GAS) of the DFE for general epidemiological models. Then, we design nonstandard finite difference (NSFD) schemes in which the Metzler matrix structure of the continuous model is carefully incorporated and both Mickens' rules (World Scientific, Singapore, 1994) on the denominator of the discrete derivative and the nonlocal approximation of nonlinear terms are used in an innovative way. As a result of these strategies, our NSFD schemes are proved to be dynamically consistent with the continuous model, i.e., they replicate their basic features, including the GAS of the DFE, the linear stability of the endemic equilibrium (EE), the positivity of the solutions, the dissipativity of the system, and its inherent conservation law. The general analysis is made detailed for the MSEIR model for which the NSFD theta method is implemented, with emphasis on the computational aspects such as its convergence, or local truncation error. Numerical simulations that illustrate the theory are provided.

AMS Subject Classification (2010): 65L12, 37M99, 92D30

1   Introduction

Dynamical systems are used extensively in the modeling of many natural phenomena; they constitute a central component in applied mathematics and their numerical simulations are of fundamental importance in gaining the correct qualitative and quantitative information on the systems (e.g., [55]). In this work, we study a class of dynamical systems that arise in epidemiology as compartmental models for the spread of various diseases. These models are in the form of initial-value problems for n-dimensional differential systems,

$$\frac{dx}{dt} = A(x)\,x + f, \qquad x(0) = x_0, \tag{1}$$

where $x = x(t) : [0, \infty) \to \mathbb{R}^n$ represents the number or density of the populations in the different compartments, $x_0 \in \mathbb{R}^n$ is the vector of initial populations, and f is a given vector function that includes the recruitment or birth rates. Usually, $A(x)$ is a nonlinear real $n \times n$ Metzler matrix. Often, it is possible to distinguish among the components of x two sub-populations: the non-infected individuals (susceptible, recovered, etc.), represented by $y \in \mathbb{R}^{n_1}$, and the infected individuals, either latent or infectious, $z \in \mathbb{R}^{n_2}$ ($n_1 + n_2 = n$). Under suitable assumptions, classical mathematical theory asserts that for each x_0 the system (1) has a unique, positive, and maximal solution (see, e.g., [33] and [55]). Furthermore, it is possible to show that the system has equilibrium points, one of which is the Disease Free Equilibrium (DFE) that is the equilibrium state without infected individuals, which is important from the epidemiological point of view. Mathematically, we denote the stability number or the spectral bound of the Jacobian matrix J of the right-hand side of (1), evaluated at the DFE by $\alpha(J)$, i.e.,

$$\alpha(J) \equiv \max\{Re(\lambda) : \ \lambda \text{ eigenvalue of } J\}.$$

Then, it can be shown that when

$$\alpha(J) < 0, \tag{2}$$

the DFE is locally asymptotically stable. Condition (2) is in practice equivalent to the Kermack and McKendrick threshold condition ([34]),

$$\mathcal{R}_0 < 1,$$

where $\mathcal{R}_0$ is the so-called 'basic reproduction number' associated with (1), which essentially is a basic stability number. Moreover, a locally asymptotically stable Endemic Equilibrium (EE) may exist when $\mathcal{R}_0 > 1$. The local behaviour of the equilibrium states has been extensively studied in epidemiological models (see, e.g., [2, 19, 29, 56, 57]) and in general dynamical systems (see, e.g., [33, 55]).

In epidemiological applications, it is important to know whether the DFE is globally asymptotically stable (GAS) under certain threshold conditions. Such information, in case of an epidemic, may be used to design an intervention procedure that would decrease the spread of the disease and eventually eradicate it. However, proving that the DFE is GAS can be very difficult. Various ways were developed to that end and the best known is to construct a Lyapunov function for the system under consideration, [35]. However, to construct such Lyapunov functions is very challenging in most problems of interest. These considerations led Kamgang and Sallet [32] to seek another way to prove that the DFE is GAS. In their paper, they obtained a necessary and sufficient condition for the global asymptotic stability of the DFE under reasonable assumptions.

In this work we follow Kamgang and Sallet [32], and reformulate (1), using the notation $x = (y, z)$, in the following manner,

$$\begin{cases} \dfrac{dy}{dt} = A_1\left(x\right)\left(y - y^*\right) + A_{12}\left(x\right)z, \\ \dfrac{dz}{dt} = A_2\left(x\right)z, \end{cases} \tag{3}$$

where $A_1\left(x\right)$ and $A_2\left(x\right)$ are square matrices of dimensions $n_1 \times n_1$ and $n_2 \times n_2$, respectively, $A_{12}\left(x\right)$ is an $n_1 \times n_2$ matrix, and $(y^*, 0) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ is the DFE of (3). They proved the following theorem, which can also be found in [18] under some restrictive assumptions:

Theorem 1 *Consider the system* (3) *on a positively invariant set* $\Omega \subset \mathbb{R}^n_+$. *Let the following assumptions hold:*

H_1. *The system is dissipative on* Ω.

H_2. *The equilibrium* y^* *of the sub-system*

$$\frac{dy}{dt} = A_1(x)(y - y^*) \tag{4}$$

is globally asymptotically stable on the canonical projection of Ω *on* $\mathbb{R}^{n_1}_+$.

H_3. *The matrix* $A_2\left(x\right)$ *is a Metzler matrix and is irreducible for each* $x \in \Omega$.

H_4. *There exists an upper bound matrix* $\bar{A}_2$ *(in the sense of pointwise order) for the set* $\mathcal{M} = \{A_2(x) : x \in \Omega\}$ *such that: either* $\bar{A}_2 \notin \mathcal{M}$, *or* $\bar{A}_2 \in \mathcal{M}$ *and for each* $\bar{x} \in \Omega$ *satisfying* $\bar{A}_2 = A_2\left(\bar{x}\right)$ *necessarily* $\bar{x} \in \mathbb{R}^{n_1}_+ \times \{0\}$.

H_5. $\alpha\left(\bar{A}_2\right) \leq 0$.

Then, the DFE of (3) *is GAS in* $\overline{\Omega}$.

Kamgang and Sallet, [32], also presented applications of Theorem 1 to various epidemiological problems, such as models for Tuberculosis and HIV, and we note that it was used in the study of models for vector-borne diseases, [21, 22].

Our study is based on Theorem 1 and aims primarily to construct numerical schemes, for systems of type (3), that preserve the global asymptotic stability of the DFE. A key feature of this study is to establish a discrete counterpart of Theorem 1. This guarantees that our numerical schemes are reliable, being dynamically consistent with respect to a wide range of properties of the continuous system (3). These include the dissipativity of the discrete schemes, positivity and boundedness of their solutions, conservation laws, among others. We note that the concept of *topological dynamic consistency* has been introduced in [12, 13], following the works [1, 49]. To that end, we use the Nonstandard Finite Difference (NSFD) method, which was initiated more than two decades ago by R.E. Mickens and has, since the publication of his monograph [44], shown great potential in replicating the dynamics of systems that arise in a variety of areas in science and engineering (see, e.g., the edited volumes [46,47], the papers [4,7,25,26], and also [31] for an overview and applications in Biosciences). Furthermore, the usefulness and reliability of the NSFD method in epidemiology is attested by the ever growing number of works over the past few years (see, e.g., [21, 22, 27, 28, 53]). The novel NSFD schemes presented here are variants of the theta-method, extending the NSFD forward and backward Euler schemes as well as related schemes that have been widely used in the literature. Moreover, our schemes reinforce the use of both Mickens' rules regarding the denominator functions of the discrete derivatives and the nonlocal approximation of functions of dependent variables, the latter rule being implemented in an innovative way. Some of the results presented here were announced in the conference papers [3, 8, 9, 14].

The rest of this work is organized as follows. In the next section, we present a discrete analogue of Theorem 1 as a general result on the global asymptotic stability of the fixed-points of a general discrete dynamical system. In Section 3, we study general discrete approximations of epidemiological models. In view of the generality of these schemes, we make the underlying rules of the nonstandard approach more transparent in Section 4, within the concrete setting of the MSEIR model for which we investigate the nonstandard theta-method. We compare theoretically our results with some NSFD schemes that have been used in the literature. This comparative study is pursued computationally in Section 5, where we present various numerical simulations, which show the advantage of the nonstandard approach over the standard schemes such as the Runge-Kutta or BDF methods. Section 6 is devoted to concluding remarks. For the convenience of the reader, we summarize in Appendix 1 various facts about nonnegative matrices that are used repeatedly in the proofs. The proof of Theorem 2 can be found in Appendix 2.

2   GAS in discrete dynamical systems

Numerical computations of the equilibria of a dynamical system, such as (3), amount to an iteration process or a discrete dynamical system of the form

$$x^{k+1} = f(x^k), \tag{5}$$

where the given function $f : \Omega \subset \mathbb{R}^n_+ \to \mathbb{R}^n_+$ is sufficiently smooth. The study of (5) is quite involved because the Banach fixed-point theorem and related results, which are commonly used for such functional equations, do not apply. Indeed, even in the case of the continuous model (3), a technical result ([32, Lemma 4.1]) that is based on LaSalle's invariance principle [16, 37], had to be used to prove the main result on the GAS of the DFE. We start by establishing a discrete analogue of an important lemma, following some of the ideas in [16]. We recall that a system is *point dissipative* if there exists a positively invariant compact set K such that for each $x^0 \in \Omega$ the trajectory originating at x^0 is eventually in K. We have,

Theorem 2 *Let x^* be a fixed-point of the dynamical system (5). Assume that system (5) is point dissipative and that there exists a continuous function $V : \overline{\Omega} \to \mathbb{R}$ such that*

   (i) V is bounded from below on Ω;

   (ii) $V(f(x)) \leq V(x)$, $x \in \Omega$;

   (iii) The fixed-point x^* is GAS when system (5) is restricted to the set $\mathcal{L}$, which is the greatest invariant set contained in $\mathcal{E} = \{x \in \Omega : V(f(x)) = V(x)\}$.

Then, the fixed-point x^* is globally asymptotically stable on the whole set Ω.

For the sake of clarity of the presentation, the proof of Theorem 2 is provided in Appendix 2.

In what follows, we develop a method of proving that a fixed-point x^* which is on the boundary of the cone $\mathbb{R}^n_+$ is globally asymptotically stable. More precisely, we reduce the question of whether or not x^* is GAS to the question of whether the vector that contains only the nonzero components of x^* is GAS for the reduced system. To that end, we assume that x^* has n_2 components which are zero and let us rearrange the dimensions so that $x^* = (y^*, 0)^T$, where $n_1 = n - n_2$, and the n_1-dimensional vector $y^* = (y_1, ..., y_{n_1})^T$ has no zeros. Denoting $x = (y, z)$, $y \in \mathbb{R}^{n_1}$, $z \in \mathbb{R}^{n_2}$, we assume that system (5) has the following form, which is possible by using Taylor's expansion about $(y^*, 0)$:

$$
\begin{aligned}
y^{k+1} &= A_{1,h}(y^k, z^k)(y^k - y^*) + A_{12,h}(y^k, z^k)z^k, &\tag{6}\\
z^{k+1} &= A_{2,h}(y^k, z^k)z^k. &\tag{7}
\end{aligned}
$$

Here, h is a parameter (the time step), $A_{1,h}(x) = A_{1,h}(y, z)$, $A_{2,h}(x) \equiv A_{2,h}(y, z)$ and $A_{12,h}(x) = A_{12,h}(y, z)$ are matrices obtained from (3) and are of the same orders.

Our discrete analogue of Theorem 1, for the dynamical system (6)–(7), is:

Theorem 3 *Let the system (6)–(7) satisfy the following conditions:*

D_1. *The system is dissipative on $\Omega \subset \mathbb{R}_+^n$.*

D_2. *The fixed-point y^* of the sub-system*

$$y^{k+1} = A_{1,h}(y^k, 0)(y^k - y^*)$$

is GAS on the set $\Omega_1 = \{y \in \mathbb{R}^{n_1} : (y, 0) \in \Omega\}$.

D_3. *The matrix $A_{2,h}(y, z)$ is nonnegative for all $(y, z) \in \Omega$.*

D_4. *There exists an irreducible matrix $\overline{A}$, which is an upper bound for the set $\mathcal{M} = \{A_{2,h}(y, z) : (y, z) \in \Omega\}$.*

D_5. *The matrix $\overline{A}$ is such that $\rho(\overline{A}) \leq 1$. In the case when $\rho(\overline{A}) = 1$, we assume that $A_{2,h}(y, z) >> 0$, $\forall (y, z) \in \Omega$ and, in addition, z reduces to the zero vector in $\mathbb{R}^{n_2}$ if there exists a vector $(y, z) \in \Omega$ such that $A_{2,h}(y, z) = \overline{A}$.*

Then, the fixed-point point $(y^*, 0)$ of system (6)–(7) is GAS on Ω.

Proof. We use Theorem 2. Since $\overline{A}$ is irreducible and nonnegative, using the Perron-Frobenius theorem (Theorem A-17 in Appendix 1) shows that there exists a positive vector $v \in \mathbb{R}^{n_2}$ such that

$$v^T \overline{A} = \rho(\overline{A}) v^T. \tag{8}$$

The smooth function $V : \overline{\Omega} \to \mathbb{R}_+$, given by

$$V(x) = (0, v)^T x, \tag{9}$$

where $0 \in \mathbb{R}^{n_1}$, meets the requirements of Theorem 2. Indeed, writing the function $f(x)$ in (5) in the form (6)–(7) we have,

$$0 \leq V(f(x)) = v^T A_{2,h}(y, z) z \leq v^T \overline{A} z = \rho(\overline{A}) v^T z \leq v^T z = V(x), \tag{10}$$

by (8), (9), and (D_5). On the other hand, if $\mathcal{L}$ is the largest invariant subset of

$$\mathcal{E} = \{x = (y, z)^T \in \Omega : V(f(x)) = V(x)\} = \{x \in \Omega : v^T A_{2,h}(y, z) z = v^T z\}, \tag{11}$$

we show below that $\mathcal{L} \subseteq \{x = (y, z)^T \in \Omega : z = 0\} = \Omega_1 \times \{0\}$. To this end, we distinguish two cases in condition (D_5).

First, $\rho(\overline{A}) < 1$. If $(y, z)^T \in \mathcal{L}$, then, it follows from (10) and (11) that $\rho(\overline{A})v^T z = v^T z$, which implies that $v^T z = 0$. Since the coordinates of v are all positive and $z \geq 0$, we have $z = 0$. When $\rho(\overline{A}) = 1$, we fix $(y, z)^T \in \mathcal{E}$. If $A_{2,h}(y, z) = \overline{A}$, then (D5) implies that $z = 0$. But if $A_{2,h}(y, z) \neq \overline{A}$, then Theorem A-18 yields,

$$\rho(A_{2,h}(x)) < \rho(\overline{A}) \leq 1,$$

which implies that the matrix $A_{2,h}(x) - I$ is nonsingular, hence, $v^T(A_{2,h}(x) - I) \neq 0$. Since v is positive, (8) implies

$$v^T(A_{2,h}(x) - I) \leq v^T(\overline{A} - I) = 0.$$

It follows from (11) that
$$0 = v^T(A_{2,h}(x) - I))z.$$

Since $v^T(A_{2,h}(x) - I)) \leq 0$ and $z \geq 0$, a least one of the components of z is zero. Thus, if $(y, z) \in \mathcal{E}$ then z has a zero component. Now, let $(y, z) \in \mathcal{L}$, and since $\mathcal{L}$ is an invariant set we have that $f(y, z) \in \mathcal{L} \subseteq \mathcal{E}$, so the part $A_{2,h}(y, z)z$ of $f(y, z)$ has at least one zero component. Since all the entries of $A_{2,h}(y, z)$ are strictly positive, we have $z = 0$. This shows that condition (iii) in Theorem 2 holds. Consequently, $(y^*, 0)$ is a GAS fixed-point on Ω. ■

Remark 4 *The major advantage of Theorem 3 is that checking the GAS of the DFE is simplified mainly to computational algebra aspects for matrices that are involved in the discrete system, instead of using Lyapunov functions. In this regard, the hypotheses in Theorem 1 compare fairly well with those in Theorem 3, apart from some refined assumptions that occur in the discrete case. In particular, the difficulty of the discrete critical case $\rho(\overline{A}) = 1$ motivates the level of technical assumptions made in (D_5) as well as in the proof of that part of Theorem 1 that deals with the continuous critical case $\alpha(\overline{A}_2) = 0$ in (H_5). In practice, the critical case happens when the basic reproduction number $\mathcal{R}_0$ is 1.*

3 Nonstandard finite difference schemes

We consider a general system, which depending on the situation, will be used either in the form (1) or in the equivalent form (3). The validity of such systems as a models for the spread of diseases are of paramount importance, and the hypotheses (H_1)-(H_5) of Theorem 1 are motivated by the need to guarantee that their predictions are meaningful. More precisely, the validity of the model will be shown in three directions. The positivity of solutions and the dissipativity of the system depend on two explicit assumptions that occur often in applications. First, we assume for the system (1), that

$$\frac{dx_i}{dt} \geq 0 \ \text{ if } \ x_i = 0, \quad i = 1, \cdots, n. \tag{12}$$

This requirement, which is met by most epidemiological models, implies that the solutions of (1) or (3) satisfy the obvious requirement from a population:

$$x_i(t) \geq 0, \quad \forall t \geq 0, \qquad i = 1, \cdots, n.$$

Secondly, we assume that $A(x)$ in (1) is a compartmental matrix, which means the following ([30]): $A(x)$ is a Metzler matrix and the entries of each column satisfy,

$$\sum_{i=1}^{n} a_{ij}(x) \leq 0 \text{ for } j = 1, 2, ..., n, \quad x \in \mathbb{R}_+^n. \tag{13}$$

We require a bit more when the vector $f \geq 0$ in (1) is different from zero, namely, we assume the following structural condition: there exists $p = const. > 0$ such that

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}(x) x_j \leq -p \sum_{j=1}^{n} x_j \quad \text{for } x \in \mathbb{R}_+^n. \tag{14}$$

Such structural conditions and more general ones are often used to show the dissipativity of a dynamical system (e.g., [55]). Then, (13) implies that $a_{jj}(x) < 0$, since the off-diagonal entries of the matrix $A(x)$ are nonnegative. We define the total mass of the system by $M(t) = \sum_{i=1}^{n} x_i(t)$, let $F \equiv \sum_{i=1}^{n} f_i$, and introduce the number

$$p_F \equiv \begin{cases} p & \text{if } F > 0, \\ 0 & \text{if } F = 0. \end{cases} \tag{15}$$

By adding the equations in (1), the requirements (13) and (14) lead to the inequality

$$\frac{dM}{dt} \leq -p_F M + F, \tag{16}$$

which, following [50], is a conservation law. Several epidemiological models, especially those with constant overall population, satisfy condition (16).

We need the constant
$$K_F \equiv \begin{cases} \frac{F}{p_F} & \text{if } F > 0, \\ K & \text{if } F = 0, \end{cases} \tag{17}$$

where K denotes the carrying capacity of the total population when $F = 0$.

Our compartmental system (1) is dissipative because, by the Gronwall inequality, the conservation law (16) implies that

$$M(t) \leq M(0)e^{-p_F t} + K_F(1 - e^{-p_F t}) \leq K_F + M(0). \tag{18}$$

Thus, the requirement (H_1) in Theorem 1 is met and the biologically-feasible region of the model is:

$$\Omega = \left\{ x \in \mathbb{R}_+^n : \sum_{i=1}^{n} x_i \leq K_F + M(0) \right\}. \tag{19}$$

8

We note that finding solutions in the domain Ω, (19), is the first key point in the validation of both the continuous model (1) and the forthcoming discrete models if they are to describe the spread of diseases. Note also that when $F = 0$, the stated dissipativity of the model could be understood in the sense that the total mass $M(t)$ is nonincreasing along the forward trajectories of the system, as is seen in (16).

Having specified the general setting, we turn to construct approximations x^k of $x(t_k)$ at times $t_k = kh$, $k = 0, 1, 2, ...,$ where $h > 0$ is the time step. It is instructive to start with the conservation law (16). On taking $t = t_{k+1}$ and setting $M^{k+1} = M(t_{k+1})$ in (18), simple manipulations show that the *exact scheme* (see [44]) for (16) is either

$$\frac{M^{k+1} - M^k}{\frac{1 - e^{-hp_F}}{p_F}} \leq -p_F M^k + F, \tag{20}$$

or,

$$\frac{M^{k+1} - M^k}{\frac{e^{hp_F} - 1}{p_F}} \leq -p_F M^{k+1} + F, \tag{21}$$

where the nonnegative functions

$$\phi(h) \equiv \frac{1 - e^{-hp_F}}{p_F} \quad \text{or} \quad \phi(h) \equiv \frac{e^{hp_F} - 1}{p_F}, \tag{22}$$

satisfy

$$\phi(h) = h + O(h^2). \tag{23}$$

The two equivalent forms of the exact scheme suggest that we use a more complex denominator function satisfying (23), instead of the standard step size h, and that we consider explicit and implicit schemes for system (1).

First, we design the following implicit scheme:

$$\frac{x_i^{k+1} - x_i^k}{\phi(h)} = \sum_{j=1}^n a_{ij}(x^k) x_j^{k+1} + f_i, \quad i = 1, \cdots, n, \tag{24}$$

which has the vector formulation

$$\left(I - \phi(h) A(x^k)\right) x^{k+1} = x^k + f.$$

Given the compartmental nature of $A(x)$, for each $x \in \mathbb{R}_+^n$, we infer that

$$I - \phi(h) A(x) \text{ is an } M\text{-matrix} \tag{25}$$

(see Definition A-19). Thus,

$$x^k \geq 0 \implies x^{k+1} \geq 0. \tag{26}$$

By adding the equations in (24), we obtain a discrete counterpart of the conservation law, which is similar to (21), with a general denominator function $\phi(h)$ as in (23). For $M^k \leq K_F$, the discrete conservation law implies that

$$M^{k+1} \leq \frac{M^k}{1 + \phi(h) p_F} + \frac{\phi(h) F}{1 + \phi(h) p_F} \leq K_F,$$

which shows the dissipativity of the discrete dynamical system (24).

Second, we want to allow for the choice of an implicit-explicit scheme. To this end, we observe that by using $a_{jj}(x) < 0$, system (1) can be written component-wise as

$$\frac{dx_i}{dt} = -|a_{ii}(x)|x_i + \sum_{j \neq i}^{n} a_{ij}(x)x_j + f_i, \quad i = 1, \cdots, n. \tag{27}$$

Therefore, we propose a new scheme, which reads as

$$\frac{x_i^{k+1} - x_i^k}{\phi(h)} = -|a_{ii}(x^k)|u_i^k + \sum_{j \neq i}^{n} a_{ij}(x^k)u_j^k + f_i, \quad i = 1, \cdots, n, \tag{28}$$

where $u_j^k = x_j^k$ or $u_j^k = x_j^{k+1}$. When $u_j^k = x_j^k$, for all $j = 1, \ldots, n$, we obtain a fully explicit scheme, and when $u_j^k = x_j^{k+1}$ for all $j = 1, \ldots, n$, it is fully implicit. Otherwise, some of the variables are computed explicitly and the rest implicitly.

In this general setting, it is difficult to make a comprehensive statement about when x_j^k or x_j^{k+1} should be used for u_j^k in (28): this will be done in the examples in the following sections. But, we can provide the following general guidelines:

- Nonlinear terms involved in the diagonal entries are approximated in a nonlocal way (see [44]) such that

$$u_i^k = \begin{cases} x_i^{k+1} & \text{if } a_{ii}(x) \text{ depends on } x, \\ x_i^k & \text{if } a_{ii} \text{ does not depend on } x, \end{cases} \tag{29}$$

and the function ϕ in (23) is chosen such that

$$\phi(h) \leq \min \left\{ \frac{1}{p_F}, \frac{1}{|a_{ii}|} : \forall a_{ii} \text{ not depending on } x \right\}. \tag{30}$$

This guarantees the positivity property (26), which is necessary for the discrete model to be meaningful.

- Overall, the values of u_j^k are chosen, in view of (13), such that

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}(x^k)u_j^k \leq -p \sum_{j=1}^{n} x_j^k \quad \text{for } x \in \mathbb{R}_+^n. \tag{31}$$

The condition (31) can be achieved because, with the mass action principle or the standard incidence that is used in most epidemiological models, it makes sense to assume that for a given j, if $a_{jj}(x)$ is a function of x, then there exist $S(j)$ indexes $i_1, i_2, \cdots, i_{S(j)}$, which are independent of x such that $a_{i_s j}(x)$ together with $a_{jj}(x)$ are the only nonzero entries in the column j of the matrix $A(x)$ and

$$a_{jj}(x) + \sum_{s=1}^{S(j)} a_{i_s j}(x) = 0.$$

10

On adding the different equations in (28), condition (31) leads to a discrete conservation law of the form (20), which in turn yields the dissipativity of the scheme (28), because

$$M^{k+1} \leq (1 - \phi(h)p_F)M^k + \phi(h)F \leq K_F,$$

for $M^k \leq K_F$.

The convergence of the nonstandard finite difference schemes (24) and (28) is not a problem due essentially to the asymptotic relation (23). For instance, the scheme (28) is of order 1 as a consequence of this asymptotic relation and Taylor expansion, which show that the local truncation error T_{k+1} of the scheme is given by

$$T_{k+1} := \{I - \phi(h)A[x(t_k)]\}x(t_{k+1}) - x(t_k) - f = O(h),$$

where $x(t)$, assumed to be smooth enough, is the exact continuous solution. More importantly, we have established the dynamic consistency of these schemes:

Proposition 5 *Let $\phi(h)$ satisfy (23). Consider the additional assumptions (30) and (31) for the convergent explicit-implicit scheme (28). Then, this scheme and the implicit scheme (24) replicate the positivity and the dissipativity properties of the dynamical system (1) on Ω, (19), irrespective of the values of the time step h.*

Remark 6 *According to the approach of [44], formalized in [7], the two numerical methods (24) and (28) are non-standard finite difference schemes for the following reasons: (a) The standard denominator h of the discrete derivatives is replaced by the more complex function $\phi(h)$, which satisfies the requirement (23). (This denominator function is expected to capture the essential features of the dynamical system); (b) Nonlinear terms are approximated in a nonlocal way by using more than one point of the mesh.*

The second direction in the validation of the discrete models is to avoid spurious or ghost solutions in the sense of the next definition [7], which describes the minimum qualitative property that a reliable scheme should have.

Definition 7 *A numerical scheme is called* elementary stable *whenever it has no other fixed points than those of the continuous system it approximates, the local stability of these fixed points is the same for both the discrete and the continuous dynamical systems for each value of h.*

The nonstandard schemes (24) and (28) should, following the philosophy of [44], eliminate elementary instabilities in the first place. Since only the implicit scheme (24) is fully defined here, we restrict our analysis of elementary stability to it.

We assume that the only equilibrium states of model (1) are the disease-free equilibrium (DFE) and the endemic equilibrium (EE), and both are hyperbolic.

Let a function $\varphi : \mathbb{R} \to \mathbb{R}$, satisfy (23) and be such that

$$0 < \varphi(z) < 1 \qquad \text{for } z > 0. \tag{32}$$

Some possible choices are $\varphi(z) = 1 - e^{-z}$ or $\varphi(z) = z/(1 + z^2)$, see [10, Eq.(24)]. The dynamics of the model (1) can be captured by any number Q that satisfies

$$Q \geq \max \left\{ \frac{|\lambda|^2}{2|Re\lambda|} \right\}, \tag{33}$$

where λ denotes the eigenvalues of the Jacobian matrices of the right-hand side of (1) at the DFE and EE. The Jacobian matrices are assumed to be diagonalizable, which is the usual case in applications. The denominator function that is needed in (24) can be taken to be

$$\phi(h) = \frac{\varphi(Qh)}{Q}, \tag{34}$$

and for this denominator, we have the following result.

Theorem 8 *The NSFD scheme (24) is elementary stable whenever $\phi(h)$ is chosen according to (33) and (34).*

Proof. It is straightforward to check that the NSFD scheme (24) has no extra fixed-points than those of (1). Let x^* denote the DFE or EE of the system (1). Let $J \equiv J(x^*)$ be the Jacobian matrix of the right-hand side of this system at x^*.

Next, we evaluate the spectral radius of the matrix $(I - \phi(h)J)^{-1}$. Since J is not necessarily a Metzler matrix, we cannot use the results in Appendix 1, but J is diagonalizable so by using the factorization

$$\Lambda^{-1} J \Lambda = diag(\lambda_1, \lambda_2, \cdots \lambda_n),$$

where Λ is a transition matrix, we can circumvent this difficulty, following the approach in [4,5] or [40]. Setting $\varepsilon = x - x^*$, the linearization of the system (1) at x^* reads as

$$\frac{d\varepsilon}{dt} = J\varepsilon, \tag{35}$$

which is equivalent to the uncoupled system

$$\frac{d\eta}{dt} = diag(\lambda_1, \lambda_2, \cdots \lambda_n)\eta. \tag{36}$$

Thus, applying the NSFD scheme (24) to the system (35) or (36), we obtain the linearized schemes

$$\varepsilon^{k+1} = (I - \phi(h)J)^{-1}\varepsilon^k,$$

or

$$\eta^{k+1} = diag\left(\frac{1}{1 - \phi(h)\lambda_1}, \cdots, \frac{1}{1 - \phi(h)\lambda_n}\right)\eta^k. \qquad (37)$$

Now, if x^* is asymptotically stable for (1), then the real parts of all the eigenvalues λ_i are negative and it follows from (37) that

$$
\begin{aligned}
\rho\left((I - \phi(h)J)^{-1}\right) &= \max_{1 \le i \le n}\left\{\frac{1}{|1 - \phi(h)\lambda_i|}\right\} \\
&= \max_{1 \le i \le n}\left\{\frac{1}{\sqrt{1 + 2\phi(h)|Re\lambda_i| + \phi^2(h)|\lambda_i|^2}}\right\} < 1,
\end{aligned}
$$

which shows that x^* is asymptotically stable for the scheme (24).

Suppose now that x^* is unstable for (1). Then, there exists at least one eigenvalue of J, say λ_1, with positive real part. We then have

$$\frac{1}{|1 - \phi(h)\lambda_1|^2} = \frac{1}{1 - 2\phi(h)Re\lambda_1 + \phi^2(h)|\lambda_1|^2} > 1$$

whenever condition (33) holds. Therefore, x^* is unstable for the scheme (24). ■

The last step in the validation of the discrete models (24) and (28) is the study of the GAS of the DFE. The novelty here is that the analysis is reduced to the study of the properties of some Metzler matrices, in contrast with the classical approach that is based on Lyapunov functions. To this end, we assume that system (1), written in the equivalent form (3), has only one equilibrium, namely the DFE, $x^* = (y^*, 0)^T$. Implicitly, this means that the basic reproduction number is such that $\mathcal{R}_0 < 1$. For the sake of consistency of the continuous model (3) and the abstract scheme (6)-(7), we assume that the semi-explicit NSFD scheme (28) can be written equivalently as

$$
\begin{aligned}
\frac{y^{k+1} - y^k}{\phi(h)} &= A_1\left(x^k\right)(y^k - y^*) + A_{12}(x^k)z^k, \\
\frac{z^{k+1} - z^k}{\phi(h)} &= A_2\left(y^{k+1}, z^k\right)z^k,
\end{aligned} \qquad (38)
$$

where we omitted the subscript h, that also reads as

$$
\begin{aligned}
y^{k+1} - y^* &= B_1(h, x^k)(y^k - y^*) + B_{12}(h, x^k)z^k, \\
z^{k+1} &= B_2(h, y^{k+1})z^k.
\end{aligned} \qquad (39)
$$

Similarly, the implicit NSFD scheme (24) may be written as

$$
\begin{aligned}
\frac{y^{k+1} - y^k}{\phi(h)} &= A_1\left(x^k\right)(y^{k+1} - y^*) + A_{12}(x^k)z^{k+1}, \\
\frac{z^{k+1} - z^k}{\phi(h)} &= A_2\left(x^k\right)z^{k+1},
\end{aligned} \qquad (40)
$$

13

which is equivalent, in view of the property (25), to

$$
\begin{aligned}
y^{k+1} - y^* &= B_1(h, x^k)(y^k - y^*) + B_{12}(h, x^k)z^k, \\
z^{k+1} &= B_2(h, x^k)z^k.
\end{aligned}
\tag{41}
$$

We now make the following assumption regarding the matrix $A_1(x)$ in (3) and in the schemes (38) and (40):

$A_1(y, 0)$ is a Metzler matrix satisfying

$$
\alpha\left(A_1(y, 0)\right) \leq -\gamma < 0, \quad \forall y \in \Omega_1 = \left\{ y \in \mathbb{R}^{n_1}_+ : (y, 0) \in \Omega \right\},
\tag{42}
$$

for some $\gamma > 0$ that does not depend on y. As a consequence of (42), the requirement (H_2) of Theorem 1 is met: y^* is a GAS equilibrium of the reduced system (4) on Ω_1. This is actually the needed requirement, since (42) is a sufficient condition. We let the matrix $\underline{A}_2$ denote a lower bound for the set $\left\{A_2(x) : x \in \bar{\Omega}\right\}$ and start with the semi-explicit NSFD scheme.

Theorem 9 *Let the assumptions (H_1)-(H_5) of Theorem 1 hold. Furthermore, choose the time-step function $\phi(h)$ in (23) and (30) such that*

$$
0 < I + \phi(h)A_1(y, 0), \quad (y, 0) \in \Omega,
\tag{43}
$$

$$
0 < I + \phi(h)\underline{A}_2,
\tag{44}
$$

$$
0 < 1 + \phi(h)\alpha\left(\bar{A}_2\right) < 1.
\tag{45}
$$

Then, $x^* = (y^*, 0)$, as a fixed point of the explicit NSFD scheme (38) or (39) on Ω, is GAS.

Proof. We just have to check that the assumptions of Theorem 3 are satisfied.

(D_1) The dissipativity of the discrete system (38) or (39) on Ω, considered in its equivalent form (28), was established in Proposition 5.

(D_2) Since A_1 is a Metzler matrix, using the assumptions (42), (43), and Theorem A-24(b), we have

$$
\rho\left(I + \phi(h)A_1(y, 0)^T\right) < 1, \quad \forall(y, 0) \in \Omega,
$$

which implies that the fixed-point y^* is GAS for the linear sub-system

$$
\frac{y^{k+1} - y^k}{\phi(h)} = A_1(y^k - y^*).
$$

(D_3) This condition follows from assumption (44).

(D_4) This condition follows from assumption (H_4) in Theorem 1.

14

(D_5) Assumption (44) implies that if the function $\phi(h)$ is chosen such that $I + \phi(h)\,\bar{A}_2 > 0$ then using Theorem A-24(b) yields

$$\rho\left(I + \phi(h)\,\bar{A}_2\right) = 1 + \phi(h)\,\alpha\left(\bar{A}_2\right).$$

Now, there are two possibilities. Firstly, let $\alpha\left(\bar{A}_2\right) < 0$. Then, by the choice of $\phi(h)$ given in (45), we have D_5 in the form $\rho\left(I + \phi(h)\,\bar{A}_2\right) < 1$.

Secondly, if $\alpha\left(\bar{A}_2\right) = 0$, then $\rho\left(I + \phi(h)\,\bar{A}_2\right) = 1$. By (44), the matrix $I + \phi(h)\,A_2(y, z)$ is positive. Moreover, if $I + \phi(h)\,A_2(y, z) = I + \phi(h)\,\bar{A}_2$ i.e., $A_2(y, z) = \bar{A}_2$ for some $(y, z) \in \Omega$, then we have $z = 0$, as a result of the assumption H_4 in Theorem 1. In this way assumption D_5 holds.

The proof of the theorem is complete. ∎

Remark 10 *In light of (32) and (34), the requirement (45) is met for the choice*

$$\phi(h) = \frac{\varphi\left(-h\alpha\left(\bar{A}_2\right)\right)}{-\alpha\left(\bar{A}_2\right)}. \tag{46}$$

In practice, the matrix $\bar{A}_2$ is chosen in such a way that $\alpha(\bar{A}_2)$ is easy to compute. In most epidemiological models the condition $\alpha\left(\bar{A}_2\right) < 0$ is equivalent to the Kermack and McKendrick (see [34]) threshold condition $\mathcal{R}_0 < 1$. Therefore, it is not surprising that the denominator function $\phi(h)$, which is expected to reflect the dynamics of the system, depends on $\mathcal{R}_0$ and other parameters of the epidemiological models. This is in line with the "philosophy" of the nonstandard approach.

The next result deals with the implicit NSFD scheme, in which (23) is the only restriction placed on $\phi(h)$.

Theorem 11 *The disease-free equilibrium $x^* = (y^*, 0)$, viewed as a fixed point of the implicit NSFD scheme (40) or (41) on Ω, is GAS.*

Proof. We only need to check the assumptions (D_1)–(D_5) of Theorem 3.

(D_1) This is established in Proposition 5.

(D_2) When $z = 0$, we obtain the reduced system

$$y^{k+1} - y^* = \left(I - \phi(h)A_1(y^k, 0)\right)^{-1}(y^k - y^*),$$

of (41), for which y^* is a fixed-point. Moreover, it follows from Theorems A-24 and A-22, and (42), that

$$\rho\{(I - \phi(h)A_1(y, 0))^{-1}\} = \frac{1}{1 - \phi(h)\alpha\left(A_1(y, 0)\right)} \leq \frac{1}{1 + \gamma\phi(h)} < 1, \quad y \in \Omega_1.$$

15

Therefore, y^* is globally asymptotically stable in Ω_1.

(D_3) Conditions (H_3)–(H_5) of Theorem 1 are used here as follows. Since $A_2(y, z)$ is a Metzler matrix satisfying $\alpha(A_2(y, z)) \leq \alpha(\overline{A}_2) \leq 0$, an application of Theorem A-24 shows that $I - \phi(h)A_2(y, z)$ is an M-matrix. Then, $(I - \phi(h)A_2(y, z))^{-1} \geq 0$ by Theorems A-24 and A-25. Furthermore, $I - \phi(h)A_2(y, z)$ is irreducible because $A_2(y, z)$ is irreducible. Hence, $(I - \phi(h)A_2(y, z))^{-1} > 0$, a fact we will use below in the proof of item D_5. This shows that $(I - \phi(h)A_2(y, z))^{-1}$ is irreducible.

(D_4) Using the fact that $A_2(y, z) \leq \overline{A}_2$ (condition (H_4) of Theorem 1) we obtain $I - \phi(h)\overline{A}_2 \leq I - \phi(h)A_2(y, z)$. It follows from Theorem A-24(a) that the matrices on both sides of the inequality are M-matrices. Therefore, their inverses are nonnegative matrices (Theorem A-25). Then, applying A-(70), we obtain:

$$
\begin{aligned}
(I - \phi(h)\overline{A}_2)(I - \phi(h)A_2(y, z))^{-1} &\leq (I - \phi(h)A_2(y, z))(I - \phi(h)A_2(y, z))^{-1}, \\
(I - \phi(h)\overline{A}_2)(I - \phi(h)A_2(y, z)]^{-1} &\leq I, \\
(I - \phi(h)A_2(y, z))^{-1} &\leq (I - \phi(h)\overline{A}_2)^{-1}.
\end{aligned}
$$

(D_5) We apply Theorem A-24(a). From condition (H_5) of Theorem 1, we have that $\alpha(\overline{A}_2) \leq 0$, which implies that $\rho\{(I - \phi(h)\overline{A}_2)^{-1}\} < 1$ if $\alpha(\overline{A}_2) < 0$ and $\rho\{(I - \phi(h)\overline{A}_2)^{-1}\} = 1$ if $\alpha(\overline{A}_2) = 0$. In the latter case, the positivity of the matrix $(I - \phi(h)A_2(y, z))^{-1}$ was shown in the proof of item D_3. Assume that there exists $(y, z) \in \Omega$ such that $(I - \phi(h)A_2(y, z))^{-1} = (I - \phi(h)\overline{A}_2)^{-1}$ and thus $A_2(y, z) = \overline{A}_2$. Then, we must have $z = 0$ by the condition (H_4) of Theorem 1.

Since the requirements (D_1)–(D_5) of Theorem 3 are met, the fixed-point $(y^*, 0)$ is GAS on Ω. ∎

4 Application to the MSEIR model

In this section, we apply the theory of the previous sections to the MSEIR epidemiological model. First, we provide the Kamgang and Sallet formulation and decomposition of this model, with explicit Metzler matrix structure. Next, we construct for this model the explicit and implicit schemes of Section 3 and extend them to the NSFD theta-method, which we investigate in details. The MSEIR model is a very general model that was applied to various diseases, including those with less compartments, such as the SIR and the SEIR models. The flow diagram of the MSEIR model is depicted in Fig. 1 for the spread of a disease in a population of size $N = N(t)$ and consists of five compartments of: M - infants with passive immunity; S - susceptibles; E - exposed individuals; I - infectives; and R - recovered individuals.

In the standard incidence formulation, the equations of the MSEIR models read as follows (see for instance [29, 56]):

$$\frac{dM}{dt} = b(N - S) - (\delta + d)M,$$

$$\frac{dS}{dt} = bS + \delta M - \beta SI/N - dS,$$

$$\frac{dE}{dt} = \beta SI/N - (\varepsilon + d)E, \tag{47}$$

$$\frac{dI}{dt} = \varepsilon E - (\gamma + d)I,$$

$$\frac{dR}{dt} = \gamma I - dR$$
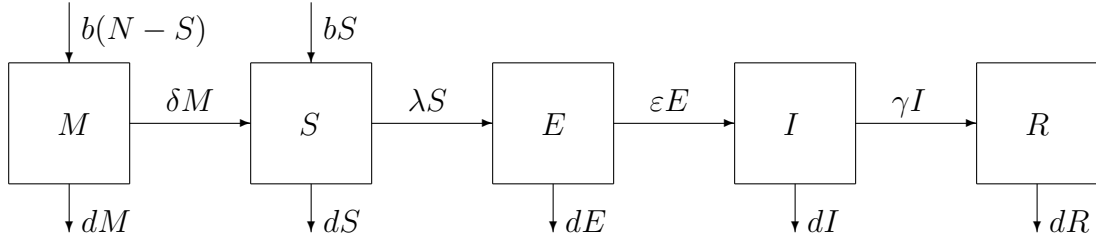
together with

$$\frac{dN}{dt} = (b - d)N.$$

Figure 1: Flow diagram of the MSEIR model

The parameters $d, q, \delta, \beta, \varepsilon$ and γ are positive numbers. In what follows, it is convenient to consider as dependent variables the fractions (divided by N) of the sub-populations in the compartments: $m = M/N, s = S/N, e = E/N, i = I/N$, and $r = R/N$. This reduces (47) to the following equivalent system for the MSEIR:

$$\frac{dm}{dt} = (d + q)(e + i + r) - \delta m,$$

$$\frac{ds}{dt} = -\beta si + \delta m,$$

$$\frac{de}{dt} = \beta si - (\varepsilon + d + q)e, \tag{48}$$

$$\frac{di}{dt} = \varepsilon e - (\gamma + d + q)i,$$

$$\frac{dr}{dt} = \gamma i - (d + q)r.$$

Equation (48) is supplemented with the initial conditions:

$$m(0) = m_0, \quad s(0) = s_0, \quad e(0) = e_0, \quad i(0) = i_0, \quad r(0) = r_0. \tag{49}$$

Our first task is to put the system (48) in the framework of Section 2 and to apply Theorem 1. Let $x = (m, s, e, i, r)^T \in \mathbb{R}_+^5$, then (48) can be rewritten as

$$\frac{dx}{dt} = A(x)x, \tag{50}$$

17

where

$$A\left(x\right) = \begin{pmatrix} -\delta & 0 & (d+q) & (d+q) & (d+q) \\ \delta & -\beta i & 0 & 0 & 0 \\ 0 & \beta i & -(\varepsilon+d+q) & 0 & 0 \\ 0 & 0 & \varepsilon & -(\gamma+d+q) & 0 \\ 0 & 0 & 0 & \gamma & -(d+q) \end{pmatrix},$$

is a compartmental matrix that satisfies the relations (13) and (14) in the form

$$\sum_{i=1}^{5} a_{ij}(x) = 0, \quad j = 1, \cdots, 5, \quad x \in \mathbb{R}_{+}^{5}.$$

Thus, the conservation law (16) reads as

$$\frac{d}{dt}(m+s+e+i+r) = 0, \tag{51}$$

and the *biologically feasible region* Ω is the simplex

$$\Omega = \{(m, s, e, i, r) \in \mathbb{R}_{+}^{5} : \quad m+s+e+i+r = 1\}. \tag{52}$$

The disease-free equilibrium is $DFE \equiv x^* = (0, 1, 0, 0, 0)^T$ and in terms of (3), the system (50) takes the equivalent formulation

$$\begin{aligned} \frac{dy}{dt} &= A_1\left(y, z\right)\left(y - y^*\right) + A_{12}(y, z)z, \\ \frac{dz}{dt} &= A_2\left(y, z\right)z, \end{aligned} \tag{53}$$

where $x = (y, z)^T$, $y = (m, s, r)^T$, $z = (e, i)^T$, $y^* = (0, 1, 0)^T$,

$$A_1\left(x\right) = \begin{pmatrix} -\delta & 0 & (d+q) \\ \delta & -\beta i & 0 \\ 0 & 0 & -(d+q) \end{pmatrix}, \quad A_{12} = \begin{pmatrix} d+q & d+q \\ 0 & -\beta \\ 0 & \gamma \end{pmatrix},$$

and

$$A_2\left(x\right) = \begin{pmatrix} -(\varepsilon+d+q) & \beta s \\ \varepsilon & -(\gamma+d+q) \end{pmatrix}.$$

The matrix $A_1(y, 0)$ is a Metzler matrix with non-positive eigenvalues,

$$\lambda_1 = 0, \quad \lambda_2 = -\delta \quad \text{and} \quad \lambda_3 = -(d+q).$$

Despite the fact that one of the eigenvalues vanishes and so $\alpha(A_1(y, 0)) = 0$, we claim that the equilibrium $y^* = (0, 1, 0)^T$ is GAS for the linear system

$$\frac{dy}{dt} = A_1\left(y, 0\right)\left(y - y^*\right). \tag{54}$$

18

Intuitively, the claim follows from the fact that the second equation in (54) is redundant, while the order two Metzler matrix

$$\tilde{A}_1(y, 0) = \begin{pmatrix} -\delta & (d+q) \\ 0 & -(d+q) \end{pmatrix},$$

associated with the first and third equations in (54) is stable. The proof follows from the explicit solution of the linear system (54), which is easily obtained by distinguishing the case $\delta \neq d+q$ from $\delta = d+q$. Thus, assumption (H_2) of Theorem 1 holds true.

On the other hand, $A_2(x)$ is an irreducible Metzler matrix such that

$$\underline{A}_2 \leq A_2(x) \leq \overline{A}_2,$$

for all $x \in \mathbb{R}_+^5$, where

$$\overline{A}_2 = \begin{pmatrix} -(\varepsilon + d + q) & \beta \\ \varepsilon & -(\gamma + d + q) \end{pmatrix}, \quad \underline{A}_2 = \begin{pmatrix} -(\varepsilon + d + q) & 0 \\ \varepsilon & -(\gamma + d + q) \end{pmatrix}.$$

Since the trace of the matrix $\overline{A}_2$ is negative, a necessary and sufficient condition for its eigenvalues to have non-positive real parts is that the determinant of $\overline{A}_2$ is nonnegative ([52]). This leads to the following criterion:

$$\alpha(\overline{A}_2) \leq 0 \quad \text{iff} \quad \mathcal{R}_0 \equiv \frac{\beta \varepsilon}{(\varepsilon + d + q)(\gamma + d + q)} \leq 1, \tag{55}$$

where $\mathcal{R}_0$ is the stability threshold parameter, or the stability number, or the basic reproduction number. Consequently, we are in the setting of Theorem 1 or other classical results, which guarantee the following well-known facts (see, e.g., [29, 56]):

Theorem 12 *The MSEIR model (48) defines a dissipative dynamical system on the set Ω given in (52). If the basic stability number satisfies $\mathcal{R}_0 \leq 1$, the dynamical system has only the disease-free equilibrium $DFE \equiv x^* = (0, 1, 0, 0, 0)$ that is GAS. If $\mathcal{R}_0 > 1$, there exists also an endemic equilibrium $EE \equiv x_e = (m_e, s_e, e_e, i_e, r_e)$, where*

$$
\begin{aligned}
s_e &= \frac{1}{\mathcal{R}_0}, \\
i_e &= \frac{\varepsilon \delta (d+q)}{(\delta + d + q)(\gamma + d + q)(\varepsilon + d + q)} \left(1 - \frac{1}{\mathcal{R}_0}\right), \\
e_e &= \frac{\delta (d+q)}{(\delta + d + q)(\varepsilon + d + q)} \left(1 - \frac{1}{\mathcal{R}_0}\right), \\
r_e &= \frac{\delta \varepsilon \gamma}{(\delta + d + q)(\gamma + d + q)(\varepsilon + d + q)} \left(1 - \frac{1}{\mathcal{R}_0}\right), \\
m_e &= \frac{d+q}{\delta + d + q} \left(1 - \frac{1}{\mathcal{R}_0}\right).
\end{aligned}
\tag{56}
$$

Then, the EE is locally asymptotically stable, while the DFE is unstable.

19

The second task ahead of us is to design and study reliable numerical schemes for the MSEIR model (48). The fully implicit scheme (24) for it is of the form

$$\frac{m^{k+1} - m^k}{\phi(h)} = -\delta m^{k+1} + (d+q)(r^{k+1} + e^{k+1} + i^{k+1}),$$

$$\frac{s^{k+1} - s^k}{\phi(h)} = -\beta s^{k+1} i^k + \delta m^{k+1},$$

$$\frac{e^{k+1} - e^k}{\phi(h)} = -(\varepsilon + d + q)e^{k+1} + \beta s^{k+1} i^k,$$

$$\frac{i^{k+1} - i^k}{\phi(h)} = \varepsilon e^{k+1} - (\gamma + d + q)i^{k+1} + i^k,$$

$$\frac{r^{k+1} - r^k}{\phi(h)} = -(d+q)r^{k+1} + \gamma i^{k+1},$$

whereas a possible explicit-implicit scheme contained in (28) is

$$\frac{m^{k+1} - m^k}{\phi(h)} = -\delta m^k + (d+q)(r^k + e^k + i^k),$$

$$\frac{s^{k+1} - s^k}{\phi(h)} = -\beta s^{k+1} i^k + \delta m^k,$$

$$\frac{e^{k+1} - e^k}{\phi(h)} = -(\varepsilon + d + q)e^k + \beta s^{k+1} i^k,$$

$$\frac{i^{k+1} - i^k}{\phi(h)} = \varepsilon e^k - (\gamma + d + q)i^k + i^k,$$

$$\frac{r^{k+1} - r^k}{\phi(h)} = -(d+q)r^k + \gamma i^k.$$

These two schemes are considered in [8]. However, they will be investigated as particular cases of a new family of schemes where the underlying guidelines stated in Section 3 on the nonlocal approximation of nonlinear terms and on complex denominator functions for the discrete derivatives are reinforced. The new family, an extension of the nonstandard theta-method developed in [40], is as follows: given $\theta, \hat{\theta} \in [0, 1]$, we consider the scheme

$$\frac{m^{k+1} - m^k}{\phi(h)} = -\delta[\theta m^{k+1} + (1-\theta)m^k] + \theta(d+q)(r^{k+1} + e^{k+1} + i^{k+1})$$
$$+ (1-\theta)(d+q)(r^k + e^k + i^k),$$

$$\frac{s^{k+1} - s^k}{\phi(h)} = -\beta[\hat{\theta} s^{k+1} i^k + (1-\hat{\theta})s^k i^{k+1}] + \delta[\theta m^{k+1} + (1-\theta)m^k],$$

$$\frac{e^{k+1} - e^k}{\phi(h)} = -(\varepsilon + d + q)[\theta e^{k+1} + (1-\theta)e^k] + \beta(\hat{\theta} s^{k+1} i^k + (1-\hat{\theta})s^k i^{k+1}), \quad (57)$$

$$\frac{i^{k+1} - i^k}{\phi(h)} = \varepsilon[\theta e^{k+1} + (1-\theta)e^k] - (\gamma + d + q)[\theta i^{k+1} + (1-\theta)i^k],$$

$$\frac{r^{k+1} - r^k}{\phi(h)} = -(d+q)[\theta r^{k+1} + (1-\theta)r^k] + \gamma[\theta i^{k+1} + (1-\theta)i^k],$$

20

where the denominator function of the discrete derivatives is chosen according to (23), with

$$\phi(h) = h + O(h^3) \quad \text{if} \ \ \theta = \hat{\theta} = 1/2. \tag{58}$$

Let us note the nonlocal approximation of the nonlinear terms in the second and the third equations. It is a weighted average of two nonlocal approximations $s^{k+1}i^k$ and $s^k i^{k+1}$. Since this approximation is essentially different from the weighted average of values at t_k and t_{k+1}, which is applied to all other terms, we use a parameter $\hat{\theta}$, which need not equal θ. This is the main difference with [40] and [10].

By adding the equations in (57), we find that the NSFD scheme replicates the conservation law contained in (52), indeed,

$$m^{k+1} + s^{k+1} + e^{k+1} + i^{k+1} + r^{k+1} = m^k + s^k + e^k + i^k + r^k = 1. \tag{59}$$

In what follows, we use (59) and substitute $i^{k+1} = 1 - m^{k+1} - s^{k+1} - r^{k+1} - e^{k+1}$ in the second equation of (57). Therefore, in practice, the implementation at each step of the NSFD scheme (57) amounts to solving the linear system

$$\tilde{C}(i^k, s^k) \left(m^{k+1} \ \ s^{k+1} \ \ r^{k+1} \ \ e^{k+1} \ \ i^{k+1} \right)^T = \tilde{D} \left(m^k \ \ s^k \ \ r^k \ \ e^k \ \ i^k \right)^T, \tag{60}$$

where $\tilde{C}(i^k, s^k)$ is the matrix

$$
\begin{pmatrix}
1+\theta\phi\delta & 0 & -\theta\phi(d+q) & -\theta\phi(d+q) & -\theta\phi(d+q) \\
-\theta\phi\delta-(1-\hat{\theta})\phi\beta s^k & 1+\hat{\theta}\phi\beta i^k-(1-\hat{\theta})\phi\beta s^k & -(1-\hat{\theta})\phi\beta s^k & -(1-\hat{\theta})\phi\beta s^k & 0 \\
0 & 0 & 1+\theta\phi(d+q) & 0 & -\theta\phi\gamma \\
0 & -\hat{\theta}\phi\beta i^k & 0 & 1+\theta\phi(\varepsilon+d+q) & -(1-\hat{\theta})\phi\beta s^k \\
0 & 0 & 0 & -\theta\phi\varepsilon & 1+\theta\phi(\gamma+d+q)
\end{pmatrix}
$$

and $\tilde{D}$ is the matrix

$$
\begin{pmatrix}
1-(1-\theta)\phi\delta & 0 & (1-\theta)\phi(d+q) & (1-\theta)\phi(d+q) & (1-\theta)\phi(d+q) \\
(1-\theta)\phi\delta & 1-(1-\hat{\theta})\phi\beta & 0 & 0 & 0 \\
0 & 0 & 1-(1-\theta)\phi(d+q) & 0 & (1-\theta)\phi\gamma \\
0 & 0 & 0 & 1-(1-\theta)\phi(\varepsilon+d+q) & 0 \\
0 & 0 & 0 & (1-\theta)\phi\varepsilon & 1-(1-\theta)\phi(\gamma+d+q)
\end{pmatrix}.
$$

Here and below, we assume that the denominator function ϕ satisfies

$$\phi(h) < \left(\max\{(1-\theta)\delta, (1-\theta)(\varepsilon+d+q), (1-\theta)(\gamma+d+q), (1-\hat{\theta})\beta\} \right)^{-1}. \tag{61}$$

Typically, this means that we may choose $\phi(h)$ as in (34), with

$$Q \geq (1-\theta)\max\{(1-\theta)\delta, (1-\theta)(\varepsilon+d+q), (1-\theta)(\gamma+d+q), (1-\hat{\theta})\beta\}. \tag{62}$$

Under the condition (61) and with $i^k \geq 0$ and $s^k \geq 0$, we note that $\tilde{C}$ is an M-matrix, since the transpose matrix $\tilde{C}^T$ is strictly diagonally dominant (see Theorem A-23).

21

By Theorem A-25, its inverse is nonnegative, $\tilde{C}^{-1} \geq 0$. Moreover, condition (61) is sufficient for the discrete scheme to preserve the positivity of solutions because $\tilde{D} \geq 0$. This, together with the property (59), implies that the scheme (57) defines a discrete dynamical system on the same domain Ω, (52), as the domain of dynamical system defined by the MSEIR model (48). Further, it is easy to see by substitution that DFE is an equilibrium of the system (57).

Remark 13 *As mentioned earlier, the analysis of the methods carried out in [8] is restricted to $\hat{\theta} = 1$, with $\theta = 0$ for the implicit-explicit scheme, and $\theta = 1$ for the fully implicit scheme of Section 3. Actually, when $\theta = \hat{\theta} = 1$ the inequality (61) always holds and can be disregarded when considering the choice of ϕ. In the absence of any other restriction, one can just take $\phi(h) = h$.*

In general, one should not expect the parameters θ and $\hat{\theta}$ to vary independently from one another if the NSFD scheme is to be dynamically consistent. The next theorem shows that the scheme (57) preserves the GAS of the DFE of (48) (in Ω), as established in Theorem 12, under some functional relationship between the values of the parameters θ and $\hat{\theta}$.

Theorem 14 *Define $\phi(h)$ according to (32), (34), and (62) so that (61) holds. Let θ and $\hat{\theta}$ be such that either $\theta = 0$ or $\theta + \hat{\theta} \geq 1$. Then, the DFE $x^* = (0, 1, 0, 0, 0)$ of the system (57) in Ω is GAS whenever $\mathcal{R}_0 < 1$.*

Proof. We apply Theorem 3, which as was emphasized above, relies on the properties of the matrices involved in the formulation of the NSFD scheme. We need to show that conditions (D_1)–(D_5) hold. Since the domain Ω of the system (57) is compact, condition (D_1) is satisfied. For the rest of the conditions, we write the system in the form (6) and (7) where $y = (m, s, r)$ and $z = (e, i)$. The reduced system obtained from (57), when $z = 0$, is

$$
\begin{pmatrix} r^{k+1} \\ m^{k+1} \end{pmatrix} = \begin{pmatrix} \dfrac{1 - (1-\theta)(d+q)\phi(h)}{1 + \theta(d+q)\phi(h)}, & 0 \\[3ex] \dfrac{1 + \theta(d+q)\phi(h)}{[1 + \theta\delta\phi(h)][1 + \theta(d+q)\phi(h)]}, & \dfrac{1 - (1-\theta)\delta\phi(h)}{1 + \theta\delta\phi(h)} \end{pmatrix} \begin{pmatrix} r^k \\ m^k \end{pmatrix}
$$

$$
= B \begin{pmatrix} r^k \\ m^k \end{pmatrix},
$$

coupled with

$$
s^{k+1} = 1 - m^{k+1} - r^{k+1}.
$$

Since $\rho(B) < 1$, the sequence (r^k, m^k) converges to $(0, 0)$ as $k \to \infty$. Then, s^k converges to 1, which implies that y^k converges to $y^* = (0, 1, 0)$, and the attracting

22

equilibrium is also stable, hence (D_2) holds. In order to deal with (D_3)–(D_5), we need to write the matrix in (7) in (53_2) explicitly. The third and fourth equations in (57) yield

$$E(s^k)z^{k+1} = G(s^{k+1})z^k, \tag{63}$$

where $z^k = (e^k, i^k)$ and

$$E(s^k) = \begin{pmatrix} 1 + \theta\phi(\varepsilon + d + q) & -(1-\hat{\theta})\phi\beta s^k \\ -\theta\phi\varepsilon & 1 + \theta\phi(\gamma + d + q) \end{pmatrix},$$

$$G(s^{k+1}) = \begin{pmatrix} 1 - (1-\theta)\phi(\varepsilon + d + q) & \hat{\theta}\phi\beta s^{k+1} \\ (1-\theta)\phi\varepsilon & 1 - (1-\theta)\phi(\gamma + d + q) \end{pmatrix}.$$

We have

$$\det(E(s^k)) = 1 + \theta\phi(\varepsilon + \gamma + 2d + 2q) + \theta\phi^2(\varepsilon + d + q)(\gamma + d + q)(\theta - (1-\hat{\theta})\mathcal{R}_0 s^k).$$

Therefore, using the assumptions of the theorem, the last term is nonnegative so,

$$\det(E(s^k)) \geq 1.$$

Hence, system (63) can be written as

$$z^{k+1} = A(s^k, s^{k+1})z^k = (E(s^k))^{-1}G(s^{k+1})z^k.$$

The matrix $A(s^k, s^{k+1})$ is a function of m^k, s^k, r^k, e^k, i^k due to (60). The assumptions on ϕ show that $E(s^k)$ is an M-matrix and thus $(E(s^k))^{-1}$ is nonnegative. Moreover, the matrix $G(s^{k+1})$ is also nonnegative, and so $A(s^k, s^{k+1})$ is nonnegative and (D_3) holds.

Next, we obtain a bound for $A(s^k, s^{k+1})$ by finding upper bounds on $(E(s^k))^{-1}$ and $G(s^{k+1})$. Let $\underline{E} = \min_{0 \leq s^k \leq 1} E(s^k)$ so $\underline{E}$ and $E(s^k)$ are M-matrices such that $\underline{E} \leq E(s^k)$, which implies that $(E(s^k))^{-1} \leq (\underline{E})^{-1}$ ([17]), i.e.,

$$(E(s^k))^{-1} \leq \frac{1}{\det(\underline{E})} \begin{pmatrix} 1 + \theta\phi(\gamma + d + q) & (1-\hat{\theta})\phi\beta \\ \theta\phi\varepsilon & 1 + \theta\phi(\varepsilon + d + q) \end{pmatrix},$$

where

$$\underline{E} = \begin{pmatrix} 1 + \theta\phi(\varepsilon + d + q) & -(1-\hat{\theta})\phi\beta \\ -\theta\phi\varepsilon & 1 + \theta\phi(\gamma + d + q) \end{pmatrix}.$$

Also,

$$G(s^{k+1}) \leq \overline{G} = \begin{pmatrix} 1 - (1-\theta)\phi(\varepsilon + d + q) & \hat{\theta}\phi\beta \\ (1-\theta)\phi\varepsilon & 1 - (1-\theta)\phi(\gamma + d + q) \end{pmatrix}.$$

Then, it follows from A-(70) that

$$A(s^k, s^{k+1}) \leq \overline{A} = (\underline{E})^{-1}\overline{G}.$$

23

It is easy to see that $\overline{A}$ is irreducible because the respective non-diagonal entries of $(\underline{E})^{-1}$ and $\overline{G}$ cannot vanish simultaneously. Therefore (D_4) holds.

We show that $\rho(\overline{A}) < 1$ by using the well-known Jury conditions ([21, 52]): The roots λ_1 and λ_2 of a polynomial $\lambda^2 + a_1\lambda + a_2$, $a_i \in \mathbb{R}$, verify $|\lambda_i| < 1$ iff

$$1 - a_2 > 0, \qquad 1 - a_1 + a_2 > 0, \qquad 1 + a_1 + a_2 > 0. \tag{64}$$

The characteristic equation of $\overline{A}$ has $a_1 = -\mathrm{tr}(\overline{A})$ and $a_2 = \det(\overline{A})$. Since,

$$\det(\underline{E}) = 1 + \theta\phi(\varepsilon + \gamma + 2d + 2q) + \theta\phi^2(\varepsilon + d + q)(\gamma + d + q)(\theta - (1 - \hat{\theta})\mathcal{R}_0) \geq 1,$$

and

$$\det(\overline{G}) = (1 - (1 - \theta)\phi(\varepsilon + d + q))(1 - (1 - \theta)\phi(\gamma + d + q)) - (1 - \theta)\hat{\theta}\phi^2\beta\varepsilon \leq 1,$$

we obtain

$$a_2 = \det(\overline{A}) = \frac{\det(\overline{G})}{\det(\underline{E})} < 1, \tag{65}$$

and the inequality is strict since $\det(\underline{E}) = 1$ holds only if $\theta = 0$, while $\det(\overline{G}) = 1$ only if $\theta = 1$.

Since the trace of $\overline{A}$ is nonnegative, it remains to prove the last inequality in (64), or equivalently that

$$\Lambda \equiv \det(\underline{E}) - \det(\underline{E})\mathrm{trace}(\overline{A}) + \det(\overline{G}) > 0. \tag{66}$$

Straightforward computations yield,

$$\begin{aligned} \Lambda &= \phi^2(\varepsilon + d + q)(\gamma + d + q) - \phi^2\beta\varepsilon \\ &= \phi^2(\varepsilon + d + q)(\gamma + d + q)(1 - \mathcal{R}_0) > 0, \end{aligned}$$

which shows that (66) holds. Therefore, (64) holds implying $\rho(\overline{A}) < 1$. Hence, condition (D_5) is satisfied. The result follows now from Theorem 3. ∎

Remark 15 *For $\hat{\theta} = \theta$, the NSFD scheme (57) for the MSEIR model is elementary stable as stated in Theorem 8. The proof of Theorem 8 can be easily adapted to this case.*

As far as the accuracy of the NSFD scheme is concerned, its local truncation error can be obtained from the Taylor expansion and the symmetric nature of the scheme when $\theta = \hat{\theta} = 1/2$, as in the case of the classical theta-method in [40] or [10]. This yields the following result:

Theorem 16 *The NSFD scheme (57) is second-order convergent for $\theta = \hat{\theta} = 1/2$ under the assumption (58) and is first-order convergent for all other values of $\theta, \hat{\theta} \in [0, 1]$.*

5   Numerical Simulations

In this section, we depict numerical simulations using schemes from the family (57) for the MSEIR model. Some of the results were announced in the conference papers [8,9]. We consider four schemes:

(A) Explicit scheme obtained from (57) for $\theta = 0$ and $\hat{\theta} = 1$. The function ϕ is chosen according to (32), (34), and (62) so that (61) holds for these values of θ and $\hat{\theta}$. More precisely, we choose

$$\phi(h) = \frac{1 - e^{-Qh}}{Q}, \qquad Q = \max\{\delta, \varepsilon + d + q, \gamma + d + q\}. \qquad (67)$$

(B) Implicit scheme obtained from (57) for $\theta = \hat{\theta} = 1$. As mentioned earlier, in this case inequality (61) does not impose any restriction on ϕ, so we choose $\phi(h) = h$.

(C) Scheme (57) for $\theta = \hat{\theta} = 0$. This is an explicit scheme similar to (A), but with a different approximation of the nonlinear term (see [9] for further details) and

$$\phi(h) = \frac{1 - e^{-Qh}}{Q}, \qquad Q = \max\{\delta, \varepsilon + d + q, \gamma + d + q, \beta\}. \qquad (68)$$

(D) Second-order scheme obtained from (57) for $\theta = \hat{\theta} = \frac{1}{2}$. In addition to (61) the denominator function has to satisfy (58). We choose

$$\phi(h) = \frac{\tanh(Qh)}{Q}, \qquad Q = \frac{1}{2}\max\{\delta, \varepsilon + d + q, \gamma + d + q, \beta\}. \qquad (69)$$

In the first series of simulations below we used the following parameter values:

$$\beta = 0.14, \ \delta = 1/180, \ \epsilon = 1/14, \ \gamma = 1/7, \ d = 1/(40 * 365), \ q = 0.$$

For these values we have
$$\mathcal{R}_0 = 0.97857.$$

Hence (Theorem 14) the DFE $(0, 1, 0, 0, 0)$ is GAS. The numerical approximations of those variables that tend to zero, i.e., m (immune), r (recovered), e (exposed), and i (infective), were obtained, respectively, by the schemes (A)–(D) and are plotted in the four graphs in Fig. 2. One can observe that, indeed, the properties of the numerical solutions as stated in Theorem 14 hold. In order to emphasize the fact that Theorem 14 holds for all step sizes h, the top two graphs in Fig. 2 were obtained with $h = 2$ while the bottom two with $h = 10$. Naturally, the step size may affect the accuracy, but it is seen that it does not affect the qualitative behavior of the numerical solutions.
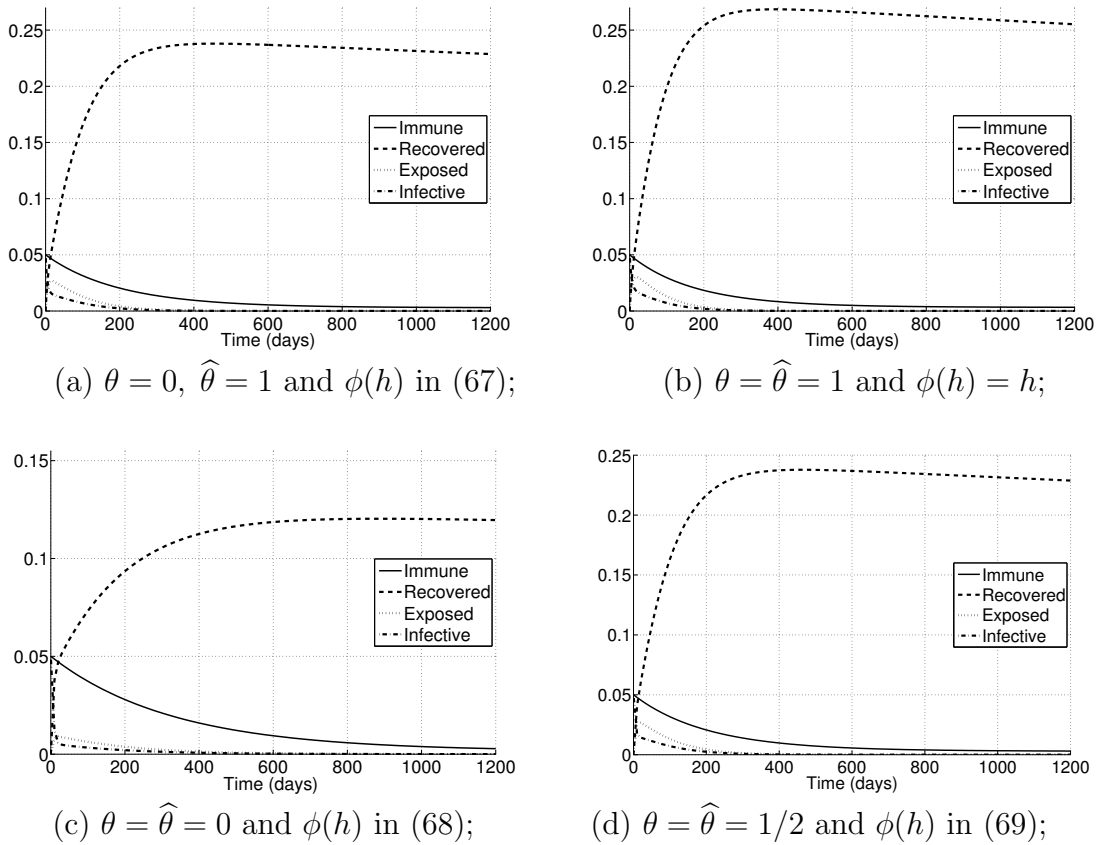
(a) $\theta = 0$, $\widehat{\theta} = 1$ and $\phi(h)$ in (67);

(b) $\theta = \widehat{\theta} = 1$ and $\phi(h) = h$;

(c) $\theta = \widehat{\theta} = 0$ and $\phi(h)$ in (68);

(d) $\theta = \widehat{\theta} = 1/2$ and $\phi(h)$ in (69);

Figure 2: Numerical solutions of the MSEIR model with scheme (57) with different values of θ, $\widehat{\theta}$, and $\phi(h)$.

Since the model assumes that the recovered r have lifelong immunity, the rate at which r tends to zero is determined by the birth/death rate. If a generation spans 40 years, in absence of infection the time period for r to vanish is much longer than that for the other vanishing variables.

(i)

(ii)
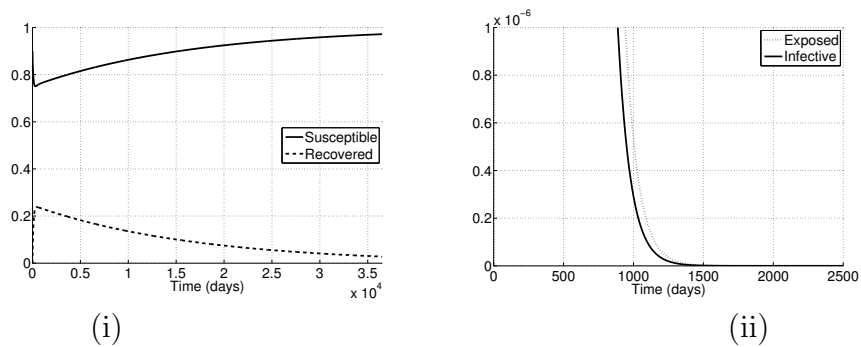
Figure 3: Numerical solution using $\theta = \widehat{\theta} = 1/2$ and $\phi(h)$ in (69): (i) r and s; (ii) e and i.

Fig. 3 depicts: (i) the computed approximations of $r(t)$ and $s(t)$ as they approach 0 and 1, respectively; (ii) even though the approximations of $e(t)$ and $i(t)$ are near zero they remain positive, i.e., they do not leave the domain of the dynamical system. We used Scheme D and $h = 10$. The graphs of the simulations done by the other

methods show similar qualitative behavior. The computational order of convergence of scheme D, tabulated in Table 1 for various step-sizes h, confirms the second-order convergence stated in Theorem 16. Here x_h denotes the approximation of the solution $x = (m, s, e, i, r)$ computed on a mesh with step size h and the norm is the composition of the Euclidian norm in space and the supremum norm in time.

h	4	2	1	0.5	0.25
$\log_2 \dfrac{\|\|x_{2h} - x_h\|\|}{\|\|x_h - x_{h/2}\|\|}$	1.905557	1.975479	1.993810	1.998449	1.999612

Table 1: Numerical order of convergence of scheme D

Finally, we note that, apart from preserving the GAS of DFE when $\mathcal{R}_0 < 1$ on which we focused most, the NSFD schemes (57) preserve the biologically relevant region Ω of the dynamical system as predicted in Theorem 14. In particular, the discrete solutions are positive and bounded from above by 1. It may happen that standard methods also preserve the stated properties of the MSEIR model. However, in general, this cannot be guaranteed or at least cannot be guaranteed for all step sizes. Examples to that effect for similar systems can be found in [12]. A general discussion of the issue can be found in [7, 46, 48]. In Fig. 4, we provide two illustrations of how these properties may be violated. The two graphs depict numerical solutions obtained via: (i) the two stage second order Runge-Kutta method with $h = 10$ and (ii) by the standard MATLAB routine ode45 ([54]) with adaptive time stepping.
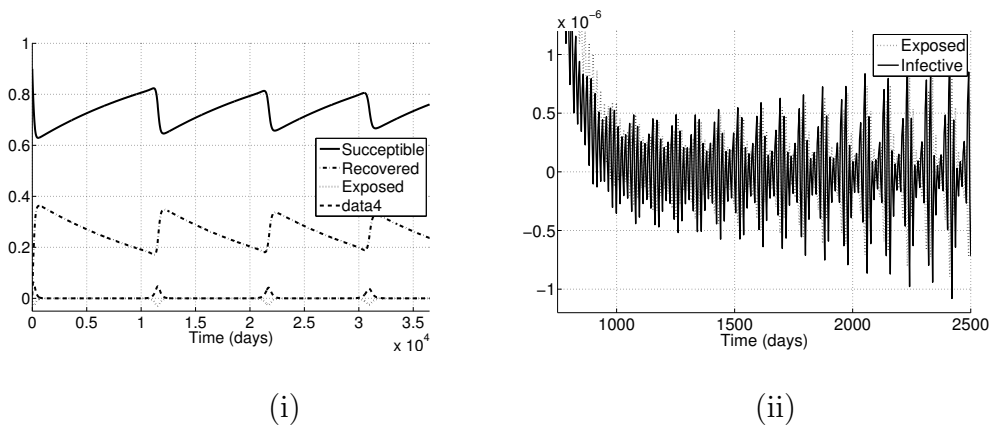

Figure 4: (i) Second order Runge-Kutta; (ii) Matlab routine ode45.

In the last simulation below (Fig. 5), obtained with the explicit scheme (D), we used the same parameter values as above for $\delta, \varepsilon, \gamma,$ and q but we choose $\beta = 0.17$ so that $R_0 \approx 1.1883$. For each h, the nonstandard scheme provides good approximations and converges to the endemic equilibrium, as predicted in Theorem 12.
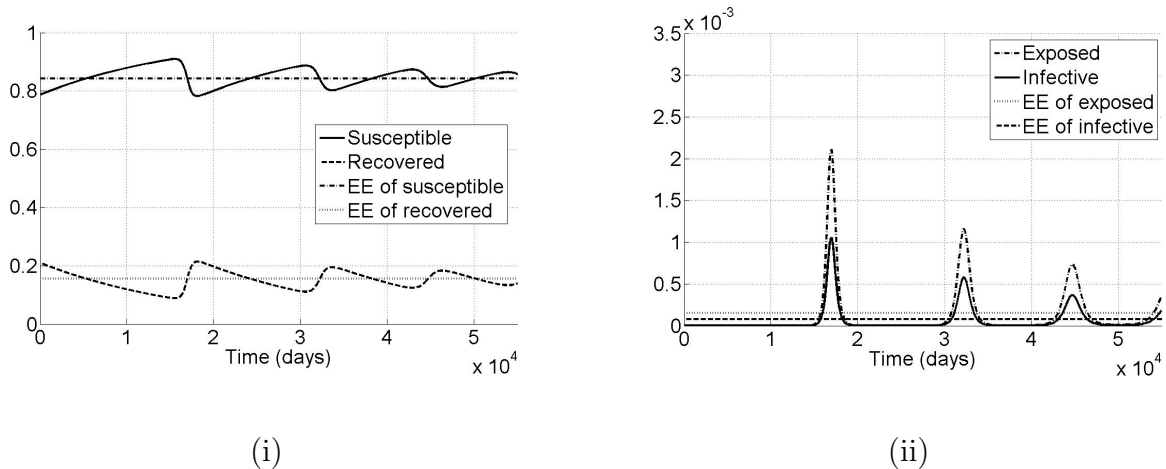
(i)   (ii)

Figure 5: Nonstandard scheme (D) for $\mathcal{R}_0 > 1$ and $h = 10$: (i) r and s; (ii) e and i.

Additional simulations of the MSEIR model are provided in [9]. In all cases, the CPU times with the ODEs routines (in particular the Runge-Kutta method) were very large, while it took only a few seconds for the nonstandard scheme, depending on h. Moreover, in the case when $\beta = 0.2$ the non-stiff routine blew-up.

6   Conclusions

This work was motivated by the paper [32] where sufficient and realistic conditions can be found for the global asymptotic stability of the disease-free equilibrium of epidemiological models associated with a class of specific matrices such as Metzler matrices. Here, we investigated how these conditions can be transposed to the discrete and computational settings. The investigation had three main steps.

In the first step, we provided the discrete counterparts of the continuous conditions in [32] for an abstract discrete dynamical system and showed that these discrete conditions guarantee the global asymptotic stability of the disease-free equilibrium, viewed as a fixed-point of the discrete dynamical system.

In the second step, still within the general setting, we designed nonstandard finite difference schemes that can be fully implicit, fully explicit, or explicit in some dependent variables and implicit in the remaining ones. The fully explicit and fully implicit schemes are nonstandard versions of the forward method and backward Euler method, respectively. The NSFD schemes are proved to be dynamically consistent with the properties of the continuous model in [32], irrespective of the values of the time step h. These properties include the positivity of the solutions, linear stability of the equilibria, global asymptotic stability of the disease-free equilibrium, dissipativity of the system, and the related conservation law.

28

In the third step, we used the general MSEIR epidemiological model as an example making the theoretical results in the previous two steps concrete. The explicit-implicit nonstandard method of the previous step was applied to this model. By considering the weighted average of the explicit-implicit method through two parameters $\theta, \hat{\theta} \in [0, 1]$, we designed a family of dynamically consistent NSFD schemes, namely the theta-methods. The schemes are second-order convergent when $\theta = \hat{\theta} = 1/2$. Special emphasis was placed on the extreme cases $\theta = 0, \quad \hat{\theta} = 1$ and $\theta = 0 =, \hat{\theta}$, both explicit methods as well as $\theta = \hat{\theta} = 1$ which is an implicit method, with regard to the need of carefully choosing the denominator function of the discrete derivative. With such a choice, we showed in numerical tests in Section 5 the superiority of the NSFD schemes over their classical counterparts.

Our plans for future studies include the extension of this work to other compartmental models, with non-compartmental matrices (see [22, 23] for a first attempt in this direction). In this case, the dissipativity of the continuous model, which is an essential assumption of the theory developed in [32], should be addressed differently. Furthermore, the design of higher-order NSFD schemes is an issue of great interest, which has been partly investigated in [39].

Acknowledgements

References

[1] H. Al-Kahby, F. Dannan and S. Elaydi, Non-standard discretization methods for some biological models, In: R.E. Mickens (Ed.), *Applications of nonstandard finite difference schemes*, World Scientific, Singapore, 2000, pp. 155–180.

[2] R. M. Anderson and R. M. May, *"Infectious Diseases of Humans: Dynamics and Control,"* Oxford University Press, 1991.

[3] R. Anguelov, Dynamically Consistent Schemes for Epidemiological Models, *Proceedings of the Annual Meeting of the Bulgarian Section of SIAM, 22-23 December 2008, Sofia*, Demetra, Sofia, 2009, 11–14

[4] R. Anguelov, J.K. Djoko, P. Kama and J.M.-S. Lubuma, On finite difference schemes having the correct linear stability and dissipative properties of dynamical systems, *University of Pretoria Technical Report, No. UPWT 2007/02.*

[5] R. Anguelov, J.K. Djoko, P. Kama and J.M-S. Lubuma, On elementary stable and dissipative finite difference schemes for dynamical systems, In: T.E. Simos, and G. Maroulis (Eds), *Proceedings of the International Conference of Computational Methods in Science and Engineering* (Crete, Greece, 27 October-1 November 2006), Lecture Series on Computer and Computational Sciences, Vol 7A, VSP International Science Publishers, Utrecht, 2006, pp. 24-27.

[6] R. Anguelov, Y. Dumont, and J. Lubuma, On Nonstandard Finite Difference Schemes in Biosciences, AMITANS 2012, june 2012, *AIP Conf. Proc.* **1487** (2012), 212–223.

[7] R. Anguelov and J.M-S. Lubuma, Contributions to the mathematics of the nonstandard finite difference method and applications, *Numerical Methods for Partial Differential Equations* **17** (2001), 518–543.

[8] R. Anguelov, Y. Dumont, J.M-S. Lubuma and M. Shillor, Dynamically consistent non-standard finite difference schemes for the MSEIR epidemiological model. In: T. Simos, G. Psihoyios, Ch. Tsitouras (eds), *Proceedings of the International Conference of Numerical Analysis and Applied Mathematics*, Crete, Greece, 18-22 September 2009, American Institute of Physics Conference Proceedings-AIP **1168**, Volume 2, 2009, pp. 1213-1216.

[9] R. Anguelov, Y. Dumont, J.M-S. Lubuma and M. Shillor, Comparison of some standard and nonstandard numerical methods for the MSEIR epidemiological model. In: T. Simos, G. Psihoyios, Ch. Tsitouras (eds), *Proceedings of the International Conference of Numerical Analysis and Applied Mathematics*, Crete, Greece, 18-22 September 2009, American Institute of Physics Conference Proceedings-AIP **1168** , Volume 2, 2009, pp. 1209-1212.

[10] R Anguelov, P Kama and J Lubuma. On non-standard finite difference models of reaction-diffusion equations, *Journal of Computational and Applied Mathematics*, **175** (2005), 11-29.

[11] R. Anguelov, J. Lubuma, J.M.-S. and K. Mahudu, Qualitatively stable finite difference schemes for advection reaction equations, *Journal of Computational and Applied Mathematics*, **158** (2003), 19-30.

[12] R. Anguelov, J. M.-S. Lubuma and M. Shillor, Dynamically consistent nonstandard finite difference schemes for continuous dynamical systems, In: X Hou, X Lu, A Miranville, J Su and J Zhu (eds), Dynamical Systems and Differential Equations, (Proceedings of the 7th AIMS Conference on Dynamical Systems and Differential Equations, 18 – 21 May 2008, University of Texas at Arlington, Texas, USA), *American Institute of Mathematical Sciences Conference Series*, 2009, pp 34-43.

[13] R. Anguelov, J. M.-S. Lubuma and M. Shillor, Topological dynamic consistency of nonstandard finite difference schemes for dynamical systems, *Journal of Difference Equations and Applications,* **17** (2011), 1769-1791.

[14] R. Anguelov and E. Popova, Reliable Simulations for Applied Dynamical Models, In: T. Simos, G. Psihoyios, Ch. Tsitouras (eds), *Proceedings of the International Conference of Numerical Analysis and Applied Mathematics*, Crete, Greece, 18-22 September 2009, American Institute of Physics Conference Proceedings-AIP **1168**, Volume 2, 2009, pp. 1205-1208.

[15] R. Anguelov and E. Popova 2010. Topological structure preserving numerical simulations of dynamical models, *Journal of Computational and Applied Mathematics*, **235** (2010), 358-365.

[16] R. Balan, An extension of Barbashin-Krasovski-LaSalle theorem to a class of nonautonomous systems, Technical Report, 1995. Available at: http://www.ima.umn.edu/ balan/academic_work.html

[17] A. Berman and R. J. Plemmons, *"Nonnegative Matrices in the Mathematical Sciences,"* Academic Press, New York, 1979.

[18] C. Castillo-Chavez, Z. Feng and W. Huang, On the computation of R_0 and its role on global stability, In: C. Castillo-Chavez, S. Blower, P. van den Driessche, D. Kirschne and A. A. Yakubu (Eds.) , *Mathematical Approaches for Emerging and Reemerging Infectious Diseases: An Introduction,* Springer, New York, 2002, pp. 229-250.

[19] O. Diekman and J. A. P. Heesterbeck, *"Mathematical Epidemiology of Infectious Diseases,"* Mathematical and Computational Biology, Wiley, New York, 2000.

[20] D.T. Dimitrov and H.V. Kojouharov, Positive and elementary stable nonstandard methods with applications to predator-prey systems, *Journal of Computational and Applied Mathematics*, **182** (2006), 98–108.

[21] Y. Dumont, F. Chiroleu and C. Domerg, On a temporal model for the Chikungunya disease: modeling, theory and numerics, *Mathematical Biosciences* **213** (2008), 80-91.

[22] Y. Dumont and F. Chiroleu, Vector control for the Chikungunya disease, *Mathematical Biosciences and Engineering* **7**(2) (2010), 315–348.

[23] Y. Dumont, J. Russel, V. Lecomte and M. Le Corre, Conservation of endangered endemic seabirds within a multi-predator context: the Barau's Petrel in Réunion Island, *Natural Resource Modelling* **23**(3) (2010), 381–436.

[24] Y. Dumont and J.M. Tchuenche, Mathematical studies on the Sterile Insect Technique for the Chikungunya Disease and Aedes albopictus, *Journal of Mathematical Biology*, **65** (5) (2012), 809–854.

[25] Y. Dumont and J.M.-S. Lubuma, Non-standard finite difference methods for vibro-impact problems, *Proceedings of the Royal Society of London, Series A: Mathematical, Physical and Engineering Sciences*, **461A** (2005), 1927–1950.

[26] Y. Dumont, Applications of the non-standard finite difference method in non-smooth mechanics, in R Mickens (Ed), *Advances in the applications of non-standard finite difference schemes*, World Sci. Publ., Hackensack, NJ, 2005, 287–341.

[27] A.B. Gumel (Ed.) Special Issue dedicated to Prof R. E. Mickens on the occasion of his 60th birthday. *Journal of Difference Equations and Applications* **9**, 2003.

[28] A.B. Gumel, C Castillo-Chavez, R Mickens and D. Clemence (Eds), *Mathematical studies on human diseases dynamics: emerging paradigms and challenges,* (Proceedings of the AMS-IMS-SIAM Joint Summer Research Conference on Modelling the Dynamics of Human Diseases: emerging paradigms and challenges, 17-21 July 2005, Utah, USA), Contemporary Mathematics Series no 410, American Mathematical Society Providence, Rhode Island 2006.

[29] H. W. Hethcote, The mathematics of infectious disease, *SIAM Review* **42** (2000), 599–653.

[30] J. A Jacquez and C. P. Simons, Qualitative theory of compartmental systems, *SIAM Review* **35** (1993), 43–79.

[31] R. Anguelov, Y. Dumont, and J. Lubuma, On nonstandard finite difference schemes in Biosciences, AMITANS 2012, June 2012, *AIP Conf. Proc.* **1487** (2012), 212–223.

[32] J. C. Kamgang and G. Sallet, Computation of threshold conditions for epidemiological models and global stability of the disease free equilibrium (DFE), *Math. Biosc.* **213** (2008), 1–12.

[33] A. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems,* Cambridge University Press, New York, 1995.

[34] W. O. Kermack and A. G. McKendrick, Contributions to the mathematical theory of epidemics, part 1, *Proc. Roy. Soc. London Ser. A* **115** (1927), 700-721.

[35] A. Korobeinikov, Lyapunov function and global stability for SIR and SIRS epidemiological models with non-linear transmission, *Bull Math. Bio.* **30** (2006), 615–626.

[36] J. D. Lambert, *Numerical methods for ordinary differential systems*, Wiley, New York, 1991.

[37] J. P. LaSalle, Stability theory for ordinary differential equations, *J. Differential Equations*, **4** (1968), 57-65.

[38] J.Lubuma and K Patidar, Uniformly convergent non-standard finite difference methods for self-adjoint singular perturbation problems, *Journal of Computational and Applied Mathematics*, **191** (2006), 228-238.

[39] J.M-S Lubuma, E. Mureithi and Y.A. Terefe, Analysis and dynamically consistent numerical schemes for the SIS model and related reaction diffusion equation, In C.I. Christosov and M.D. Todorov (eds): Applications of Mathematics in Technical and Natural Sciences (Proceedings of the 3rd International Conference AMiTaNS'11, Albena, Bulgaria, 20-25 June 2011), *American Institute of Physics (AIP) Conference Proceedings* **1404** (2011), pp. 168-175.

[40] J.M-S. Lubuma and A. Roux, An improved theta method for systems of ordinary differential equations, *Journal of Difference Equations and Applications* **9**, 1023–1035 (2003).

[41] A.S. De Markus and R.E. Mickens, Suppression of numerically induced chaos with nonstandard finite difference schemes, *Journal of Computational and Applied Mathematics*, **106** (1999), 317–324S.

[42] E. Matthias and R.E. Mickens, Solutions to the discrete Airy equation: application to parabolic equation calculations, *Journal of Computational and Applied Mathematics*, **172** (2004), 183–206

[43] T. Meis and U. Marcowitz, *"Numerical Solution of Partial Differential Equations,"* Springer, Berlin, 1981.

[44] R. E. Mickens, *Nonstandard finite difference models of differential equations,* World Scientific, Singapore, 1994.

[45] R.E. Mickens, Discretizations of nonlinear differential equations using explicit nonstandard methods, *Journal of Computational and Applied Mathematics*, **110** (1999), 167-171.

[46] R. E. Mickens (Ed.), *Applications of Nonstandard Finite Difference Schemes,* World Scientific, Singapore, 2000.

[47] R. E. Mickens (Ed.), *Advances in the applications of nonstandard finite difference schemes,* World Scientific, Singapore, 2005

[48] R. E. Mickens, Nonstandard finite difference methods, In: R. E. Mickens (Ed), *Advances in the applications of nonstandard finite difference schemes*, World Scientific, Singapore, 2005, pp.1-9.

[49] R. E. Mickens, Discrete models of differential equations: the roles of dynamic consistency and positivity, In: L. J. S. Allen, B. Aulbach, S. Elaydi and R. Sacker (Eds), *Difference equations and discrete dynamical systems* (Proceedings of the 9th International Conference, Los Angeles, USA, 2-7 August 2004), World Scientific, Singapore, 2005, pp. 51-70.

[50] R. E. Mickens, Calculation of denominator functions for nonstandard finite difference schemes for differential equations satisfying a positivity condition, *Numerical Methods for Partial Differential Equations*, **23** (2007), 672-691.

[51] Moghadas and A Gumel, A mathematical study of a model for childhood diseases with non-permanent immunity, *Journal of Computational and Applied Mathematics*, **157** (2003), 347–363.

[52] J. D. Murray, *Mathematical biology*, Springer-Verlag, 1989.

[53] W. Piyawong, E. H. Twizell and A. B. Gumel, An unconditionally convergent finite-difference scheme for the SIR model, *Applied Math. Comp.* **146** (2003), 611–625.

[54] L. F. Shampine and M. W. Reichelt, The Matlab ODE Suite. SIAM J. Sci. Comput. **18** (1997), 1–22

[55] A. M. Stuart and A. R. Humphries, *Dynamical systems and numerical analysis*, Cambridge University Press, New York, 1998.

[56] H. R. Thieme, *Mathematics of population biology*, Princeton University Press, Princeton and Oxford, 2003.

[57] P. Van den Driessche and J. Watmough, Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission, *Mathematical Biosciences* **180** (2002), 29–48.

[58] R.S. Varga, *Matrix iterative analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

A   Nonnegative matrices

Nonnegative matrices play an important role in many areas of applied mathematics, including Numerical Analysis. They are discussed in some of the well known books on Matrix Computations such as [17] or [58]. The formulations and proof of some of the main theorems in this paper rely on the properties of nonnegative matrices. In order to avoid frequent interruptions of the exposition with statements about nonnegative matrices and also to make the paper as self-contained as possible, we include in this Appendix some basic results. Proofs are omitted even for the theorems for which we could not find references in the literature.

The order in the set of $n \times n$ matrices is the point-wise one, namely,

$$A \leq B \Longleftrightarrow a_{ij} \leq b_{ij}, \quad i, j = 1, ..., n.$$

Hence, a matrix A is nonnegative if $A \geq 0$. Let us note that multiplication by a nonnegative matrix preserves the order. More precisely,

$$(A \leq B, \ C \geq 0) \Longrightarrow (AC \leq BC, \ CA \leq CB). \tag{70}$$

We also use the following notation

$$A < B \iff A \leq B, \ A \neq B$$
$$A << B \iff a_{ij} < b_{ij}, \quad i,j = 1, ..., n$$

As usual, for a given matrix A, $\rho(A)$ is the spectral radius and $\alpha(A) = \max\{Re(\lambda) : \lambda \text{ eigenvalue of } A\}$.

The following theorem, known as the Perron-Frobenius Theorem, gives one of the important properties of nonnegative matrices.

Theorem 17 *([17, Theorem 2.1.3, p.27], [58]) Let A be a real $n \times n$ matrix with nonnegative entries. Then $\rho(A)$ is an eigenvalue and there exists a nonnegative eigenvector v associated with $\rho(A)$. Furthermore, if A is irreducible, then $\rho(A)$ is a simple eigenvalue and A has a positive eigenvector associated with $\rho(A)$.*

Theorem 18 *[17, Corollary 2.1.5, p.27] Let A and B be real nonnegative $n \times n$ matrices such that $A < B$ and $A + B$ is irreducible. Then $\rho(A) < \rho(B)$.*

Related to nonnegative matrices are the three kinds given in the next definitions and which are frequently used in the body of the paper.

Definition 19 *([58]) A matrix A is called an M -matrix if it can be written in the form $A = \mu I - B$, where $B \geq 0$ and $\mu > \rho(B)$. This is equivalent to saying that A, with non-positive off-diagonal entries, is nonsingular, and $A^{-1} \geq 0$.*

Definition 20 *[30] A matrix A is called a Metzler matrix if all its off-diagonal entries are nonnegative.*

Definition 21 *[30] Let $A \in \mathbb{R}^{n \times n}$. A is called a compartmental matrix if A is a Metzler matrix and $\sum_{i=1}^{n} A_{ij} \leq 0$, $j = 1, 2, ..., n$.*

Theorem 22 *Let A and B be two Metzler matrices such that $A < B$ and $A + B$ is irreducible. Then $\alpha(A) < \alpha(B)$.*

Theorem 23 *[43, Theorem 13.9 & Condition 13.10] Let $A = D - B$, where $D = diag(A) \geq 0$ and $B \geq 0$. If A is strictly diagonally dominant then A is an M-matrix.*

The next theorem gives an essential connection between Metzler matrices and M-matrices.

Theorem 24 *Let A be a Metzler matrix and $\alpha(A) \leq 0$.*

a) For each $\mu > 0$ the matrix $I - \mu A$ is a nonsingular M-matrix and

$$\rho\left((I - \mu A)^{-1}\right) = \frac{1}{1 - \mu\alpha(A)} \leq 1.$$

b) If $\mu > 0$ is such that $I + \mu A \geq 0$ then $\rho(I + \mu A) = 1 + \mu\alpha(A) \leq 1$.

Furthermore, if $\alpha(A) < 0$ then the inequalities in both a) and b) are strict.

Theorem 25 *[17, Theorem 6.3.11] If A is an M-matrix that is irreducible, then $A^{-1} \gg 0$.*

B   Proof of Theorem 2

Proof. The global asymptotic stability of x^* on Ω entails two properties:
(a) The fixed-point x^* attracts all trajectories with Ω being the basin of attraction;
(b) The fixed-point x^* is stable.
We will prove the attractiveness first. Let (x^k) be any trajectory of (5). Since

$$V(x^{k+1}) = V(f(x^k)) \leq V(x^k),$$

it follows from assumptions (i) and (ii) that the sequence $(V(x^k))$ is decreasing and is bounded from below, so it converges. Let $m = \lim_{k \to \infty} V(x^k)$. Due to the dissipativity property of the dynamical system, the sequence (x^k) is bounded and the set A of its accumulation points is a compact subset of the absorbing set K. Let $a \in A$. Then, there exists a subsequence (x^{n_k}) of (x^k) such that $\lim_{k \to \infty} x^{n_k} = a$. Using the continuity of V and f we have $V(a) = \lim_{k \to \infty} V(x^{n_k}) = m$ and $V(f(a)) = \lim_{k \to \infty} V(f(x^{n_k})) = \lim_{k \to \infty} V(x^{n_k+1}) = m$. Therefore $V(a) = V(f(a))$ which implies that $a \in \mathcal{E}$. Since a is an arbitrary accumulation point, this also implies that $A \subset \mathcal{E}$. Next, for the sake of convenience, we show that the set A is invariant. Indeed consider the sequence $(f(x^k))$. The accumulation points of this sequence are in $f(A)$. However, this sequence is actually the original sequence (x^k) without the first term, so the set of its accumulation points is A. Hence,

$$A = f(A). \tag{71}$$

Since by assumption, $\mathcal{L}$ is the largest (in terms of inclusion) invariant set of (5), which is a subset of $\mathcal{E}$, we have $A \subseteq \mathcal{L}$. Using the fact that x^* is GAS on $\mathcal{L}$, we also have $x^* \in A$.

Equation (71) shows that f defines a complete generalized dynamical system on A, that is, any trajectory $(a^k)_{k \in \mathbb{N}}$ on A can be continued infinitely in the negative direction as well, so that we have

$$(a^k)_{k \in \mathbb{Z}} = (..., a^{-2}, a^{-1}, a^0, a^1, a^2, a^3, ...) \,,$$

where $a^{-k-1} \in f^{-1}(a^{-k}) \cap A$. The fact that $f^{-1}(a^{-k}) \cap A \neq \emptyset$ is guaranteed by (71). Note that there is no uniqueness in the negative direction.

Let $a \in A$. Assume that $a \neq x^*$. Let $\varepsilon = \frac{1}{2}||a - x^*||$, where $|| \cdot ||$ is the Euclidean norm on $\mathbb{R}^n$. Since x^* is stable on A, there exists $\delta > 0$ such that the positive trajectory initiated at any point in $B_\delta(x^*) \cap A$ is contained in $B_\varepsilon(x^*) \cap A$. Here $B_\delta(x^*)$ and $B_\varepsilon(x^*)$ denote the open balls in Ω with center x^* and radii δ and ε, respectively, e.g. $B_\varepsilon(x^*) = \{x \in \Omega : ||x - x^*|| < \varepsilon\}$. Consider a negative trajectory $(a^{-k})_{k \in \mathbb{N}}$ initiated at a, that is, $a^0 = a$ and $f(a^{-k}) = a^{-k+1}$. Denote by B the set of accumulation points of the sequence $(a^{-k})_{k \in \mathbb{N}}$. Since the set of the accumulation points of $(f(a^{-k}))_{k \in \mathbb{N}}$ (this is the sequence $(a^{-k})_{k \in \mathbb{N}}$ with one more term in front) is also B we have that $B = f(B)$. However, any invariant subset of $\mathcal{L}$ must contain x^* by virtue of its global asymptotic stability. Hence x^* is an accumulation point of $(a^{-k})_{k \in \mathbb{N}}$. Therefore, there exists k_0 such that $a^{-k_0} \in B_\delta(x^*)$. Consider the sequence

$$a^{-k_0}, a^{-k_0+1}, a^{-k_0+2}, ..., a^{-1}, a^0, a^1, ...$$

Since its initial point a^{-k_0} is in $B_\delta(x^*)$, it follows that the whole sequence is contained in $B_\varepsilon(x^*)$. This is a contradiction since $a^0 = a \notin B_\varepsilon(x^*)$. Hence the assumption $a \neq x^*$ is false. Therefore x^* is the only point in A, which implies that the sequence x^k converges to x^*. This completes the proof that x^* is attractive on Ω.

It remains to show that x^* is stable. Since x^* is attractive, it is easy to see that

$$\min_{x \in \Omega} V(x) = V(x^*) = m, \qquad \mathcal{L} = \{x \in \Omega : V(x) = m\}. \tag{72}$$

Assume that x^* is not stable. Then there exists $\hat{\varepsilon}$ such that for every $\delta > 0$ there exists $x^0 \in B_\delta(x^*)$ such that the trajectory initiated at x^0 is not in $B_{\hat{\varepsilon}}(x^*)$. Since x^* is stable on $\mathcal{L}$, there exists $\hat{\delta} < \hat{\varepsilon}$ such that any trajectory initiated in $B_{\hat{\delta}}(x^*) \cap \mathcal{L}$ is contained in $B_{\hat{\varepsilon}}(x^*) \cap \mathcal{L}$. Choose a sequence (δ_i) such that $\lim_{i \to \infty} \delta_i = 0$ and $\delta_i < \tilde{\delta} = \frac{1}{2}\hat{\delta}$. Then for every $i \in \mathbb{N}$ there exist $x_i^0 \in B_{\delta_i}(x^*)$ and $p_i, q_i \in \mathbb{N}$ such that

$$\begin{aligned} &x_i^{p_i} \in B_{\tilde{\delta}}(x^*), \\ &x_i^{p_i+s} \in B_{\hat{\varepsilon}}(x^*) \setminus B_{\tilde{\delta}}(x^*), \ 1 \leq s < q_i, \\ &x_i^{p_i+q_i} \notin B_{\hat{\varepsilon}}(x^*). \end{aligned}$$

The sequence $(x_i^{p_i})$ is contained is the compact set $\overline{B_{\tilde{\delta}}(x^*)}$. Therefore, it has a convergent subsequence. To avoid re-indexing we assume that $(x_i^{p_i})$ is convergent. Let $\lim_{i \to \infty} x_i^{p_i} = w^0$. We have that $V(w^0) = \lim_{i \to \infty} V(x_i^{p_i}) \leq \lim_{i \to \infty} V(x_i^0) = m$. Hence

$$w^0 \in \overline{B_{\tilde{\delta}}(x^*)} \cap \mathcal{L} \subseteq B_{\hat{\delta}}(x^*) \cap \mathcal{L}.$$

Therefore, the trajectory (w^k) initiated at w^0 is contained in $B_{\hat{\varepsilon}}(x^*)$. Let $k \in \mathbb{N}$ be arbitrary and consider the set $\{w^1, ..., w^k\} \subseteq B_{\hat{\varepsilon}}(x^*)$. Since $w^s = \lim_{i \to \infty} x_i^{p_i+s}$, $0 \leq s \leq k$, there exists i_0 such that $q_i > k$ for $i > i_0$. Hence $w^k \notin B_{\tilde{\delta}}(x^*)$ and so, (w^k) does not converge to x^*. This contradicts the global asymptotic stability of x^* on $\mathcal{L}$. Thus x^* is stable. ∎