IMPLICIT-EXPLICIT METHODS FOR A CLASS OF NONLINEAR NONLOCAL GRADIENT FLOW EQUATIONS MODELLING COLLECTIVE BEHAVIOUR

RAIMUND BÜRGER[A,*], DANIEL INZUNZA[A], PEP MULET[B], AND LUIS M. VILLADA[C]

Abstract. Nonlinear convection-diffusion equations with nonlocal flux and possibly degenerate diffusion describe interacting gases, flow in porous media, collective behaviour in biology, and other phenomena. Their numerical solution by an explicit finite difference method is costly since one needs to discretize a local spatial convolution for each evaluation of the convective numerical flux, and moreover the diffusion term gives rise to a disadvantageous Courant-Friedrichs-Lewy (CFL) condition. More efficient numerical methods are obtained by applying second-order implicit-explicit (IMEX) Runge-Kutta time discretizations to an available explicit scheme for such models [J.A. Carrillo, A. Chertock, Y. Huang, Commun. Comput. Phys. 17 (2015) 233–258]. The resulting IMEX-RK methods avoid the restrictive time step limitation of explicit schemes since the diffusion term is handled implicitly, but one needs to solve nonlinear algebraic systems in every time step. It is proven, for a general number of space dimensions, that this method is well defined. Numerical experiments for spatially two-dimensional problems motivated by models of collective behaviour are conducted with several alternative choices of the pair of Runge-Kutta schemes defining an IMEX-RK method. Results illustrate that for fine discretizations IMEX-RK methods are more efficient in terms of reduction of error versus CPU time than the original explicit method.

1. Introduction

1.1. **Scope.** This contribution is concerned with the efficient numerical solution of the following initial value problem for a nonlinear nonlocal partial differential equation (PDE) with a gradient flow structure in d space dimensions:

$$u_t = \nabla \cdot \big(u\nabla(H'(u) + V(\boldsymbol{x}) + W * u)\big), \quad \boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d, \quad t > 0, \tag{1.1}$$

$$u(\boldsymbol{x}, 0) = u_0(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}^d. \tag{1.2}$$

Here the sought solution $u = u(\boldsymbol{x}, t)$ is an unknown probability distribution function or population density, $H(u)$ is a density of internal energy, $V(\boldsymbol{x})$ is a confinement potential, $W(\boldsymbol{x})$ is an interaction potential, which is assumed to be symmetric and we recall that

$$\big(W * u(\cdot, t)\big)(\boldsymbol{x}) = \int_{\mathbb{R}^d} W(\boldsymbol{y})u(\boldsymbol{x} - \boldsymbol{y}, t)\,\mathrm{d}\boldsymbol{y} = \int_{\mathbb{R}^d} W(\boldsymbol{x} - \boldsymbol{y})u(\boldsymbol{y}, t)\,\mathrm{d}\boldsymbol{y}.$$

Date: March 4, 2019.

2010 *Mathematics Subject Classification.* 35K15, 35K55, 35K65, 65M06.

Key words and phrases. Nonlocal partial differential equation, gradient flow, implicit-explicit numerical method, collective behaviour.

*Corresponding author.

[A]CI[2]MA and Departamento de Ingeniería Matemática, Facultad de Ciencias Físicas y Matemáticas, Universidad de Concepción, Casilla 160-C, Concepción, Chile. E-Mail: rburger@ing-mat.udec.cl, dinzunza@ing-mat.udec.cl.

[B]Departament de Matemàtiques, Universitat de València, Av. Vicent Andrés Estellés, E-46100 Burjassot, Spain. E-Mail: mulet@uv.es.

[C]CI[2]MA, Universidad de Concepción, and Departamento de Matemàtica, Facultad de Ciencias, Universidad del Bío-Bío, Casilla 5-C, Concepción, Chile. E-Mail: lvillada@ubiobio.cl.

Clearly, if $W = 0$ and $H(u) = u \log u - u$ or $H(u) = u^m$, the classical heat equation and porous medium/fast diffusion equation are recovered, respectively [30]. The function W may be as singular as the Newtonian potential in the chemotaxis system [19] or as smooth as $W(\boldsymbol{x}) = |\boldsymbol{x}|^\alpha$ with $\alpha > 2$ in granular flow [3]. For the further discussion it is useful to recall that (1.1) can be written as a nonlinear, nonlocal convection-diffusion equation

$$u_t + \nabla \cdot \big(u \boldsymbol{v}[u]\big) = \Delta \Phi(u),$$

where

$$\Phi(u) = \int_0^u s H''(s)\, \mathrm{d}s, \tag{1.3}$$

so that $\Phi'(u) = uH''(u)$ and $\boldsymbol{v}[u](\boldsymbol{x}) = -\nabla(W * u + V)(\boldsymbol{x})$. Here the notation $\boldsymbol{v}[u] = \boldsymbol{v}[u(\cdot, t)]$ means that the velocity $\boldsymbol{v}$ depends on $u(\cdot, t)$ as a function of $\boldsymbol{x}$ as a whole.

Although there is no closed existence, uniqueness and well-posedness theory for nonlocal convection-diffusion equations such as (1.1), it is plausible to perform simulations with appropriate numerical methods. Explicit schemes for hyperbolic first-order conservation laws and related equations can be rather slow for some steady-state computations due to CFL stability restrictions on the time step size, but their use for unsteady computations is deemed practical in many situations. If diffusion terms are present, then these may be treated implicitly to overcome the drastic step size stability restrictions imposed by their alternative explicit treatment. It is the purpose of the present work to demonstrate the benefits of applying implicit-explicit (IMEX) schemes for the efficient solution of (1.1), (1.2) in several space dimensions, under specific assumptions on the diffusive term. The proposed schemes, based on IMEX Runge-Kutta (IMEX-RK) time discretizations, turn out to be more efficient, in terms of error reduction versus CPU time, than the explicit scheme of [12].

1.2. **Related work.** Equation (1.1) arises in many contexts including interacting gases [20], granular flows [28], flow in porous media [13, 23], and collective behavior in biology [26, 27] (see also [12, 21, 25] for further references). In one space dimension, (1.1) can also be understood as a model of the aggregation of populations [4], as we elaborate in [9]. Of particular interest is the relation with a self-propelled swarming model of Cucker-Smale type [2] with diffusion, for which the corresponding space-homogeneous version can be formulated in terms of a nonlinear Fokker-Planck equation that can be expressed as (1.1) under suitable definitions of W, V and Φ, and where the solution u is the density distribution of individuals (say, birds or fish) that have velocity $\boldsymbol{x} \in \mathbb{R}^d$ at time $t > 0$. We refer to [2, 26] and the references cited in these papers for details.

Concerning numerical methods for (1.1), we mention that finite element approximations have been proposed in the literature which are positivity preserving and entropy decreasing at the expense of constructing them by an implicit discretization in time but continuous in space [11]. Numerical methods for equation (1.1) and its variants are mostly explicit schemes for convection-diffusion equations. In fact, Carrillo et al. [12] propose both one- and two-dimensional finite volume schemes for (1.1) and prove their positivity preserving and entropy dissipation properties along with error estimates and convergence results. These schemes follow a method of lines and are explicit by the choice of explicit SSP Runge-Kutta ODE integrators.

Assume for simplicity that (1.1) is discretized in time with a time step Δt and in space by a Cartesian grid with meshwidth Δx in each direction. Then the detailed stability restriction for its explicit discretization [12] implies that $\Delta t \propto \Delta x^2$. In fact, this restriction stems from the explicit treatment of the diffusive term, so this motivates the main goal of our work, which is to propose using an IMEX-RK method that treats the diffusive term implicitly and the convective

term explicitly. To explain the main idea, we assume that

$$\frac{\mathrm{d}\boldsymbol{u}(t)}{\mathrm{d}t} = \boldsymbol{\mathcal{C}}\big(\boldsymbol{u}(t)\big) + \boldsymbol{\mathcal{D}}\big(\boldsymbol{u}(t)\big), \tag{1.4}$$

represents a method-of-lines semi-discretization of (1.1), where $\boldsymbol{u} = \boldsymbol{u}(t)$ is a spatial discretization of the solution and $\boldsymbol{\mathcal{C}}(\boldsymbol{u})$ and $\boldsymbol{\mathcal{D}}(\boldsymbol{u})$ are discretizations of the convective and diffusive terms, respectively. Assume, for simplicity, that the spatial mesh width is $\Delta x > 0$. Then the stability restriction on the time step Δt that explicit schemes impose when applied to (1.4) is very severe (Δt must be proportional to Δx^2), due the presence of $\boldsymbol{\mathcal{D}}(\boldsymbol{u})$. Implicit treatment of both $\boldsymbol{\mathcal{C}}(\boldsymbol{u})$ and $\boldsymbol{\mathcal{D}}(\boldsymbol{u})$ would remove any stability restriction on Δt. However, the upwind nonlinear discretization of the convective terms contained in $\boldsymbol{\mathcal{C}}(\boldsymbol{u})$ that is needed for stability, makes its implicit treatment extremely involved. In fact, after the pioneering work of Crouzeix [14], numerical integrators that deal implicitly with $\boldsymbol{\mathcal{D}}(\boldsymbol{u})$ and explicitly with $\boldsymbol{\mathcal{C}}(\boldsymbol{u})$ can be used with a time step restriction dictated by the convective term alone. References to applications of IMEX methods to convection-diffusion problems, convection problems with stiff reaction terms, hyperbolic systems with relaxation, and the solution of semidiscretized PDEs include [1, 7, 8, 10, 16, 24].

1.3. **Outline of the paper.** The remainder of the paper is organized as follows. The numerical method is introduced in Section 2, starting in Section 2.1 with a statement of assumptions on properties of the function Φ and a description of the index notation, which is partly inspired by the usage in [18], that allows us to write the scheme in a general number d of space dimensions. Then, in Section 2.2 we outline the spatial discretization of (1.1), for which we prove that under a suitable CFL condition, a fully discrete scheme based on first-order explicit Euler time stepping generates a positivity-preserving scheme (Theorem 2.1 in Section 2.2). Next, in Section 2.3, the IMEX-RK time discretization is introduced. That section contains our main theoretical results, namely Theorem 2.2 (and its proof), which states that under a suitable CFL condition the IMEX-RK schemes are well defined (the algebraic systems arising within each time step are uniquely solvable with non-negative solutions), along with Corollary 2.1 that ensures that the Euler IMEX-RK scheme is positivity preserving. Finally, in Section 2.4 the damped linear solver required to handle the algebraic problems within each iteration of the IMEX-RK scheme is described. In Section 3 numerical results are presented. To this end, we first specify (in Section 3.1) the particular IMEX-RK schemes to be used in the numerical experiments along with their corresponding CFL conditions limiting the time step, and then describe (in Section 3.2) the computation of the approximate numerical error (defined in each case by a suitable reference solution). Then, in Section 3.3 five numerical examples are presented, where we limit ourselves to the case of $d = 2$ space dimensions. These include numerical experiments proposed by Carrillo et al. [12] (Examples 1 and 3), Topaz et al. (Example 2), and Pareschi and Zanella [26] (Example 5). (These examples are all motivated by models of collective behaviour, which motivates the title of the paper.) Some conclusions are collected in Section 4.

2. Numerical method

2.1. **Some assumptions and notation.** We assume that $H''(u) \geq 0$ for all $u \in (0,\infty)$, $H'' \in C^1(\mathbb{R} \setminus \{0\})$, so that

$$\begin{aligned} &\Phi \in C^1([0,\infty)) \cap C^2((0,\infty)), \\ &\Phi(0) = \Phi'(0) = 0, \\ &\Phi(u) \geq 0, \Phi'(u) \geq 0 \text{ for } u \in [0,\infty). \end{aligned} \tag{2.1}$$

We limit the treatment to the spatial domain given by the d-dimensional open interval

$$\Omega := (-L_1, L_1) \times \cdots \times (-L_d, L_d) \tag{2.2}$$

and denote by $u : \Omega \times (0, \infty) \to [0, \infty)$ the solution of (1.1). Each coordinate interval $(-L_l, L_l)$, $l = 1, \ldots, d$, is subdivided into M_l subintervals of size $\Delta x_l = 2L_l/M_l$. This creates a number $M_* := M_1 M_2 \cdots M_d$ of finite volumes or cells $C_{\boldsymbol{i}}$, which we indicate by $\boldsymbol{i} = (i_1, \ldots, i_d) \in \mathcal{M}$, where we define the index set

$$\mathcal{M} := \{1, \ldots, M_1\} \times \cdots \times \{1, \ldots, M_d\} \subset \mathbb{N}^d. \tag{2.3}$$

The center of $C_{\boldsymbol{i}}$ is denoted by $\boldsymbol{x}_{\boldsymbol{i}}$, so the coordinates of $\boldsymbol{x}_{\boldsymbol{i}}$ are

$$\boldsymbol{x}_{\boldsymbol{i}} = (x_{1,i_1}, \ldots, x_{d,i_d}) = \big((i_1 - o_1)\Delta x_1, \ldots, (i_d - o_d)\Delta x_d\big), \tag{2.4}$$

where $o_l = (M_l + 1)/2$ for $l = 1, \ldots, d$, such that $x_{l,1} = -L_l + \Delta x_l/2$ and $x_{L,M_l} = L_l - \Delta x_l/2$, $l = 1, \ldots, d$, and we utilize d-dimensional unit vectors $\boldsymbol{e}_1 = (1, 0, \ldots, 0)$ to $\boldsymbol{e}_d = (0, \ldots, 0, 1)$ to address neighboring grid points, for instance $\boldsymbol{x}_{\boldsymbol{i}+\boldsymbol{e}_1} = \boldsymbol{x}_{i_1+1,i_2,\ldots,i_d}$. A similar notation is used to address flux values associated with cell interfaces. Furthermore, we assume that the velocity vector $\boldsymbol{v}$ is given by components $\boldsymbol{v} = (v^1, \ldots, v^d)^{\mathrm{T}}$.

2.2. **Spatial semi-discretization.** To define the spatial (semi-)discretization, assume first that at an instant t the solution is given through the cell averages $u_{\boldsymbol{i}} = u_{\boldsymbol{i}}(t)$ for $\boldsymbol{i} \in \mathcal{M}$. As in the one-dimensional treatment [9], we use MUSCL reconstructions [29], which amounts for each cell $C_{\boldsymbol{i}}$ to calculating the following reconstructed values at the boundaries of $C_{\boldsymbol{i}}$:

$$u_{\boldsymbol{i}}^{l,\pm} = u_{\boldsymbol{i}} \pm \frac{\Delta x_l}{2}\sigma_{\boldsymbol{i}}^{(l)}, \quad \boldsymbol{i} \in \mathcal{M}, \quad l = 1, \ldots, d, \tag{2.5}$$

where the slope $\sigma_{\boldsymbol{i}}^{(l)}$ for the extrapolation of $u_{\boldsymbol{i}}$ in coordinate direction l is defined by using a so-called slope limiter that guarantees that the reconstructed point values are nonnegative as long as the cell averages $u_{\boldsymbol{i}}$ are nonnegative. Specifically, we utilize the slope limiter defined by

$$\sigma_{\boldsymbol{i}}^{(l)} = \begin{cases} \dfrac{1}{2\Delta x_l}\left(u_{\boldsymbol{i}+\boldsymbol{e}_l} - u_{\boldsymbol{i}-\boldsymbol{e}_l}\right) & \text{if } u_{\boldsymbol{i}} \geq |u_{\boldsymbol{i}+\boldsymbol{e}_l} - u_{\boldsymbol{i}-\boldsymbol{e}_l}|/4, \\ \theta \operatorname{minmod}\left\{\dfrac{1}{\Delta x_l}\left(u_{\boldsymbol{i}+\boldsymbol{e}_l} - u_{\boldsymbol{i}}\right), \dfrac{1}{\Delta x_l}\left(u_{\boldsymbol{i}} - u_{\boldsymbol{i}-\boldsymbol{e}_l}\right)\right\} & \text{otherwise,} \end{cases}$$

where the standard minmod function is given by

$$\operatorname{minmod}\{z_1, z_2\} := \begin{cases} \operatorname{sgn}(z_1)\min\{|z_1|, |z_2|\} & \text{if} \quad \operatorname{sgn}(z_1) = \operatorname{sgn}(z_2), \\ 0 & \text{otherwise.} \end{cases}$$

The parameter $\theta \in (0, 2]$ is used to control the numerical viscosity of the scheme. The value $\theta = 2$ is used in [9, 12], and we adopt it here in all numerical examples.

To approximate $\boldsymbol{v}[u](x_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l})$, we use the formula

$$\left.\frac{\partial z[u]}{\partial x_l}\right|_{\boldsymbol{x}=\boldsymbol{x}_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l}} \approx \frac{1}{\Delta x_l}\big(z[u](\boldsymbol{x}_{\boldsymbol{i}+e_l}) - z[u](\boldsymbol{x}_{\boldsymbol{i}})\big). \tag{2.6}$$

If we assume that $u(t)$ is compactly supported in $(-L, L)^d$, then the discrete approximations of the convolutions $z[u](\boldsymbol{x}_{\boldsymbol{i}}) := (W * u + V)(\boldsymbol{x}_{\boldsymbol{i}})$ (shifted by the function V) are given by

$$(W * u + V)(\boldsymbol{x}_{\boldsymbol{i}}) \approx \tilde{z}[u]_{\boldsymbol{i}} := \prod_{l=1}^{d} \Delta x_l \sum_{-\rho \leq p_1, \ldots, p_d \leq \rho} W_{\boldsymbol{p}}\, u^*_{\boldsymbol{i}-\boldsymbol{p}} + V_{\boldsymbol{i}}, \tag{2.7}$$

where $V_{\boldsymbol{i}} = V(\boldsymbol{x}_{\boldsymbol{i}})$ for $\boldsymbol{p} = (p_1, \dots, p_d) \in \mathbb{Z}^d$, we define $W_{\boldsymbol{p}} := W(p_1\Delta x_1, \dots, p_d\Delta x_d)$, and

$$u^*_{\boldsymbol{i}-\boldsymbol{p}} := \begin{cases} u_{\boldsymbol{i}-\boldsymbol{p}} & \text{if } \boldsymbol{i}-\boldsymbol{p} \in \mathcal{M}, \\ 0 & \text{otherwise,} \end{cases}$$

where the radius of the stencil $\rho \in \mathbb{N}_0$ is computed to retain second-order accuracy. To select ρ, we proceed as in [9] by choosing ρ as the smallest integer such that

$$1 - \frac{\sum_{n_1=-\rho}^{\rho} \cdots \sum_{n_d=-\rho}^{\rho} W(n_1\Delta x_1, \dots, n_d\Delta x_d)}{\sum_{n_1=-\infty}^{\infty} \cdots \sum_{n_d=-\infty}^{\infty} W(n_1\Delta x_1, \dots, n_d\Delta x_d)} \le \xi \Delta x^2, \quad \Delta x = \min\{\Delta x_1, \dots, \Delta x_d\},$$

where we have taken $\xi = 10^{-8}$ in all our numerical examples and the term

$$\sum_{n_1=-\infty}^{\infty} \cdots \sum_{n_d=-\infty}^{\infty} W(n_1\Delta x_1, \dots, n_d\Delta x_d)$$

is approximated by

$$\sum_{n_1=-N_1}^{N_1} \cdots \sum_{n_d=-N_d}^{N_d} W(n_1\Delta x_1, \dots, n_d\Delta x_d)$$

for very large $N_1, \dots, N_d$. Here we used that W is a symmetric function. Clearly, the discrete convolution in (2.7) causes a computational bottleneck. This is a classical problem in scientific computing that is effectively evaluated using fast convolution algorithms, mainly based on Fast Fourier Transforms [31].

Moreover, we apply upwinding based on the sign of the l-th component $\hat{v}^l_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l}$ of the vector

$$\boldsymbol{v}_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l} = \frac{1}{\Delta x}\big(\boldsymbol{v}[u]_{\boldsymbol{i}+\boldsymbol{e}_l} - \boldsymbol{v}[u]_{\boldsymbol{i}}\big) = \big(\hat{v}^1_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l}, \dots, \hat{v}^d_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l}\big)^{\mathrm{T}},$$

where, in agreement with (2.6), the l-th component of $\hat{\boldsymbol{v}}[u]$, the discrete version of $\boldsymbol{v}[u]$, is given by

$$\hat{v}^l_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l} = \frac{1}{\Delta x_l}\big(\tilde{z}[u]_{\boldsymbol{i}+\boldsymbol{e}_l} - \tilde{z}[u]_{\boldsymbol{i}}\big).$$

The upwind procedure now consists in choosing the u-value associated with the cell interface $\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l$ as follows:

$$u_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l} = \begin{cases} u^{l,+}_{\boldsymbol{i}} & \text{if } \hat{v}^l_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l} \ge 0, \\ u^{l,-}_{\boldsymbol{i}+\boldsymbol{e}_l} & \text{if } \hat{v}^l_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l} < 0. \end{cases} \tag{2.8}$$

Solution values are extended by zero outside the domain, i.e., we set $u_{\boldsymbol{i}} := 0$ for $\boldsymbol{i} \in \mathbb{Z}^d \backslash \mathcal{M}$.

Combining all ingredients we may write the semidiscrete scheme in compact form as (1.4), where $\boldsymbol{u} : [0, \infty) \to \mathbb{R}^{M_*}$ and we recall that $\boldsymbol{\mathcal{C}}(\boldsymbol{u})$ and $\boldsymbol{\mathcal{D}}(\boldsymbol{u})$ represent the spatial discretizations of the convective and the diffusive terms, i.e., the respective entries of $\boldsymbol{\mathcal{C}}(\boldsymbol{u}) = (C(\boldsymbol{u})_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathcal{M}}$ and $\boldsymbol{\mathcal{D}}(\boldsymbol{u}) = (D(\boldsymbol{u})_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathcal{M}}$ are given by

$$C(\boldsymbol{u})_{\boldsymbol{i}} = -\sum_{l=1}^{d} \frac{1}{\Delta x_l}\big(u_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l}\hat{v}^l_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l} - u_{\boldsymbol{i}-\frac{1}{2}\boldsymbol{e}_l}\hat{v}^l_{\boldsymbol{i}-\frac{1}{2}\boldsymbol{e}_l}\big), \tag{2.9}$$

$$D(\boldsymbol{u})_{\boldsymbol{i}} = \sum_{l=1}^{d} \frac{1}{\Delta x_l^2}\big(\Phi(u_{\boldsymbol{i}+\boldsymbol{e}_l}) - 2\Phi(u_{\boldsymbol{i}}) + \Phi(u_{\boldsymbol{i}-\boldsymbol{e}_l})\big). \tag{2.10}$$

In closed form the finite volume semidiscretization on cells centered at $\boldsymbol{x_i}$ can be written as the system of ODEs

$$\frac{\mathrm{d}u_{\boldsymbol{i}}}{\mathrm{d}t} = -\sum_{l=1}^{d} \frac{1}{\Delta x_l}\big(u_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l}\hat{v}^l_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l} - u_{\boldsymbol{i}-\frac{1}{2}\boldsymbol{e}_l}\hat{v}^l_{\boldsymbol{i}-\frac{1}{2}\boldsymbol{e}_l}\big) + \sum_{l=1}^{d}\frac{1}{\Delta x_l^2}\big(\Phi(u_{\boldsymbol{i}+\boldsymbol{e}_l}) - 2\Phi(u_{\boldsymbol{i}}) + \Phi(u_{\boldsymbol{i}-\boldsymbol{e}_l})\big).$$

Theorem 2.1. *If $\Phi' \geq 0$ on $(0,\infty)$, $u_{\boldsymbol{i}} \geq 0$ for all $\boldsymbol{i} \in \mathcal{M}$ and the CFL condition*

$$\frac{1}{d}\max_{0\leq u\leq \eta(\boldsymbol{u})}\Phi'(u)\sum_{q=1}^{d}\frac{\Delta t}{\Delta x_q^2} + \max_{1\leq l\leq d}\left\{\frac{\Delta t}{\Delta x_l}\max_{\boldsymbol{k}\in\mathcal{M}}\big|v^l_{\boldsymbol{k}+\frac{1}{2}\boldsymbol{e}_l}\big|\right\} \leq \frac{1}{2d}, \quad \eta(\boldsymbol{u}) := \max_{\boldsymbol{j}\in\mathcal{M}} u_{\boldsymbol{j}} \tag{2.11}$$

is satisfied, then the quantity

$$\mathcal{E}(\boldsymbol{u})_{\boldsymbol{i}} := u_{\boldsymbol{i}} + \Delta t\big(C(\boldsymbol{u})_{\boldsymbol{i}} + D(\boldsymbol{u})_{\boldsymbol{i}}\big) \tag{2.12}$$

satisfies $\mathcal{E}(\boldsymbol{u})_{\boldsymbol{i}} \geq 0$ for all $\boldsymbol{i} \in \mathcal{M}$, i.e., the explicit Euler method applied to the semi-discrete scheme (1.4) *yields a fully discrete positivity preserving scheme.*

Proof. From (2.5) there hold

$$\begin{aligned} &\frac{1}{2d}\sum_{l=1}^{d}\big(u^{l,+}_{\boldsymbol{i}} + u^{l,-}_{\boldsymbol{i}}\big) = u_{\boldsymbol{i}} \quad \text{for } \boldsymbol{i}\in\mathcal{M}, \\ &u^{l,\pm}_{\boldsymbol{i}} \geq 0 \quad \text{for } \boldsymbol{i}\in\mathcal{M},\ l = 1,\dots,d. \end{aligned} \tag{2.13}$$

Moreover for all $\boldsymbol{i} \in \mathcal{M}$ there exist convex combinations

$$\hat{u}_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l} = \theta_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l}u_{\boldsymbol{i}} + \big(1-\theta_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l}\big)u_{\boldsymbol{i}+\boldsymbol{e}_l}, \quad \theta_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l}\in(0,1), \quad l = 1,\dots,d,$$

such that for all $\boldsymbol{i}\in\mathcal{M}$ and $l = 1,\dots,d$,

$$\Phi(u_{\boldsymbol{i}+\boldsymbol{e}_l}) - \Phi(u_{\boldsymbol{i}}) = \Delta x_l^2\beta_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l}(u_{\boldsymbol{i}+\boldsymbol{e}_l} - u_{\boldsymbol{i}}), \quad \beta_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l} = \Phi'(\hat{u}_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l})/\Delta x_l^2. \tag{2.14}$$

Then $D(\boldsymbol{u})_{\boldsymbol{i}}$, given by (2.10), can be written as

$$D(\boldsymbol{u})_{\boldsymbol{i}} = \sum_{l=1}^{d}\big(\beta_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l}u_{\boldsymbol{i}+\boldsymbol{e}_l} + \beta_{\boldsymbol{i}-\frac{1}{2}\boldsymbol{e}_l}u_{\boldsymbol{i}-\boldsymbol{e}_l}\big) + 2d\gamma_{\boldsymbol{i}}u_{\boldsymbol{i}}, \tag{2.15}$$

where we use the notation

$$\gamma_{\boldsymbol{i}} := \frac{1}{2d}\sum_{l=1}^{d}\big(\beta_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l} + \beta_{\boldsymbol{i}-\frac{1}{2}\boldsymbol{e}_l}\big). \tag{2.16}$$

Therefore, we obtain from (2.16) and (2.14) that

$$\max_{\boldsymbol{j}\in\mathcal{M}}\gamma_{\boldsymbol{j}} \leq \frac{1}{d}\max_{0\leq u\leq\eta(\boldsymbol{u})}\Phi'(u)\sum_{q=1}^{d}\frac{1}{\Delta x_q^2}. \tag{2.17}$$

From (2.8) we obtain

$$u_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l}\hat{v}^l_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l} = u^{l,+}_{\boldsymbol{i}}\hat{v}^{l,+}_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l} + u^{l,-}_{\boldsymbol{i}+\boldsymbol{e}_l}\hat{v}^{l,-}_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l}, \quad \hat{v}^{l,\pm}_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l} = \big(\hat{v}^l_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l}\big)^{\pm}.$$

By (2.12), (2.13), (2.9) and (2.15) we may write $\mathcal{E}(\boldsymbol{u})_{\boldsymbol{i}}$ as follows:

$$\mathcal{E}(\boldsymbol{u})_{\boldsymbol{i}} = \frac{1-2d\Delta t\gamma_{\boldsymbol{i}}}{2d}\sum_{l=1}^{d}\big(u^{l,+}_{\boldsymbol{i}} + u^{l,-}_{\boldsymbol{i}}\big)$$

$$+\Delta t\sum_{l=1}^{d}\frac{1}{\Delta x_l}\big(-u_{\boldsymbol{i}}^{l,+}v_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l}^{l,+}-u_{\boldsymbol{i}+\boldsymbol{e}_l}^{l,-}v_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l}^{l,-}+u_{\boldsymbol{i}-\boldsymbol{e}_l}^{l,+}v_{\boldsymbol{i}-\frac{1}{2}\boldsymbol{e}_l}^{l,+}+u_{\boldsymbol{i}}^{l,-}v_{\boldsymbol{i}-\frac{1}{2}\boldsymbol{e}_l}^{l,-}\big)$$
$$+\Delta t\sum_{l=1}^{d}\big(\beta_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l}u_{\boldsymbol{i}+\boldsymbol{e}_l}+\beta_{\boldsymbol{i}-\frac{1}{2}\boldsymbol{e}_l}u_{\boldsymbol{i}-\boldsymbol{e}_l}\big).$$

Taking into account that $u_{\boldsymbol{i}}\geq 0$, $u_{\boldsymbol{i}}^{l,\pm}\geq 0$, $v_{\boldsymbol{i}\pm\frac{1}{2}\boldsymbol{e}_l}^{l,\pm}\geq 0$, and (2.17), we deduce

$$\begin{aligned}
\mathcal{E}(\boldsymbol{u})_{\boldsymbol{i}} &\geq \left(\frac{1}{2d}-\Delta t\gamma_{\boldsymbol{i}}\right)\sum_{l=1}^{d}\big(u_{\boldsymbol{i}}^{l,+}+u_{\boldsymbol{i}}^{l,-}\big)+\Delta t\sum_{l=1}^{d}\frac{1}{\Delta x_l}\big(-u_{\boldsymbol{i}}^{l,+}v_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l}^{l,+}+u_{\boldsymbol{i}}^{l,-}v_{\boldsymbol{i}-\frac{1}{2}\boldsymbol{e}_l}^{l,-}\big)\\
&\geq \sum_{l=1}^{d}\left[\left(\frac{1}{2d}-\Delta t\gamma_{\boldsymbol{i}}-\frac{\Delta t}{\Delta x_l}\big|v_{\boldsymbol{i}-\frac{1}{2}\boldsymbol{e}_l}^{l}\big|\right)u_{\boldsymbol{i}}^{l,-}+\left(\frac{1}{2d}-\Delta t\gamma_{\boldsymbol{i}}-\frac{\Delta t}{\Delta x_l}\big|v_{\boldsymbol{i}+\frac{1}{2}\boldsymbol{e}_l}^{l}\big|\right)u_{\boldsymbol{i}}^{l,+}\right]\\
&\geq \sum_{l=1}^{d}\left(\frac{1}{2d}-\Delta t\max_{\boldsymbol{j}\in\mathcal{M}}\gamma_{\boldsymbol{j}}-\frac{\Delta t}{\Delta x_l}\max_{\boldsymbol{j}\in\mathcal{M}}\big|v_{\boldsymbol{j}-\frac{1}{2}\boldsymbol{e}_l}^{l}\big|\right)\big(u_{\boldsymbol{i}}^{l,-}+u_{\boldsymbol{i}}^{l,+}\big)\\
&\geq \left[\frac{1}{2d}-\frac{1}{d}\max_{0\leq u\leq\eta(\boldsymbol{u})}\Phi'(u)\sum_{q=1}^{d}\frac{\Delta t}{\Delta x_q^2}+\max_{1\leq l\leq d}\left\{\frac{\Delta t}{\Delta x_l}\max_{\boldsymbol{k}\in\mathcal{M}}\big|v_{\boldsymbol{k}+\frac{1}{2}\boldsymbol{e}_l}^{l}\big|\right\}\right]\sum_{l=1}^{d}\big(u_{\boldsymbol{i}}^{l,+}+u_{\boldsymbol{i}}^{l,-}\big)\geq 0.
\end{aligned}$$

This concludes the proof. □

2.3. **Time discretization.** We will use IMEX-RK integration for the system of ODEs (1.4), and where the components of $\boldsymbol{\mathcal{C}}(\boldsymbol{u})$ and $\boldsymbol{\mathcal{D}}(\boldsymbol{u})$ are given by (2.9) and (2.10), respectively. Only the diffusion term $\boldsymbol{\mathcal{D}}(\boldsymbol{u})$ will be treated implicitly. For the diffusive part $\boldsymbol{\mathcal{D}}(\boldsymbol{u})$ we utilize an s-stage diagonally implicit (DIRK) scheme with coefficients $\boldsymbol{A}\in\mathbb{R}^{s\times s}$, $\boldsymbol{b},\boldsymbol{c}\in\mathbb{R}^s$, in the common Butcher notation, where $\boldsymbol{A}=(a_{ij})$ with $a_{ij}=0$ for $j>i$. For the convective term $\boldsymbol{\mathcal{C}}(\boldsymbol{u})$ we employ an s-stage explicit scheme with coefficients $\tilde{\boldsymbol{A}}\in\mathbb{R}^{s\times s}$, $\tilde{\boldsymbol{b}},\tilde{\boldsymbol{c}}\in\mathbb{R}^s$ and $\tilde{\boldsymbol{A}}=(\tilde{a}_{ij})$ with $\tilde{a}_{ij}=0$ for $j\geq i$. We denote the corresponding Butcher arrays by

$$\begin{array}{c|c}\boldsymbol{c}&\boldsymbol{A}\\\hline&\boldsymbol{b}^{\mathrm{T}}\end{array}\quad\text{and}\quad\begin{array}{c|c}\tilde{\boldsymbol{c}}&\tilde{\boldsymbol{A}}\\\hline&\tilde{\boldsymbol{b}}^{\mathrm{T}}\end{array}. \tag{2.18}$$

If applied to (1.4), then the IMEX-RK scheme gives rise to the following algorithm (see [1]):

Algorithm 3.1: IMEX-RK scheme

Input: approximate solution vector $\boldsymbol{u}^n$ for $t=t_n$
do $m=1,\dots,s$
solve for $\boldsymbol{u}^{(m)}$ the nonlinear equation
$$\boldsymbol{u}^{(m)}=\boldsymbol{u}^n+\Delta t\left(\sum_{j=1}^{m-1}a_{mj}K_j+\sum_{j=1}^{m-1}\tilde{a}_{mj}\tilde{K}_j\right)+a_{mm}\Delta t\boldsymbol{\mathcal{D}}(\boldsymbol{u}^{(m)})$$
$K_m\leftarrow\boldsymbol{\mathcal{D}}(\boldsymbol{u}^{(m)})$, $\tilde{K}_m\leftarrow\boldsymbol{\mathcal{C}}(\boldsymbol{u}^{(m)})$
enddo
$$\boldsymbol{u}^{n+1}\leftarrow\boldsymbol{u}^n+\Delta t\sum_{j=1}^{s}b_jK_j+\Delta t\sum_{j=1}^{s}\tilde{b}_j\tilde{K}_j$$
Output: approximate solution vector $\boldsymbol{u}^{n+1}$ for $t=t^{n+1}=t^n+\Delta t$.

Algorithm 3.1 requires solving for the vector $\boldsymbol{u} = \boldsymbol{u}^{(m)}$ a nonlinear system of scalar equations of the form

$$\boldsymbol{F}_m(\boldsymbol{u}) := \boldsymbol{u} - a_{mm}\Delta t\boldsymbol{\mathcal{D}}(\boldsymbol{u}) - \boldsymbol{r}_m = \boldsymbol{0}, \quad m = 1, \ldots, s, \tag{2.19}$$

where the vector $\boldsymbol{r}_m$ is given by

$$\boldsymbol{r}_m = \boldsymbol{u}^n + \Delta t \left(\sum_{j=1}^{m-1} a_{mj} K_j + \sum_{j=1}^{m-1} \tilde{a}_{mj} \tilde{K}_j \right).$$

The solution of (2.19) is positive as long as $\boldsymbol{r}_m$ is positive. The following result, which is a generalization of [9, Th. 2.2], deals with the solution of (2.19). (This theorem is formulated for a general system of nonlinear equations of a particular form; in the subsequent Corollary 2.1, we will apply it to the special case of (2.19).)

Theorem 2.2. *Let $\boldsymbol{G}$ be a symmetric invertible diagonally dominant $M \times M$ matrix, with positive diagonal entries and non-positive off-diagonal entries and $\boldsymbol{w} \in \mathbb{R}^M$, $\boldsymbol{w} \geq \boldsymbol{0}$, where such inequalities for vectors and matrices are understood in the component-wise sense. If Φ satisfies* (2.1) *and $\boldsymbol{\Phi}$ denotes its vectorial component-wise extension $\boldsymbol{\Phi}(\boldsymbol{u})_i = \Phi(u_i)$, then the equation*

$$\boldsymbol{z} + \boldsymbol{G}\boldsymbol{\Phi}(\boldsymbol{z}) = \boldsymbol{w} \tag{2.20}$$

has a unique solution $\boldsymbol{z} \in \mathbb{R}^M$ satisfying $\boldsymbol{z} \geq \boldsymbol{0}$.

Proof. We define the function

$$L(u) := \begin{cases} \Phi(u)/u & \text{if } u > 0, \\ 0 & \text{if } u = 0. \end{cases}$$

In view of the requirements in (2.1), L is continuous in $[0, \infty)$ and $L(u), \Phi(u) \geq 0$ for $u \geq 0$. Let

$$\boldsymbol{E}(\boldsymbol{z}) := \operatorname{diag}\big(L(z_1), \ldots, L(z_M)\big),$$

then $\boldsymbol{\Phi}(\boldsymbol{z}) = \boldsymbol{E}(\boldsymbol{z})\boldsymbol{z}$. We denote by $\boldsymbol{I}$ the $M \times M$ identity matrix. For $\boldsymbol{z} \geq \boldsymbol{0}$, the matrix $\boldsymbol{I} + \boldsymbol{G}\boldsymbol{E}(\boldsymbol{z})$ is strictly diagonally dominant (by columns) with positive diagonal entries and non-positive off-diagonal entries, and therefore $(\boldsymbol{I} + \boldsymbol{G}\boldsymbol{E}(\boldsymbol{z}))^{-1}$ is a non-negative matrix and it is a continuous function of $\boldsymbol{z}$. Then, the solution of equation (2.20) is reduced to finding fixed points of the mapping $\boldsymbol{z} \mapsto \boldsymbol{\varphi}(\boldsymbol{z}) = (\boldsymbol{I} + \boldsymbol{G}\boldsymbol{E}(\boldsymbol{z}))^{-1}\boldsymbol{w}$. To assess existence of fixed points, we aim to apply Brouwer's theorem to $\boldsymbol{\varphi}$ and the compact and convex set $\mathcal{K} := \{\boldsymbol{z} \in \mathbb{R}^M \mid \boldsymbol{z} \geq \boldsymbol{0} \text{ and } \|\boldsymbol{z}\|_1 \leq \|\boldsymbol{w}\|_1\}$. Clearly, $(\boldsymbol{I} + \boldsymbol{G}\boldsymbol{E}(\boldsymbol{z}))^{-1} \geq \boldsymbol{0}$ and $\boldsymbol{w} \geq \boldsymbol{0}$ immediately yield $\boldsymbol{\varphi}(\boldsymbol{z}) \geq \boldsymbol{0}$ for all $\boldsymbol{z} \in \mathcal{K}$, so, to prove that $\boldsymbol{\varphi}(\mathcal{K}) \subseteq \mathcal{K}$, there only remains to prove that

$$\big\|\boldsymbol{\varphi}(\boldsymbol{z})\big\|_1 \leq \|\boldsymbol{w}\|_1 \quad \text{for all } \boldsymbol{z} \in \mathcal{K}. \tag{2.21}$$

To this end, we take into account that

$$\big\|\boldsymbol{\varphi}(\boldsymbol{z})\big\|_1 \leq \big\|\big(\boldsymbol{I} + \boldsymbol{G}\boldsymbol{E}(\boldsymbol{z})\big)^{-1}\big\|_1 \|\boldsymbol{w}\|_1.$$

Thus, to establish (2.21) it is sufficient to prove that

$$\big\|\big(\boldsymbol{I} + \boldsymbol{G}\boldsymbol{E}(\boldsymbol{z})\big)^{-1}\big\|_1 \leq 1 \quad \text{for all } \boldsymbol{z} \in \mathcal{K}.$$

For this purpose, we use the auxiliary matrix $\tilde{\boldsymbol{G}} = (\tilde{G}_{ij})_{1 \leq i,j \leq M}$ defined by

$$\tilde{G}_{ij} := \begin{cases} G_{ij} & \text{if } i \neq j, \\ -\sum_{k \neq i} G_{ik} & \text{if } i = j \end{cases}$$

and the notation $\boldsymbol{H} := \boldsymbol{I} + \boldsymbol{G}\boldsymbol{E}(\boldsymbol{z})$ and $\tilde{\boldsymbol{H}} := \boldsymbol{I} + \tilde{\boldsymbol{G}}\boldsymbol{E}(\boldsymbol{z})$. Since $\tilde{\boldsymbol{H}}$ is also a strictly diagonally dominant matrix (by columns) with positive diagonal entries and non-positive off-diagonal entries, $\tilde{\boldsymbol{H}}^{-1} \geq \boldsymbol{0}$. Now, for $\boldsymbol{e} := (1, \dots, 1)^{\mathrm{T}} \in \mathbb{R}^M$ it follows that $\boldsymbol{e}^{\mathrm{T}}\tilde{\boldsymbol{G}} = \boldsymbol{0}$, so $\boldsymbol{e}^{\mathrm{T}}\tilde{\boldsymbol{H}} = \boldsymbol{e}^{\mathrm{T}}$ and $\boldsymbol{e}^{\mathrm{T}}\tilde{\boldsymbol{H}}^{-1} = \boldsymbol{e}^{\mathrm{T}}$. If we assume that $\boldsymbol{H}^{-1} = (\bar{\eta}_{ij})_{1\leq i,j\leq M}$ and $\tilde{\boldsymbol{H}}^{-1} = (\bar{\mu}_{ij})_{1\leq i,j\leq M}$, then this is equivalent to $\bar{\mu}_{1j} + \dots + \bar{\mu}_{Mj} = 1$ for $j = 1, \dots, M$. Furthermore, since $\boldsymbol{H}^{-1} \geq \boldsymbol{0}$, $\tilde{\boldsymbol{H}}^{-1} \geq \boldsymbol{0}$,

$$\boldsymbol{H} - \tilde{\boldsymbol{H}} = (\boldsymbol{G} - \tilde{\boldsymbol{G}})\boldsymbol{E}(\boldsymbol{z}) = \operatorname*{diag}_{1\leq i\leq M}\left(\left(G_{ii} - \sum_{j\neq i} G_{ij}\right)L(\boldsymbol{z}_i)\right) \geq 0 \quad \text{for } \boldsymbol{z} \in \mathcal{K}$$

and $\boldsymbol{H}^{-1} = \tilde{\boldsymbol{H}}^{-1} - \boldsymbol{H}^{-1}(\boldsymbol{H} - \tilde{\boldsymbol{H}})\tilde{\boldsymbol{H}}^{-1}$, it follows that $\boldsymbol{H}^{-1} \leq \tilde{\boldsymbol{H}}^{-1}$. This yields that

$$\|\boldsymbol{H}^{-1}\|_1 = \max_{1\leq j\leq M} \sum_{i=1}^{M} \bar{\eta}_{ij} \leq \max_{1\leq j\leq M} \sum_{i=1}^{M} \bar{\mu}_{ij} = 1.$$

Applying Brouwer's fixed point theorem to the continuous function $\boldsymbol{\varphi}\colon \mathcal{K} \to \mathcal{K}$ we deduce the existence of a fixed point of $\boldsymbol{\varphi}$, i.e. a non-negative solution to equation (2.20).

For uniqueness, we adapt an argument that can be found in [22] and define

$$\Psi(\boldsymbol{z}) := \sum_{i=1}^{M} N\big(z_i\big), \quad N(u) := \int_0^{|u|} \Phi(s)\,\mathrm{d}s, \quad \text{and} \quad f(\boldsymbol{z}) := \frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}^{-1}\boldsymbol{z} + \Psi(\boldsymbol{z}) - \boldsymbol{z}^{\mathrm{T}}\boldsymbol{G}^{-1}\boldsymbol{w}.$$

Since $\Phi(0) = 0$, it follows from the definition that $N'(u) = \operatorname{sgn}(u)\Phi(|u|)$ and $N''(u) = \Phi'(|u|)$ for any $u \in \mathbb{R}$, so $N \in C^2(\mathbb{R})$. Therefore, Ψ is twice continuously differentiable. Thus, f is also twice continuously differentiable and its gradient $f'(\boldsymbol{z})$ and Hessian $f''(\boldsymbol{z})$ are given by the respective expressions

$$\begin{aligned} f'(\boldsymbol{z})^{\mathrm{T}} &= \boldsymbol{G}^{-1}\boldsymbol{z} + \big(\operatorname{sgn}(z_1)\Phi(|z_1|), \dots, \operatorname{sgn}(z_M)\Phi(|z_M|)\big)^{\mathrm{T}} - \boldsymbol{G}^{-1}\boldsymbol{w}, \\ f''(\boldsymbol{z}) &= \boldsymbol{G}^{-1} + \operatorname{diag}\big(\Phi'(|z_1|), \dots, \Phi'(|z_M|)\big). \end{aligned}$$

Since $\boldsymbol{G}^{-1}$ is symmetric and positive definite and $\Phi'(|z_i|) \geq 0$, it follows that $f''(\boldsymbol{z})$ is symmetric and positive definite, therefore f is strictly convex, so any critical point (at which $f'(\boldsymbol{z}) = \boldsymbol{0}$) is the unique global minimum. Now, if $\boldsymbol{z} + \boldsymbol{G}\boldsymbol{\Phi}(\boldsymbol{z}) = \boldsymbol{w}$ with $\boldsymbol{z} \geq \boldsymbol{0}$, then $f'(\boldsymbol{z}) = \boldsymbol{0}$ and $\boldsymbol{z} \in \mathcal{K}$, so positive solutions of (2.20) are critical points of f, so uniqueness is proven. □

Corollary 2.1. *If Φ satisfies the conditions* (2.1) *and*

$$\max_{1\leq l\leq d}\left\{\frac{\Delta t}{\Delta x_l}\max_{\boldsymbol{k}\in\mathcal{M}}\big|v^l_{\boldsymbol{k}+\frac{1}{2}\boldsymbol{e}_l}\big|\right\} \leq \frac{1}{2d}, \tag{2.22}$$

then the Euler IMEX method

$$\boldsymbol{u}^{n+1} = \boldsymbol{u}^n + \Delta t\big(\boldsymbol{\mathcal{C}}(\boldsymbol{u}^n) + \boldsymbol{\mathcal{D}}(\boldsymbol{u}^{n+1})\big) \tag{2.23}$$

is a positivity-preserving scheme.

Proof. To be able to apply Theorem 2.2 to the situation at hand, we must be explicit on how we write the semi-discrete formulation as a system of ordinary differential equations (ODEs). To this end, we assume that $\boldsymbol{u}(t)$ is an M_*-dimensional vector that represents an arrangement of the M_* unknown functions $u_{\boldsymbol{i}}$, $\boldsymbol{i} \in \mathcal{M}$. Specifically, we fix a bijective map $\nu : \mathcal{M} \ni \boldsymbol{i} \mapsto \nu(\boldsymbol{i}) \in \{1, \dots, M_*\}$, with inverse $\eta : \{1, \dots, M_*\} \ni m \mapsto \eta(m) \in \mathcal{M}$, that maps the d-dimensional index $\boldsymbol{i}$ to the corresponding position within the vector $\boldsymbol{u}(t)$. Now the nonlinear equation (2.19) can be written in the form (2.20) for $M = M_*$ as follows. Suppose that $\boldsymbol{u}^n = (u^n_{\boldsymbol{i}})_{\boldsymbol{i}\in\mathcal{M}}$, then we set

$$\boldsymbol{z} = (z_1, \dots, z_{M_*})^{\mathrm{T}} = \big(u^{n+1}_{\eta(1)}, \dots, u^{n+1}_{\eta(M_*)}\big)^{\mathrm{T}}.$$

Moreover, if we set correspondingly

$$\boldsymbol{\Phi}(\boldsymbol{z}) = \big(\Phi\big(u^{n+1}_{\eta(1)}\big), \ldots, \Phi\big(u^{n+1}_{\eta(M_*)}\big)\big)^{\mathrm{T}},$$

then (2.10) stipulates that (2.23) or equivalently,

$$\boldsymbol{u}^{n+1} - \Delta t \boldsymbol{\mathcal{D}}(\boldsymbol{u}^{n+1}) = \boldsymbol{u}^n + \Delta t \boldsymbol{\mathcal{C}}(\boldsymbol{u}^n),$$

can be written as (2.20) if we define $\boldsymbol{G} = (G_{ij})_{1\leq i,j\leq M_*}$ by

$$G_{ij} = \Delta t \cdot \begin{cases} 2d & \text{if } i = j, \\ -1 & \text{if } i \neq j \text{ and } j = \nu(\eta(i) \pm \boldsymbol{e}_l) \text{ for } l \in \{1, \ldots, d\}, \\ 0 & \text{otherwise,} \end{cases}$$

and analogously define the entries of $\boldsymbol{w} = (w_1, \ldots, w_{M_*})$ by

$$w_i = u^n_{\eta(i)} + \Delta t C(\boldsymbol{u}^n)_{\eta(i)}, \quad i = 1, \ldots, M_*.$$

By Theorem 2.1 with $\Phi = 0$, condition (2.22) guarantees that $w_i \geq 0$ for $i = 1, \ldots, M_*$, so the statement of the corollary follows if we apply Theorem 2.2 to the system (2.20) under the present interpretations of $\boldsymbol{z}$, $\boldsymbol{G}$ and $\boldsymbol{w}$. □

Unfortunately, Corollary 2.1 cannot be directly applied to higher-order IMEX-RK schemes, since there cannot exist Runge-Kutta implicit schemes in SSP form of order higher than one (see [17]), so Corollary 2.1 cannot be in principle applied for second-order accuracy in time. We have nevertheless used Newton-Raphson method, together with a line search algorithm (see [10]) to solve (2.19). At each step of this algorithm a particular sparse system is solved (apart from the diagonal entry in each row, only $2d$ off-diagonal entries are occupied; details in Section 2.4). We have not experienced any troubles in solving these systems under a stability restriction as (2.22).

2.4. **Linear solver.** The (damped) Newton's method applied to (2.20) or, equivalently, to

$$\boldsymbol{F}(\boldsymbol{z}) := \boldsymbol{z} + \boldsymbol{G}\boldsymbol{\Phi}(\boldsymbol{z}) - \boldsymbol{w} = \boldsymbol{0} \tag{2.24}$$

consists in the iteration

$$\boldsymbol{F}'(\boldsymbol{z}^\nu)\boldsymbol{\delta}^\nu = -\boldsymbol{F}(\boldsymbol{z}^\nu), \quad \boldsymbol{z}^{\nu+1} = \boldsymbol{z}^\nu + \alpha^\nu \boldsymbol{\delta}^\nu, \quad \nu = 0, 1, 2, \ldots,$$

where the scalar α^ν is selected using a line-search algorithm that enforces sufficient decrease of the function $\alpha \mapsto \|\boldsymbol{F}(\boldsymbol{z}^{\boldsymbol{\nu}} + \alpha\boldsymbol{\delta}^\nu)\|_2^2$ (see [15]). The structure of the Jacobian matrix associated with (2.24) is particularly simple, namely $\boldsymbol{J} := \boldsymbol{F}'(\boldsymbol{z}) := \boldsymbol{I} + \boldsymbol{G}\boldsymbol{D}$, where $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{\Phi}'(\boldsymbol{z}))$. To get a favorable structure when solving $\boldsymbol{J}\boldsymbol{\delta} = -\boldsymbol{F}(\boldsymbol{z})$, we notice that the columns corresponding to $z_k = 0$ are zero. In particular, the equations for those k are explicit, so the only equations to be solved are those for $z_k \neq 0$. In algebraic terms, we define $M_* := M_1 M_2 \cdots M_d$ and assume that the vector $\boldsymbol{z} \in (\mathbb{R}_0^+)^{M_*}$ is an arrangement of $\{u_{\boldsymbol{i}}\}_{\boldsymbol{i}\in\mathcal{M}}$, taking into account (2.4) and (2.3). For a fixed vector $\boldsymbol{z} = (z_1, \ldots, z_{M_*})^{\mathrm{T}}$ of this dimension, we define the index set

$$\mathcal{I} = \mathcal{I}(\boldsymbol{z}) := \{k \mid \Phi'(z_k) > 0\} = \{k_1 < k_2 < \ldots < k_r\} \subseteq \{1, \ldots, M_*\}$$

and its complement

$$\mathcal{L} := \{1, \ldots, M_*\} \backslash \mathcal{I}(\boldsymbol{z}) = \{k \mid \Phi'(\boldsymbol{z}_k) = 0\} = \{j_1 < j_2 < \ldots < j_{\bar{r}}\}.$$

and consider the permutation of $(1, \ldots, M_*)$ given by $(k_1, k_2, \ldots, k_r, j_1, \ldots, j_{\bar{r}})$, with associated permutation matrix $\boldsymbol{P}$. If $\boldsymbol{\mathcal{A}} = (\mathcal{A}_{ij})_{1\leq i,j\leq M_*}$ is any $M_* \times M_*$ matrix, then

$$\boldsymbol{P}\boldsymbol{\mathcal{A}}\boldsymbol{P}^{\mathrm{T}} = \begin{bmatrix} \boldsymbol{\mathcal{A}}_{\mathcal{I},\mathcal{I}} & \boldsymbol{\mathcal{A}}_{\mathcal{I},\mathcal{L}} \\ \boldsymbol{\mathcal{A}}_{\mathcal{L},\mathcal{I}} & \boldsymbol{\mathcal{A}}_{\mathcal{L},\mathcal{L}} \end{bmatrix}$$

with the submatrices

$$(\boldsymbol{\mathcal{A}}_{\mathcal{I},\mathcal{I}})_{p,q} = \mathcal{A}_{k_p,k_q}, \quad (\boldsymbol{\mathcal{A}}_{\mathcal{I},\mathcal{L}})_{p,m} = \mathcal{A}_{k_p,j_m}, \quad (\boldsymbol{\mathcal{A}}_{\mathcal{L},\mathcal{I}})_{l,q} = \mathcal{A}_{j_l,k_q}, \quad (\boldsymbol{\mathcal{A}}_{\mathcal{L},\mathcal{L}})_{l,m} = \mathcal{A}_{j_l,j_m}.$$

Consequently, the matrix $\boldsymbol{PJP}^{\mathrm{T}}$ has the following block structure, where $\boldsymbol{I}_r$ and $\boldsymbol{I}_{\bar{r}}$ denote the $r \times r$ and $\bar{r} \times \bar{r}$ identity matrix, respectively:

$$\boldsymbol{PJP}^{\mathrm{T}} = \boldsymbol{I} + \boldsymbol{PGP}^{\mathrm{T}}\boldsymbol{PDP}^{\mathrm{T}} = \begin{bmatrix} \boldsymbol{I}_r & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{\bar{r}} \end{bmatrix} + \begin{bmatrix} \boldsymbol{G}_{\mathcal{I},\mathcal{I}} & \boldsymbol{G}_{\mathcal{I},\mathcal{L}} \\ \boldsymbol{G}_{\mathcal{L},\mathcal{I}} & \boldsymbol{G}_{\mathcal{L},\mathcal{L}} \end{bmatrix} \begin{bmatrix} \boldsymbol{D}_{\mathcal{I},\mathcal{I}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I}_r + \boldsymbol{G}_{\mathcal{I},\mathcal{I}}\boldsymbol{D}_{\mathcal{I},\mathcal{I}} & \boldsymbol{0} \\ \boldsymbol{G}_{\mathcal{L},\mathcal{I}}\boldsymbol{D}_{\mathcal{I},\mathcal{I}} & \boldsymbol{I}_{\bar{r}} \end{bmatrix},$$

Therefore, the solution of $\boldsymbol{J\delta} = -\boldsymbol{F}(\boldsymbol{z})$ can be obtained by solving $\boldsymbol{PJP}^{\mathrm{T}}\boldsymbol{P\delta} = -\boldsymbol{PF}(\boldsymbol{z})$, that is,

$$\begin{bmatrix} \boldsymbol{I}_r + \boldsymbol{G}_{\mathcal{I},\mathcal{I}}\boldsymbol{D}_{\mathcal{I},\mathcal{I}} & 0 \\ \boldsymbol{G}_{\mathcal{L},\mathcal{I}}\boldsymbol{D}_{\mathcal{I},\mathcal{I}} & \boldsymbol{I}_{\bar{r}} \end{bmatrix} \begin{pmatrix} \boldsymbol{\delta}_{\mathcal{I}} \\ \boldsymbol{\delta}_{\mathcal{L}} \end{pmatrix} = - \begin{pmatrix} \boldsymbol{F}(\boldsymbol{z})_{\mathcal{I}} \\ \boldsymbol{F}(\boldsymbol{z})_{\mathcal{L}} \end{pmatrix},$$

which means that in each iteration, we first determine $\boldsymbol{\delta}_{\mathcal{I}}$ by solving

$$(\boldsymbol{I}_r + \boldsymbol{G}_{\mathcal{I},\mathcal{I}}\boldsymbol{D}_{\mathcal{I},\mathcal{I}})\,\boldsymbol{\delta}_{\mathcal{I}} = -\boldsymbol{F}(\boldsymbol{z})_{\mathcal{I}}, \tag{2.25}$$

and then calculate $\boldsymbol{\delta}_{\mathcal{L}}$ by evaluating

$$\boldsymbol{\delta}_{\mathcal{L}} = -\boldsymbol{F}(\boldsymbol{z})_{\mathcal{L}} - \boldsymbol{G}_{\mathcal{L},\mathcal{I}}\boldsymbol{D}_{\mathcal{I},\mathcal{I}}\boldsymbol{\delta}_{\mathcal{I}}.$$

The matrix of the system (2.25) can be written as $\hat{\boldsymbol{J}} = \boldsymbol{I}_r + \hat{\boldsymbol{G}}\hat{\boldsymbol{D}}$, where $\hat{\boldsymbol{G}}$ and $\hat{\boldsymbol{D}}$ are the corresponding submatrices of $\boldsymbol{G}$ and $\boldsymbol{D}$, respectively. Since the diagonal entries of $\hat{\boldsymbol{D}}$ are positive, $\hat{\boldsymbol{J}}$ can be transformed into a symmetric and positive definite matrix by

$$\hat{\boldsymbol{D}}^{1/2}\hat{\boldsymbol{J}}\hat{\boldsymbol{D}}^{-1/2} = \boldsymbol{I}_r + \hat{\boldsymbol{D}}^{1/2}\hat{\boldsymbol{G}}\hat{\boldsymbol{D}}^{1/2}.$$

Therefore system (2.25) can be solved by solving first

$$\big(\boldsymbol{I}_r + \hat{\boldsymbol{D}}^{1/2}\hat{\boldsymbol{G}}\hat{\boldsymbol{D}}^{1/2}\big)\hat{\boldsymbol{\delta}} = -\hat{\boldsymbol{D}}^{1/2}\boldsymbol{F}(\boldsymbol{z})_{\mathcal{I}} \tag{2.26}$$

and then evaluating

$$\boldsymbol{\delta}_{\mathcal{I}} = \hat{\boldsymbol{D}}^{1/2}\hat{\boldsymbol{\delta}}.$$

The solution of (2.26) can be obtained by applying the conjugate gradient method.

3. Numerical examples

3.1. **IMEX-RK schemes and CFL condition.** We solve numerically (1.1), (1.2) for $0 \le t \le T$ and $\boldsymbol{x} \in \Omega$ (see (2.2)). We compare numerical results obtained by the IMEX approach proposed herein with those obtained by the explicit scheme of [4]. To demonstrate that the IMEX schemes are more efficient than the explicit one independently of the particular choice of the specific IMEX-RK scheme as given by its pair of Butcher arrays (2.18), we utilize and partly compare results produced by three different IMEX-RK schemes, namely the scheme

$$\begin{array}{c|c} \boldsymbol{c} & \boldsymbol{A} \\ \hline & \boldsymbol{b}^{\mathrm{T}} \end{array} = \begin{array}{c|cc} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \\ \hline & 1/2 & 1/2 \end{array}, \quad \begin{array}{c|c} \tilde{\boldsymbol{c}} & \tilde{\boldsymbol{A}} \\ \hline & \tilde{\boldsymbol{b}}^{\mathrm{T}} \end{array} = \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & 1/2 & 1/2 \end{array}, \tag{3.1}$$

denoted by H-CN(2,2,2) in [5] since it is a natural choice when dealing with convection-diffusion problems, since Heun's method is an SSP explicit RK one [17], and the Crank-Nicolson method is

A-stable and widely used for diffusion problems; the classical second-order IMEX-RK method

$$\begin{array}{c|c} \boldsymbol{c} & \boldsymbol{A} \\ \hline & \boldsymbol{b}^{\mathrm{T}} \end{array} = \begin{array}{c|ccc} 1/4 & 1/4 & 0 & 0 \\ 1/4 & 0 & 1/4 & 0 \\ 1 & 1/3 & 1/2 & 1/3 \\ \hline & 1/3 & 1/3 & 1/3 \end{array}, \quad \begin{array}{c|c} \tilde{\boldsymbol{c}} & \tilde{\boldsymbol{A}} \\ \hline & \tilde{\boldsymbol{b}}^{\mathrm{T}} \end{array} = \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 1 & 1/2 & 1/2 & 0 \\ \hline & 1/3 & 1/3 & 1/3 \end{array} \tag{3.2}$$

due to Pareschi and Russo [24], denoted by IMEX-SSP2(3,3,2) as in [5], and the third-order scheme

$$\begin{array}{c|c} \boldsymbol{c} & \boldsymbol{A} \\ \hline & \boldsymbol{b}^{\mathrm{T}} \end{array} = \begin{array}{c|ccccc} \alpha & \alpha & 0 & 0 & 0 \\ 0 & -\alpha & \alpha & 0 & 0 \\ 1 & 0 & 1-\alpha & \alpha & 0 \\ 1/2 & \beta & \eta & 1/2-\beta-\eta-\alpha & \alpha \\ \hline & 0 & 1/6 & 1/6 & 2/3 \end{array}, \quad \begin{array}{c|c} \tilde{\boldsymbol{c}} & \tilde{\boldsymbol{A}} \\ \hline & \tilde{\boldsymbol{b}}^{\mathrm{T}} \end{array} = \begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1/2 & 0 & 1/4 & 1/4 & 0 \\ \hline & 0 & 1/6 & 1/6 & 2/3, \end{array},$$

where $\alpha = 0.24169426078821, \quad \beta = \alpha/4, \quad \eta = 0.12915286960590,$

also introduced in [24], and which is here denoted by IMEX-SSP3(4,3,3) following [6].

For each iteration, the time step $\Delta t = \Delta t_n$ is determined by the formula

$$\Delta t \left(\frac{1}{d} \max_{0 \le u \le \eta(\boldsymbol{u})} \Phi'(u) \sum_{q=1}^{d} \frac{1}{\Delta x_q^2} + \max_{1 \le l \le d} \left\{ \frac{1}{\Delta x_l} \max_{\boldsymbol{k} \in \mathcal{M}} \left| v^{l,n}_{\boldsymbol{k}+\frac{1}{2}\boldsymbol{e}_l} \right| \right\} \right) = C_{\mathrm{cfl}_1}, \quad \eta(\boldsymbol{u}) := \max_{\boldsymbol{j} \in \mathcal{M}} u_{\boldsymbol{j}}, \tag{3.3}$$

for the explicit scheme and by

$$\Delta t \max_{1 \le l \le d} \left\{ \frac{1}{\Delta x_l} \max_{\boldsymbol{k} \in \mathcal{M}} \left| v^{l,n}_{\boldsymbol{k}+\frac{1}{2}\boldsymbol{e}_l} \right| \right\} = C_{\mathrm{cfl}_2} \tag{3.4}$$

for the IMEX-RK scheme. (Note that the left-hand sides of (3.3) and (3.4) are identical to those of (2.11) and (2.22), respectively, for $u_{\boldsymbol{i}} = u^n_{\boldsymbol{i}}$ for all $\boldsymbol{i} \in \mathcal{M}$.) In the numerical examples and for each time discretization we choose C_{cfl_1} and C_{cfl_2} as the largest multiple of 0.05 that yields oscillation-free solutions. This strategy leads to $C_{\mathrm{cfl}_1} = 0.25$ for the explicit scheme, $C_{\mathrm{cfl}_2} = 0.25$ for the H-CN(2,2,2) and IMEX-SSP3(4,3,3) schemes, and $C_{\mathrm{cfl}_1} = 0.2$ for the IMEX-SSP2(3,3,2) scheme, respectively.

3.2. **Approximate numerical error.** The approximate numerical error is measured for four of the five numerical examples, all of which are defined on a square domain with $M_1 = M_2 =: M$. In all these cases we compute a reference solution with $M = M_{\mathrm{ref}}$ utilizing the IMEX-RK H-CN(2,2,2) scheme. In each case the reference solution allows us to compute approximate L^1 errors at different times as follows. We have $\mathcal{M} = \{1, \ldots, M\}^2$, define $\mathcal{M}_{\mathrm{ref}} := \{1, \ldots, M_{\mathrm{ref}}\}^2$, and denote by

$$\left(u^M_{i,j}(t)\right)_{(i,j) \in \mathcal{M}} \quad \text{and} \quad \left(u^{M_{\mathrm{ref}}}_{i,j}(t)\right)_{(i,j) \in \mathcal{M}_{\mathrm{ref}}}$$

the numerical solution at time t calculated with M^2 and M^2_{ref} cells, respectively. We assume that $R := M_{\mathrm{ref}}/M$ is an integer and compute the projection of the reference solution

$$\tilde{u}^{\mathrm{ref},M}_{j,k}(t) = \frac{1}{R^2} \sum_{p,q=1}^{R} u^{M_{\mathrm{ref}}}_{R(j-1)+p,R(k-1)+q}(t).$$

The approximate L^1 error $e_M(t)$ associated with the numerical solution on the mesh with M^2 cells at time t is given by

$$e_M(t) = \frac{1}{M^2} \sum_{j,k=1}^{M} \left| \tilde{u}^{\mathrm{ref},M}_{j,k}(t) - u^M_{j,k}(t) \right|.$$

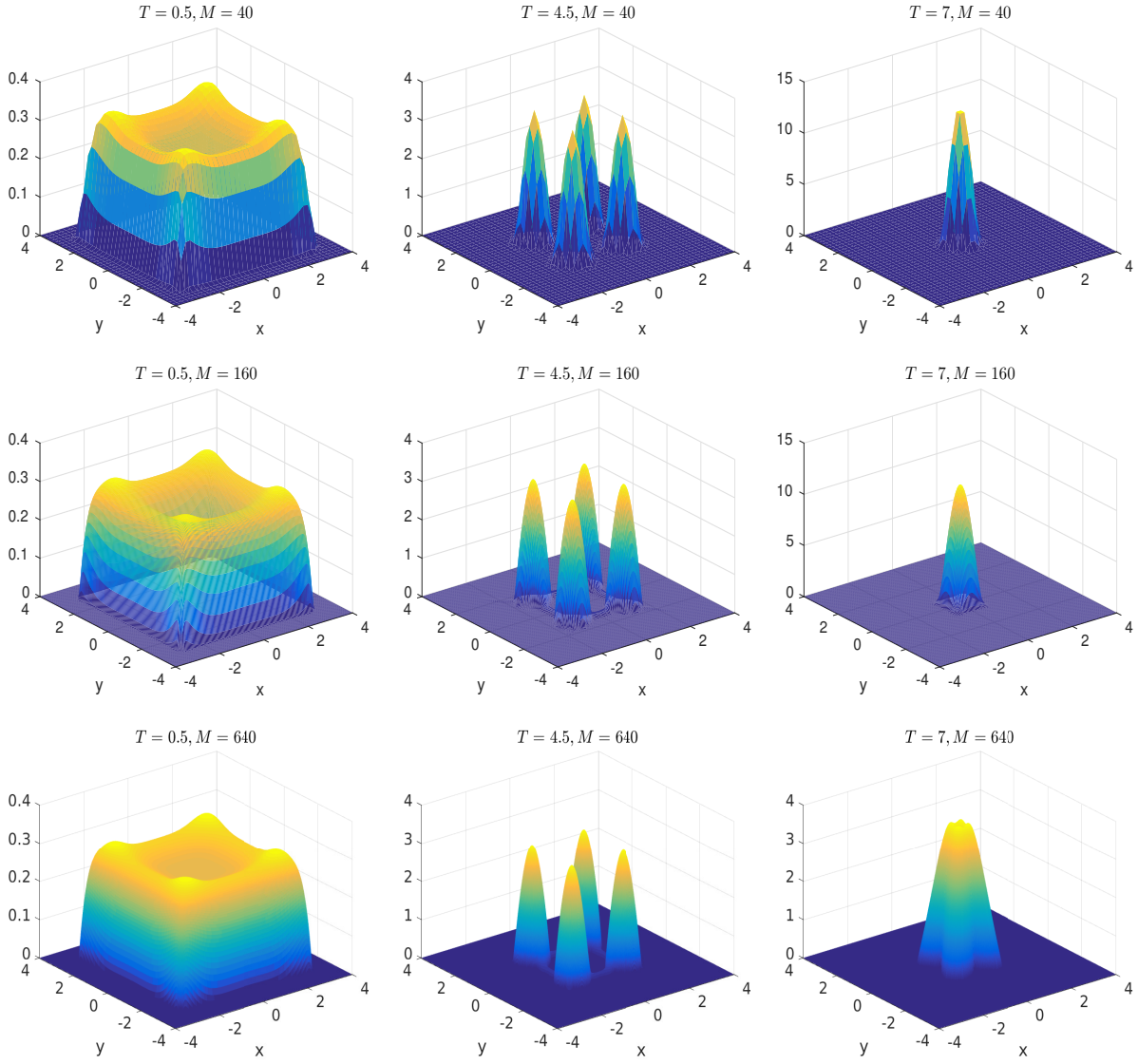

FIGURE 1. Example 1: numerical solutions with $\Delta x = 2L/M$ and $L = 4$ for (top) $M = 40$, (middle) $M = 160$, and (bottom) $M = 640$, at simulated times $T = 0.5$, 4.5, and 7. The IMEX-RK scheme used is H-CN(2,2,2) given by (3.1).

A numerical order of convergence can be calculated from pairs $e_{M/2}(t)$ and $e_M(t)$ by

$$\theta_M(t) := \log_2 \left(e_M(t)/e_{M/2}(t) \right).$$

3.3. **Numerical examples.**

3.3.1. *Example 1.* Following a numerical experiment proposed in [12], we solve (1.1) for

$$u_0(\boldsymbol{x}) = 0.25\chi_{[-3,3]\times[-3,3]}(\boldsymbol{x}), \quad W(\boldsymbol{x}) = -\frac{1}{\pi}\exp\left(-|\boldsymbol{x}|^2\right), \quad V \equiv 0,$$

$$H(u) = \frac{\nu}{m}u^m \Rightarrow \Phi(u) = \frac{\nu(m-1)}{m}u^m, \tag{3.5}$$

	IMEX-RK H-CN(2,2,2)			Explicit			IMEX-RK H-CN(2,2,2)			Explicit		
M	e_M	θ_M	cpu [s]	e_M	θ_M	cpu [s]	e_M	θ_M	cpu [s]	e_M	θ_M	cpu [s]
	$T = 0.5$						$T = 1$					
40	6923	—	0.50	6183	—	0.44	14391	—	0.69	13817	—	0.95
80	3437	1.01	0.73	3104	0.99	3.26	7089	1.02	1.80	6973	0.99	7.70
160	1595	1.11	5.50	1443	1.11	35.5	3355	1.08	12.0	3292	1.08	80.5
320	709.9	1.17	53.3	651.1	1.15	577	1507	1.15	103	1480	1.15	1267
640	290.3	1.29	466	270.8	1.27	10010	636.7	1.24	985	625.8	1.24	21849
	$T = 4.5$						$T = 5$					
40	35067	0.60	2.30	41592	0.72	12.2	40057	—	2.58	47500	—	14.7
80	23158	0.88	14.3	25320	0.95	135	24684	0.70	16.5	26910	0.82	164
160	12619	1.02	87.1	13137	1.05	1562	12629	0.97	100	13225	1.02	1945
320	6226	1.18	716	6340	1.20	21548	6032	1.07	821	6171	1.10	26936
640	2739	0.00	8065	2766	0.00	338985	2614	1.21	8982	2649	1.22	422759
	$T = 6.5$						$T = 7$					
40	269041	—	3.80	275375	—	25.9	172173	—	4.70	183920	0.00	36.2
80	165465	0.70	24.2	177743	0.63	252	178006	-0.05	31.3	183770	0.15	362
160	63488	1.38	141	65775	1.43	3050	163004	0.13	170	165729	0.93	3771
320	26180	1.28	1159	26637	1.30	42585	85463	0.93	1374	86850	1.31	48479
640	10577	1.31	12377	10691	1.32	665871	34566	1.31	14069	34931	0.00	741821

TABLE 1. Example 1: approximate L^1 errors (e_M, figures to be multiplied by 10^{-6}), convergence rates (θ_M), and CPU times (cpu).

where χ_A is the characteristic function of a set A and $\Phi(u)$ is defined in (1.3). In this example we choose $\nu = 0.1$ and $m = 2.1$, and limit ourselves to the IMEX-RK scheme H-CN(2,2,2) given by (3.1). The numerical results for various values of M are displayed in Figure 1. The approximate errors, convergence rates, and CPU times are provided in Table 1, where we compare the error of approximation with respect to a reference solution with $M_{\text{ref}} = 2560$ cells per direction. Figure 2 contains the efficiency plots for the end times T for which the errors are measured. According to Table 1, for $T \geq 4.5$ the errors and CPU times produced by IMEX-RK scheme are smaller than for the explicit version, and Figure 2 indicates that for these simulated times the IMEX-RK scheme is more efficient than the explicit version.

3.3.2. *Example 2.* We consider the numerical example in two dimensions proposed in [27]. Here u represents the population density and $W * u$ is the velocity. Specifically, we choose the functions

$$W(\boldsymbol{x}) = -\frac{1}{2\pi}\exp(-|\boldsymbol{x}|), \quad V \equiv 0,$$

and $\Phi(u)$ given by (3.5) with $\nu = 1/2$ and $m = 3$. The initial datum u_0 consists of two disjoint discs of radius 5 and centered at $(7, 0)$ and $(-7, 0)$, both with randomly distributed population with values between 0 and 1 such that the total population size is given by $\int_{\mathbb{R}^2} u_0(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 600$. The initial condition is defined for a 75×75 discretization, which is also utilized for finer discretizations with $M = 150, 300$ and 600, so the initial condition is exactly the same in all cases. The numerical solution for three different discretizations at four different times is displayed in Figure 3. The approximate errors, convergence rates, and CPU times for these times are provided in Table 2. The reference solution is computed with $M_{\text{ref}} = 2400$ cells per dimension. Figure 4 contains the efficiency plots for three different end times. We observe that for a given discretization M and

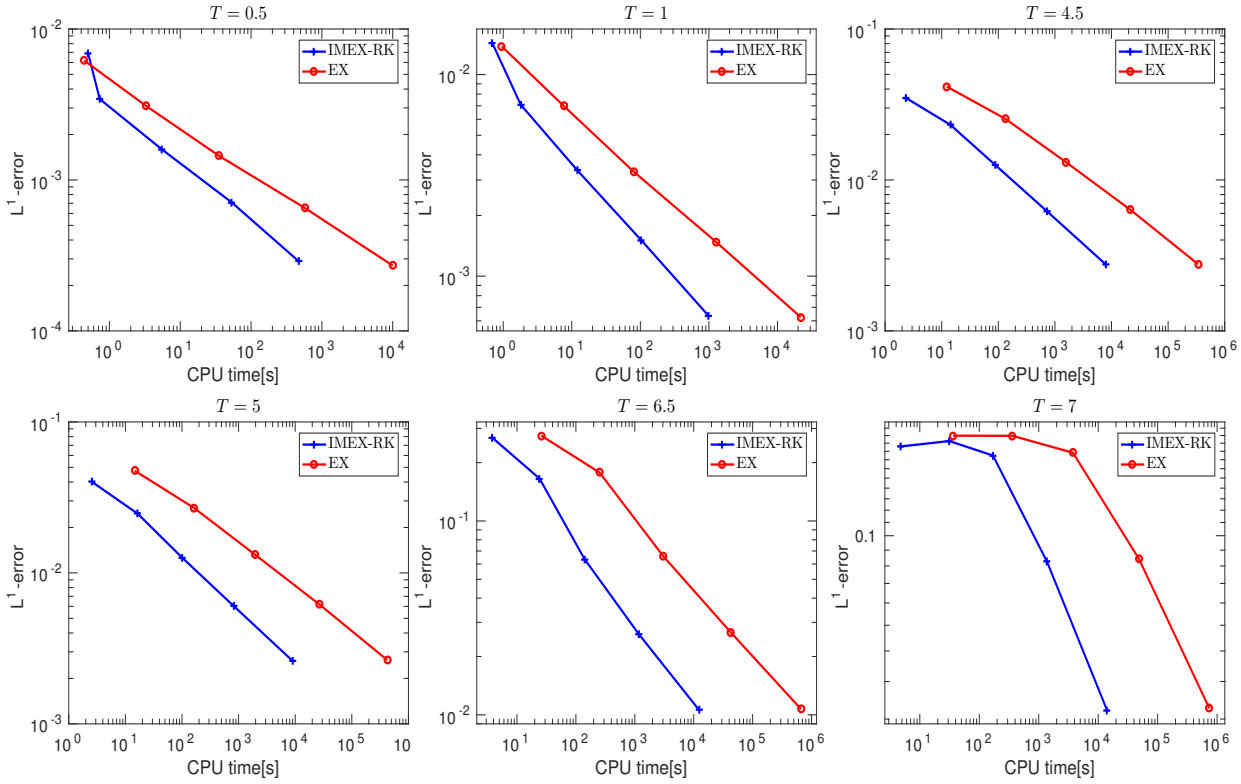

FIGURE 2. Example 1: efficiency plots corresponding to six simulated times. The IMEX-RK scheme employed is the scheme H-CN(2,2,2) given by (3.1).

with the exception of simulated time $T = 0.01$, the errors produced by the IMEX-RK schemes are smaller than those of the explicit scheme. Also, with the exception of some of the cases of $M = 75$, the CPU times for the IMEX-RK schemes are substantially smaller than for the explicit scheme. Thus, the IMEX-RK schemes are most efficient in all instances.

3.3.3. *Example 3.* This example is a 2D version of [12, Example 2]. More precisely, we utilize

$$u_0(\boldsymbol{x}) = 0.05\chi_{[-3,3]\times[-3,3]}(\boldsymbol{x}), \quad W(\boldsymbol{x}) = -\left(1 - |\boldsymbol{x}|\right)_+, \quad V \equiv 0,$$

and $\Phi(u)$ given by (3.5) with $\nu = 1.48$ and $m = 3$. The numerical solution for three different discretizations at four different times is displayed in Figure 5. The approximate errors, convergence rates, and CPU times for these times are provided in Table 3, where we compare the error of approximation with respect to the reference solution with $M_{\text{ref}} = 1280$ cells per dimension, and Figure 6 contains the efficiency plots for three different end times. For this example, the IMEX-RK schemes produce slightly smaller errors but are faster than the explicit scheme, and therefore turns out significantly more efficient.

3.3.4. *Example 4.* In this example we utilize the functions

$$W(\boldsymbol{x}) = \frac{1}{2\pi}\big(\exp(-|\boldsymbol{x} - \boldsymbol{x}_1|) + \exp(-|\boldsymbol{x} - \boldsymbol{x}_2|)\big),$$

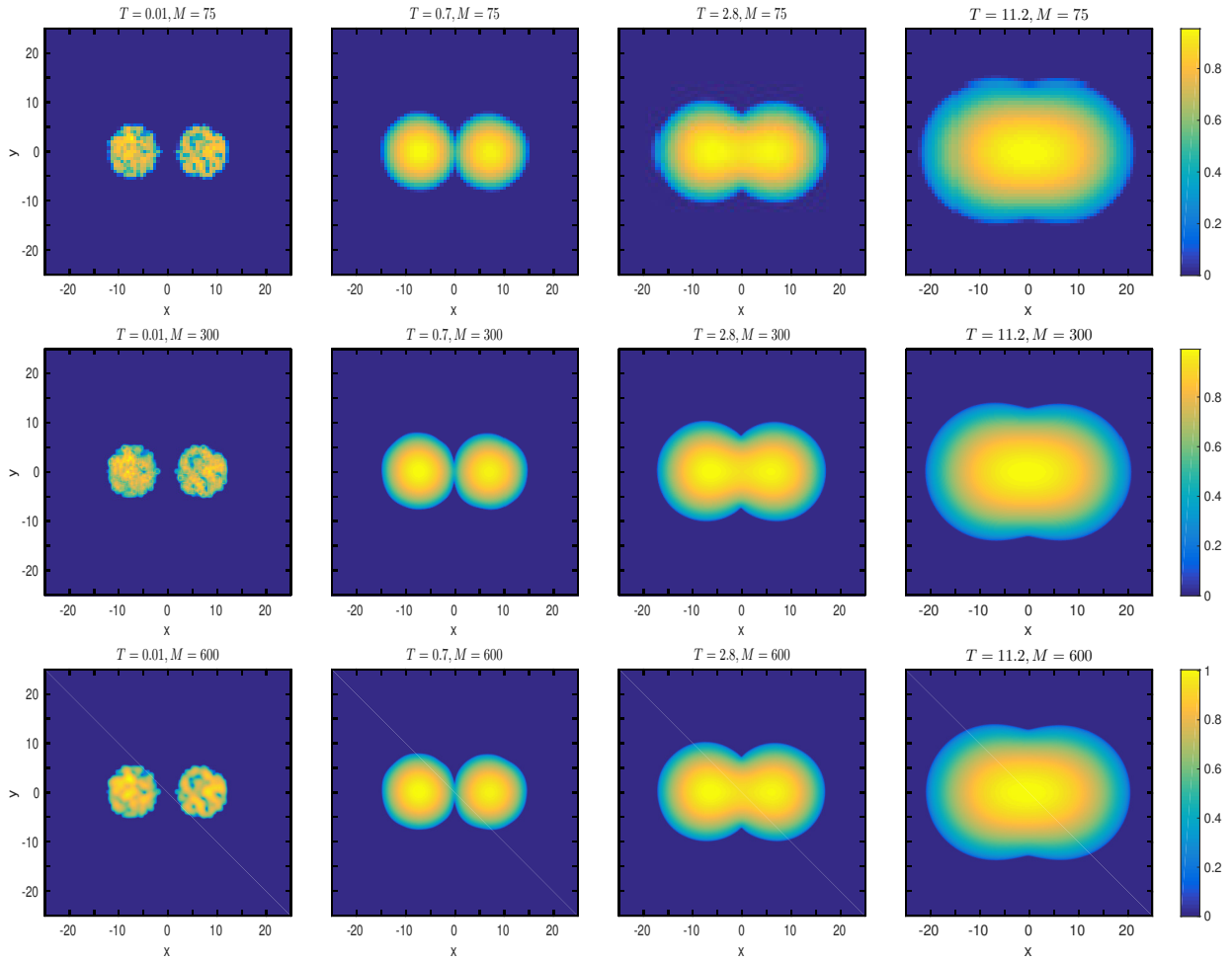

FIGURE 3. Example 2: numerical solutions with $\Delta x = 2L/M$ and $L = 20$ for (top) $M = 75$, (middle) $M = 300$ and (bottom) $M = 600$, at simulated times $T = 0.01$, 0.7, 2.8, and 11.2. The IMEX-RK scheme is H-CN(2,2,2) given by (3.1).

where $\boldsymbol{x}_1 = (7, 0)$ and $\boldsymbol{x}_2 = (-7, 0)$, $V \equiv 0$, and $\Phi(u)$ given by (3.5) with $\nu = 1$ and $m = 4$. The initial condition u_0 is a random function with values between 0 and 1 distributed over the (x_1, x_2)-square $[-30, 30] \times [-30, 30]$. The initial condition is defined for a discretization with $M = 40$, which is also used for finer discretizations as in Example 2. Numerical solutions for three different discretizations at four different simulations times are displayed in Figure 7. For this example the initial solution evolves until a steady state that consists of vertical stripes. The numerical solutions for the different discretizations converge to the same steady state solution. This can be observed in Table 4, where we compare the error of approximation with respect the reference solution which is computed with $M_{\text{ref}} = 1280$. Figure 8 contains the efficiency plots for three different end times. Table 4 indicates that the errors produced by IMEX-RK schemes are slightly smaller than those of the explicit scheme, but again the IMEX-RK schemes use less CPU time than the explicit scheme, and we conclude that the IMEX-RK schemes turn out more efficient than the explicit version.

It is instructive to compare the gain in efficiency of this example with that of Example 2. In view of the CFL conditions (3.3) and (3.4), the gain in efficiency by IMEX-RK schemes with respect

	IMEX-RK H-CN(2,2,2)			IMEX-RK IMEX-SSP2(3,3,2)			IMEX-RK IMEX-SSP3(4,3,3)			Explicit		
M	e_M	θ_M	cpu [s]	e_M	θ_M	cpu [s]	e_M	θ_M	cpu [s]	e_M	θ_M	cpu [s]
	$T = 0.01$											
75	27430	—	1.84	28858	—	0.87	30323	—	1.96	29394	—	1.49
150	19532	0.49	1.78	14933	0.95	1.57	10070	1.59	2.90	17487	0.75	14.9
300	34651	-0.83	17.6	26977	-0.85	28.6	15781	-0.65	21.7	19571	-0.16	185
600	49383	-0.51	381	12475	1.11	484	23922	-0.60	386	22320	-0.19	2626
	$T = 0.7$											
75	10480	—	12.5	10496	—	8.64	10545	—	15.5	11106	—	20.6
150	6787	0.63	56.8	6818	0.62	57.3	6833	0.63	72.8	7052	0.66	265
300	3188	1.09	594	3207	1.09	503	3218	1.09	629	3294	1.10	3647
600	1096	1.54	4878	1105	1.54	7367	1112	1.53	4822	1138	1.53	47957
	$T = 2.8$											
75	14212	—	39.4	14217	—	28.6	14231	—	52.3	14481	—	40.1
150	9044	0.65	182	9052	0.65	197	9057	0.65	236	9165	0.66	523
300	4244	1.09	1675	4249	1.09	1457	4252	1.09	1839	4296	1.09	7359
600	1461	1.54	11360	1463	1.54	15493	1465	1.54	11235	1482	1.54	123643
	$T = 11.2$											
75	17527	—	134	17528	—	128	17532	—	182	17631	—	72.4
150	11171	0.65	749	11173	0.65	853	11174	0.65	1034	11223	0.65	946
300	5264	1.09	4840	5265	1.09	5565	5266	1.09	5915	5288	1.09	13319
600	1817	1.53	27290	1818	1.53	44158	1818	1.53	31011	1827	1.53	288459

TABLE 2. Example 2: approximate L^1 errors (e_M, figures to be multiplied by 10^{-6}), convergence rates (θ_M) and CPU times (cpu).

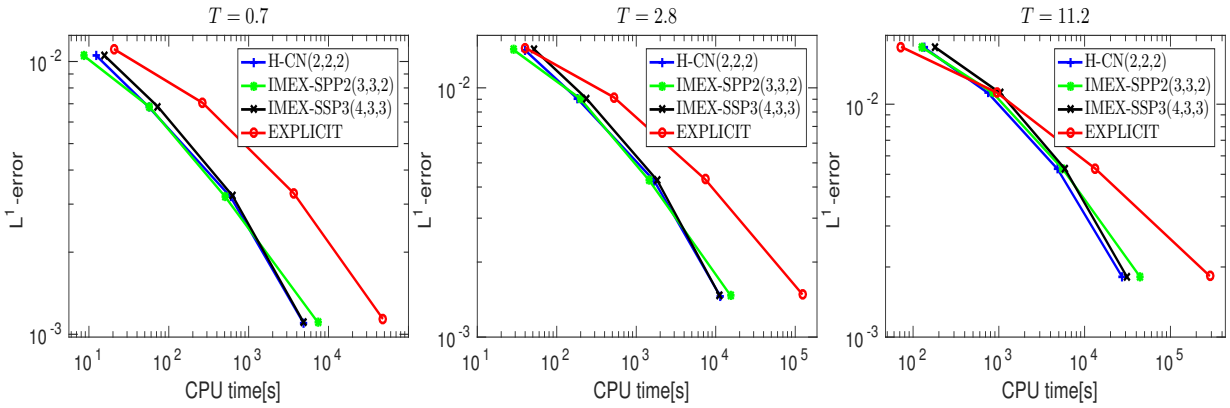

FIGURE 4. Example 2: efficiency plots based on numerical solutions for $M = 75, 150, 300, 600$.

to their explicit counterpart is likely to appear earlier (when discretization is successively refined) whenever the diffusion term is dominant, that is $\max_{0\le u\le \eta(\boldsymbol{u}^n)} \Phi'(u)$ (arising in (3.3)) is large in comparison with the maximum on the convective velocities (the term that arises in both (3.3) and (3.4)). In Example 2 we have $\Phi'(u) = u^2$ and in Example 4 there holds $\Phi'(u) = u^3$, with u-values ranging between 0 and 1 in Example 2 (see Figure 3) but only between 0 and 0.6 for $M = 160$ and

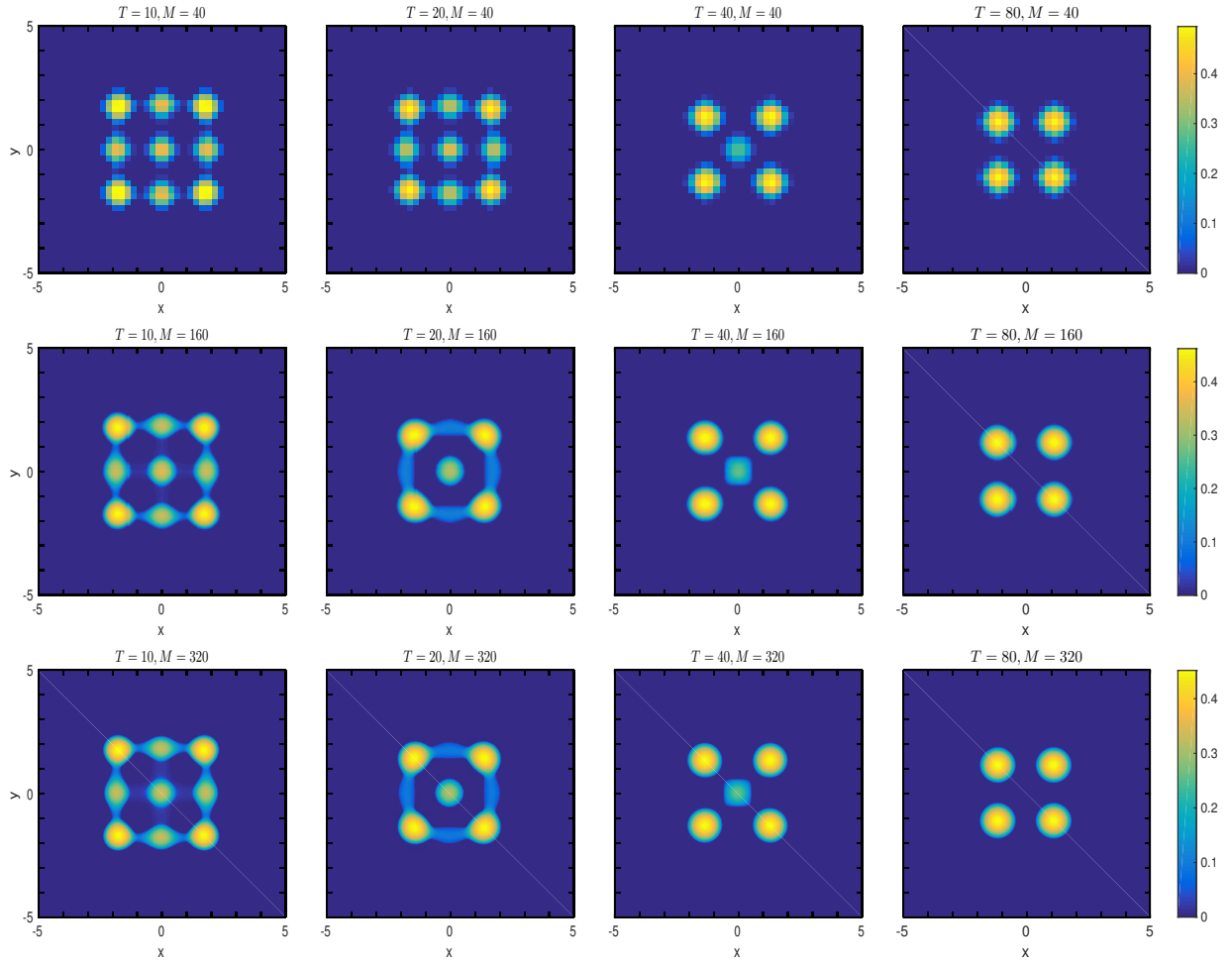

FIGURE 5. Example 3: numerical solutions with $\Delta x = 2L/M$ and $L = 5$ for (top) $M = 40$, (middle) $M = 160$ and (bottom) $M = 320$, at simulated times $T = 10$, 20, 40, and 80. The IMEX-RK scheme is given by (3.1).

$M = 320$ in Example 4 (see Figure 7). Since the dimensions of the domain of Examples 2 and 4 are the same and the interaction potential W produces similar values in both cases, one can roughly say that Example 2 is more diffusion dominant than Example 4, which explains why the gain in efficiency is better visible for the discretizations considered in the plots of Figure 4 (for Example 2) than for those of Figure 8.

3.3.5. *Example 5.* Finally, we present one additional example without error analysis but to demonstrate that (1.1), (1.2), and numerical methods developed for the approximation of its solutions, capture a model of swarming with diffusion [26]. In this context (1.1) represents the Fokker-Planck equation of the space-homogeneous version of a swarming model (see Section 1.2) whose solution $u = u(\boldsymbol{x}, t)$ is the density distribution of individuals having velocity $\boldsymbol{x} \in \mathbb{R}^d$ at time $t > 0$. For our example, the functions

$$W(\boldsymbol{x}) = \frac{1}{|\boldsymbol{x}|}, \quad V(\boldsymbol{x}) = \alpha \left(\frac{|\boldsymbol{x}|^4}{4} - \frac{|\boldsymbol{x}|^2}{2} \right), \quad \text{and } \Phi(u) = \nu u,$$

	IMEX-RK H-CN(2,2,2)			IMEX-RK IMEX-SSP2(3,3,2)			IMEX-RK IMEX-SSP3(4,3,3)			Explicit		
M	e_M	θ_M	cpu [s]	e_M	θ_M	cpu [s]	e_M	θ_M	cpu [s]	e_M	θ_M	cpu [s]
	$T = 10$											
40	4276	—	1.23	4294	—	4.62	4264	—	7.37	6528	—	11.0
80	3428	0.32	5.91	3417	0.33	42.7	3419	0.32	84.5	4110	0.67	118
160	2142	0.68	44.5	2140	0.68	310	2140	0.68	547	2262	0.86	1423
320	1042	1.04	301	1041	1.04	2797	1042	1.04	4325	1071	1.08	19170
	$T = 20$											
40	14161	—	2.60	14433	—	10.7	14202	—	16.9	20045	—	28.9
80	5298	1.42	15.2	5566	1.37	105	5352	1.41	213	4128	2.28	349
160	960	2.46	118	959	2.54	788	961	2.48	1299	1130	1.87	4738
320	580	0.73	819	571	0.75	7236	580	0.73	11774	735	0.62	67428
	$T = 40$											
40	2109	—	6.42	2139	—	26.6	2216	—	40.4	24943	—	64.4
80	3031	-0.52	38.0	3130	-0.55	273	3071	-0.47	578	2226	3.49	1148
160	1228	1.30	277	1235	1.34	1896	1233	1.32	3052	1170	0.93	15782
320	496	1.31	2000	497	1.31	16784	498	1.31	27367	506	1.21	228229
	$T = 80$											
40	2109	—	15.0	2076	—	63.8	2133	—	96.2	32239	—	135.3
80	1912	0.14	86.0	1919	0.11	626	1988	0.10	1306	2562	3.65	2892
160	1152	0.73	593	1154	0.73	4061	1163	0.77	6546	1154	1.15	39366
320	567	1.02	4335	567	1.03	35241	568	1.03	45427	537	1.10	571621

TABLE 3. Example 3: approximate L^1 errors (e_M, figures to be multiplied by 10^{-6}), convergence rates (θ_M), and CPU times (cpu).

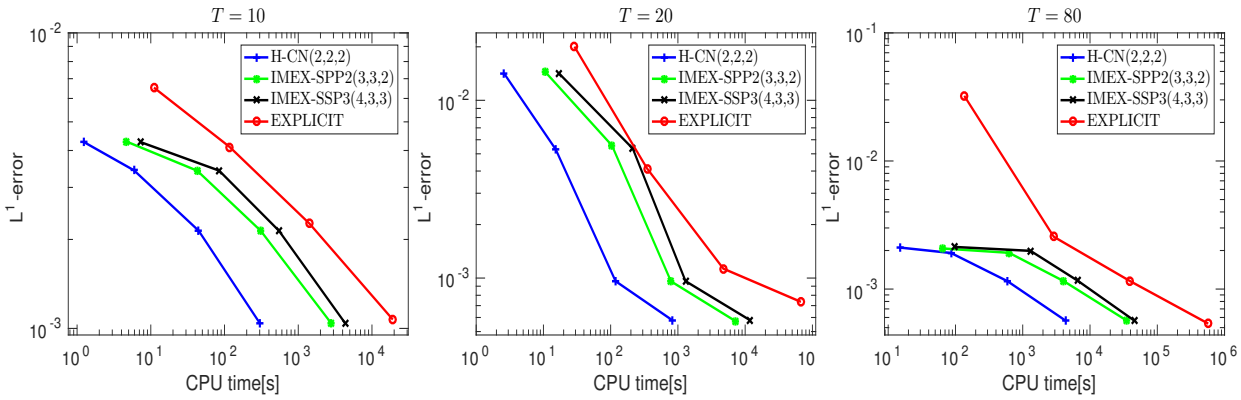

FIGURE 6. Example 3: efficiency plots based on numerical solutions for $\Delta x = 2L/M$ with $M = 40, 80, 160, 320$.

the parameters $\alpha = 2$ or $\alpha = 4$ and $\nu = 0.3$, 0.1 and 0.5, and the initial condition is given by

$$u_0(\boldsymbol{x}) = \frac{1}{\pi}\exp(-|\boldsymbol{x} - \boldsymbol{x}_3|^2), \quad \text{where } \boldsymbol{x}_3 = (2, 2), \tag{3.6}$$

are chosen precisely as in [26, Example 3]. We obtained numerical solutions for $M_1 = M_2 = 80$ cells at four different simulated times shown in Figure 9. The pairs $(\alpha, \nu) = (2, 0.3)$, $(4, 0.1)$, and

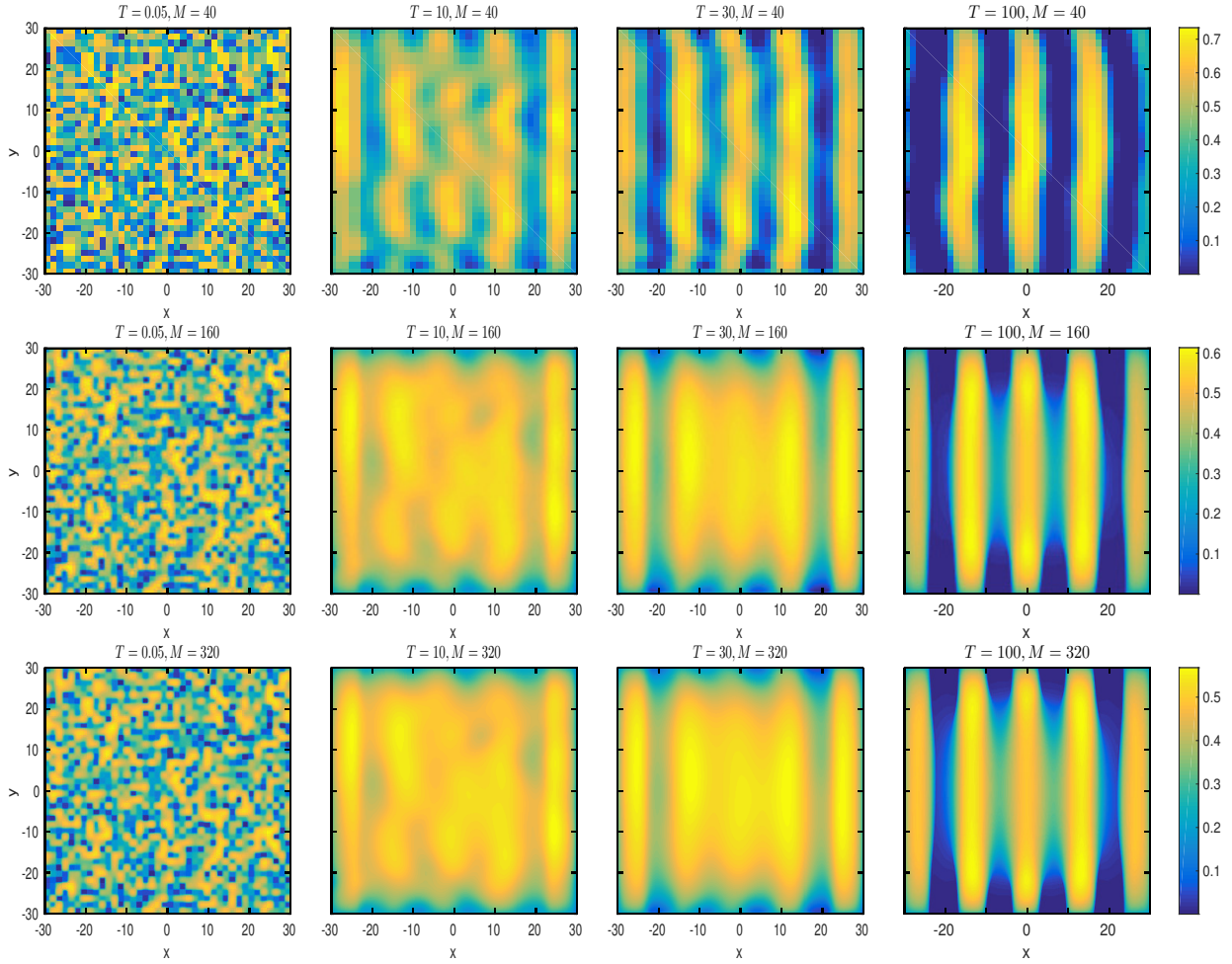

FIGURE 7. Example 4: numerical solutions with $\Delta x = 2L/M$ and $L = 30$ for (top) $M = 40$, (middle) $M = 160$ and (bottom) $M = 320$, at simulated times $T = 0.05$, 10, 30, and 100. The IMEX-RK scheme employed is IMEX-SSP2(3,3,2) given by (3.2).

$(4, 0.5)$, corresponding to the top, middle, and bottom rows of Figure 9, respectively, correspond to the scenarios for which the corresponding steady-state solution is shown in plot (e), (g), and (i) of [26, Figure 5], respectively. It is worth noting that only in the case $(\alpha, \nu) = (4, 0.5)$, due to the relative high value of ν (cf. [26, Th. 4]), the steady-state solution will be radially symmetric with mean $(0, 0)$ so the swarm will not propel into any preferential direction while in the two other cases that mean will have two equal positive components, as is stipulated by the initial condition (3.6).

4. Conclusions

Through a series of numerical examples we have reconfirmed that IMEX schemes, based on IMEX-RK time discretizations, represent a serious alternative to the explicit scheme introduced in [12] for the efficient numerical solution of (1.1). These results reconfirm those of [9] for the one-dimensional (1D) case, (where we remark that the 1D case had been studied separately since

	IMEX-RK H-CN(2,2,2)			IMEX-RK IMEX-SSP2(3,3,2)			IMEX-RK IMEX-SSP3(4,3,3)			Explicit		
M	e_M	θ_M	cpu [s]	e_M	θ_M	cpu [s]	e_M	θ_M	cpu [s]	e_M	θ_M	cpu [s]
	$T = 0.05$											
40	55627	—	1.05	55685	—	0.55	55661	—	0.54	51846	—	0.19
80	34136	0.70	0.82	34446	0.69	0.83	34462	0.69	0.72	33340	0.64	0.87
160	17462	0.97	14.3	15969	1.11	8.32	13870	1.31	6.63	12353	1.43	11.8
320	9056	0.95	83.6	3392	2.24	104	3664	1.92	75.3	2921	2.08	190
	$T = 10$											
40	55446	—	2.44	55572	—	1.97	55472	—	1.55	58363	—	3.47
80	19039	1.54	30.7	19020	1.55	27.4	19017	1.54	19.7	19283	1.60	41.3
160	7408	1.36	614	7406	1.36	466	7406	1.36	337	7539	1.35	595
320	2384	1.64	4511	2383	1.64	4974	2383	1.64	3625	2477	1.61	9817
	$T = 30$											
40	168099	—	5.23	170601	—	4.83	170061	—	3.44	184135	—	10.7
80	41992	2.00	72.0	42051	2.02	63.7	42052	2.02	45.6	42685	2.11	102
160	13931	1.59	1455	13930	1.59	968	13932	1.59	797	13983	1.61	1380
320	4191	1.73	10804	4191	1.73	11134	4191	1.73	7890	4231	1.72	20746
	$T = 60$											
40	222432	—	10.5	224421	—	10.2	224201	—	6.99	250099	—	27.7
80	101733	1.13	132	102380	1.13	118	102265	1.13	90.2	104855	1.25	209
160	32695	1.64	2451	32743	1.64	1578	32736	1.64	1335	32847	1.67	2427
320	9388	1.80	18426	9391	1.80	18536	9390	1.80	13027	9389	1.81	34878
	$T = 100$											
40	213524	—	16.9	214386	—	16.7	214315	—	11.3	220105	—	49.2
80	135826	0.65	233	136635	0.65	211	136502	0.65	164	142007	0.63	407
160	56102	1.28	3790	56290	1.28	2612	56259	1.28	2133	57156	1.31	4359
320	17536	1.68	26139	17569	1.68	27088	17563	1.68	19782	17668	1.69	55105

TABLE 4. Example 4: approximate L^1 errors (e_M, figures to be multiplied by 10^{-6}), convergence rates (θ_M), and CPU times (cpu).

in that setting one may transform a particular nonlocal of aggregation [4] into a local PDE that can be solved to obtain a reference solution). Here, we observe for Examples 1 to 4 that according to Tables 1 to 4, and with the possible exception of very small simulated times and the coarsest discretization in each case, the approximate numerical errors for a given simulated time and discretization stay very close to each other for all numerical schemes tested, and the efficiency plots (Figures 2, 4, 6, and 8) indicate that at least for finer discretizations, there is a clear gain of efficiency of the IMEX-RK schemes in comparison to the explicit version. Of course, in view of (3.3) as compared to (3.4), the gain in efficiency is likely to appear earlier (for moderately fine discretizations) whenever the diffusion term is dominant. With respect to CPU times, we remark that the maxima arising in (3.3) and (3.4) have been evaluated in each iteration; in practice one would possibly devise alternative and less cost intensive time stepping strategies (say, keep Δt fixed over larger periods).

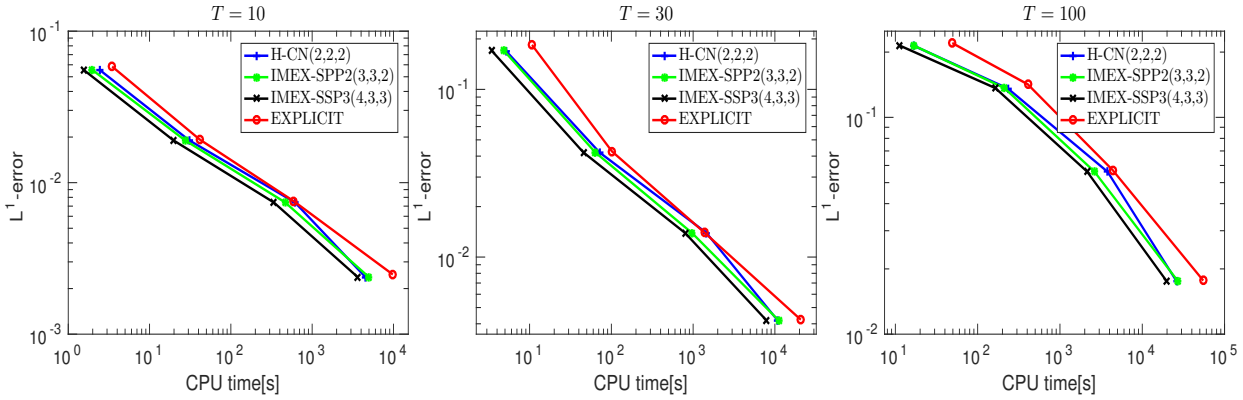

FIGURE 8. Example 4: efficiency plots based on numerical solution for $\Delta x = 2L/M$ with $M = 40, 80, 160, 320$.

ACKNOWLEDGMENTS

RB is supported by CRHIAM, project CONICYT/FONDAP/15130015 and Fondecyt project 1170473. DI acknowledges CONICYT scholarship CONICYT-PCHA/Doctorado Nacional/2014-21140362. PM is supported by Spanish MINECO project MTM2017-83942-P and Conicyt (Chile), project PAI-MEC, folio 80150006. LMV is supported by Fondecyt project 1181511. RB, DI and LMV are also supported by CONICYT/PIA/Concurso Apoyo a Centros Científicos y Tecnológicos de Excelencia con Financiamiento Basal AFB170001, and by the INRIA Associated Team "Efficient numerical schemes for non-local transport phenomena" (NOLOCO; 2018–2020).

REFERENCES

[1] U. Ascher, S. Ruuth, J. Spiteri, Implicit-explicit Runge-Kutta methods for time dependent partial differential equations, Appl. Numer. Math. 25 (1997) 151–167.
[2] A.B.T. Barbaro, J.A. Cañizo, J.A. Carrillo, P. Degond, Phase transitions in a kinetic model of Cucker-Smale type, Multiscale Model. Simul. 14 (2016) 1063–1088.
[3] D. Benedetto, E. Caglioti, M. Pulvirenti, A kinetic equation for granular media, RAIRO Modél. Math. Anal. Numér. 31 (1997) 615–641.
[4] F. Betancourt, R. Bürger, K.H. Karlsen, A strongly degenerate parabolic aggregation equation, Commun. Math. Sci. 9 (2011) 711–742.
[5] S. Boscarino, R. Bürger, P. Mulet, G. Russo, L.M. Villada, Linearly implicit IMEX Runge-Kutta methods for a class of degenerate convection-diffusion problems, SIAM J. Sci. Comput. 37 (2015) B305–B331.
[6] S. Boscarino, R. Bürger, P. Mulet, G. Russo, L.M. Villada, On linearly implicit IMEX Runge-Kutta Methods for degenerate convection-diffusion problems modelling polydisperse sedimentation, Bull. Braz. Math. Soc. (N. S.) 47 (2016) 171–185.
[7] S. Boscarino, F. Filbet, G. Russo, High order semi-implicit schemes for time dependent partial differential equations, J. Sci. Comput. 68 (2016) 975–1001.
[8] S. Boscarino, P.G. LeFloch, G. Russo, High order asymptotic-preserving methods for fully nonlinear relaxation problems, SIAM J. Sci. Comput. 36 (2014) A377–A395.
[9] R. Bürger, D. Inzunza, P. Mulet, L.M. Villada, Implicit-explicit schemes for nonlinear nonlocal equations with a gradient flow structure in one space dimension. Numer. Methods Partial Differential Equations, in press.
[10] R. Bürger, P. Mulet, L.M. Villada, Regularized nonlinear solvers for IMEX methods applied to diffusively corrected multi-species kinematic flow models, SIAM J. Sci. Comput. 35 (2013) B751–B777.
[11] M. Burger, J.A. Carrillo, M.-T. Wolfram, A mixed finite element method for nonlinear diffusion equations, Kinet. Relat. Models 3 (2010) 59–83.

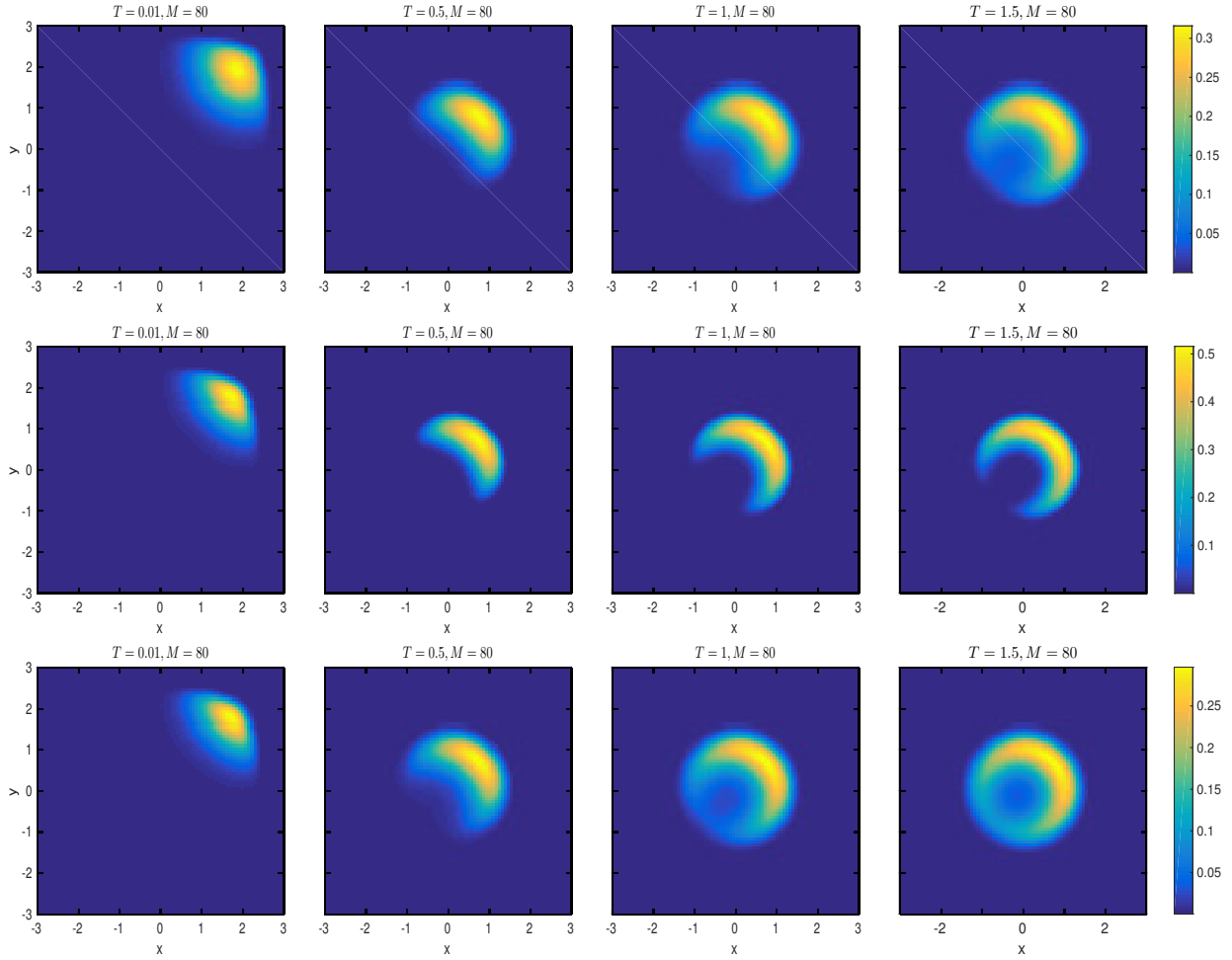

FIGURE 9. Example 5: numerical solutions with $\Delta x = 2L/M$ and $L = 3$ for $M = 80$ at simulated times $T = 0.01$, 0.5, 1, and 1.5 produced by the H-CN(2,2,2) scheme for (top) $\alpha = 2$, $\nu = 0.3$, (middle) $\alpha = 4$, $\nu = 0.1$, and (bottom) $\alpha = 4$, $\nu = 0.5$.

[12] J.A. Carrillo, A. Chertock, Y. Huang, A finite-volume method for nonlinear nonlocal equations with a gradient flow structure, Commun. Comput. Phys. 17 (2015) 233–258.
[13] J.A. Carrillo, G. Toscani, Asymptotic L^1-decay of solutions of the porous medium equation to self-similarity, Indiana Univ. Math. J. 49 (2000) 113–142.
[14] M. Crouzeix, Une méthode multipas implicite-explicite pour l'approximation des équations d'évolution paraboliques, Numer. Math. 35 (1980) 257–276.
[15] J.E. Dennis Jr., R.B. Schnabel, Numerical Methods for Unconstrained Optimization and Nonlinear Equations, Classics in Applied Mathematics vol. 16, SIAM, 1996.
[16] R. Donat, F. Guerrero, P. Mulet, Implicit-explicit methods for models for vertical equilibrium multiphase flow, Comput. Math. Appl. 68 (2014) 363–383.
[17] S. Gottlieb, C.-W. Shu, E. Tadmor, Strong stability-preserving high-order time discretization methods, SIAM Rev. 43 (2001) 89–112.
[18] K.H. Karlsen, N.H. Risebro, Convergence of finite difference schemes for viscous and inviscid conservation laws with rough coefficients, ESAIM: Math. Model. Numer. Anal. 35 (2001) 239–269.
[19] E.F. Keller, L.A. Segel, Initiation of slime mold aggregation viewed as an instability, J. Theor. Biol 26 (1970) 399–415.

[20] R.J. McCann, A convexity principle for interacting gases, Adv. Math. 128 (1997) 153–179.
[21] G. Naldi, L. Pareschi, G. Toscani, Mathematical Modeling of Collective Behavior in Socio-Economic and Life Sciences, Birkhäuser, Boston, 2010.
[22] J.M. Ortega, W.C. Rheinboldt, Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, New York-London, 1970.
[23] F. Otto, The geometry of dissipative evolution equations: the porous medium equation, Comm. Partial Diff. Eqns. 26 (2001), 101–174.
[24] L. Pareschi, G. Russo, Implicit-Explicit Runge-Kutta schemes and applications to hyperbolic systems with relaxation, J. Sci. Comput. 25 (2005) 129–155.
[25] L. Pareschi, G. Toscani, Interacting Multiagent Systems: Kinetic Equations and Monte Carlo Methods. Oxford University Press, Oxford, 2013.
[26] L. Pareschi, M. Zanella, Structure preserving schemes for nonlinear Fokker-Planck equations and applications, J. Sci. Comput. 74 (2018) 1575–1600.
[27] C.M. Topaz, A.L. Bertozzi, M.A. Lewis, A nonlocal continuum model for biological aggregation, Bull. Math. Biol. 68 (2006) 1601–1623.
[28] G. Toscani, One-dimensional kinetic models of granular flows, ESAIM: Math. Model. Numer. Anal. 34 (2000) 1277–1291.
[29] B. van Leer, Towards the ultimate conservative finite difference scheme, V. A second order sequel to Godunov's method, J. Comput. Phys. 32 (1979) 101–136.
[30] J.L. Vázquez, The Porous Medium Equation. Oxford University Press, Oxford, 2007.
[31] J. von zur Gathen, J. Gerhard, Modern Computer Algebra, Cambridge University Press, Cambridge, Second Edition, 2003.