



**FACULTY  
OF MATHEMATICS  
AND PHYSICS**  
Charles University

**DOCTORAL THESIS**

Shahin Heydari

**Development and analysis of monotone  
numerical schemes**

Department of Numerical Mathematics

Supervisor of the doctoral thesis: doc. Mgr. Petr Knobloch, Dr., DSc.

Study programme: Computational mathematics

Prague 2024

I declare that I carried out this doctoral thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date .....

Author's signature

To my lovely family

This thesis is the result of over five years studies, a time period during which I had the opportunity to receive both mathematical and personal support from many wonderful people, whom I would like to thank sincerely.

First of all, I would like to express my deepest gratitude toward my supervisor Prof. Petr Knobloch, who has provided me with countless hours in advising and encouraging me in my research. You were always kind, supportive, patient and managed to give me sufficient amount of time and space to read and work independently while at the same time always being available and open for questions and fruitful discussions. In fact, I can not even remember the numbers of my emails which have received helpful answers, hints, and advice almost immediately. I hope you know how much I appreciate your guidance, motivation, and supports during all these years; Děkuji moc Petr.

I would also like to express my thankfulness to all my colleagues starting from our kind secretary Hanka, to Prof. Zdeněk Strakoš, Prof. Václav Kučera, Prof. Vít Dolejší,  $\dots$ , all through the head of our department Prof. Miroslav Tůma for providing me with such a nice academic environment and countless support. And also to my friends and colleagues, whom I shared an office with, in particular I am thinking of Eva, Lukáš, Petr, Neil, Eda, Sunčica, and Shazma from the Department of Numerical Mathematics and Petr, Jiří, and Martin from the Division of Mathematical Modeling.

Moreover, I would like to thank Prof. Thomas Wick and his group for a very warm welcome and hospitality at Leibniz University Hannover and for numerous interesting mathematical exchange and collaborations. I am also very grateful to Mario whom I shared an office with while in there, not only for countless help, kindness, and friendly discussions but also for joint works and providing me with an excellent opportunity to get in touch with Prof. Johanness Lankeit with whom we had a very fruitful discussions and dialogues related to cross-diffusion problems which at some point developed into a nice collaboration.

Additionally, I would like to thank everyone serving on thesis committee for their time and effort.

Last and most importantly, I would like to thank my family and love ones, every single one of them, for all their love, support, understanding and encouragement, and for always being by my side, it would not have been possible to finish this work without you. You are and will always be the most important aspect of my life.

Title: Development and analysis of monotone numerical schemes

Author: Shahin Heydari

Department: Department of Numerical Mathematics

Supervisor: doc. Mgr. Petr Knobloch, Dr., DSc., Department of Numerical Mathematics

Abstract: In this thesis, we investigate various systems of strongly-coupled nonlinear partial and ordinary differential equations, which mainly originate from bio-science, both theoretically and numerically. For the main part of this work, systems of parabolic equation with cross-diffusion is considered. It is well-known that, the systems of these types usually suffer from low regularity due to the nature of the cross-diffusion term(s). Lack of regularity may also be caused due to the structure of the other equations present in the system. We address these difficulties and establish the existence of global classical solutions for different cross-diffusion systems. Next, we show that the analytical investigation may get very difficult or simply fail to solve or capture the behavior of the solutions of the considered systems and it is necessary to approximate the respective solutions by means of numerical methods. We show that the behavior of numerical solutions heavily depends on the effect of the cross-diffusion term(s), i.e., when these terms are dominant the standard numerical methods become unstable and the approximate solutions are usually polluted by spurious oscillations. We present high-resolution nonlinear finite element flux-corrected transport (FE-FCT) methods to overcome this problem. Then, we analyze the proposed schemes and address their solvability, positivity, and satisfaction of discrete maximum principle. The theoretical and numerical results are validated by several numerical experiments in various spatial dimensions.

In the last part of this work, we investigate the qualitative and quantitative behavior of a strongly-coupled nonlinear system of ordinary differential equations. We employ a nonstandard finite difference scheme to approximate the solutions of the system under consideration and address the questions of positivity, elementary stability and conservation.

Keywords: cross diffusion, existence of solutions, FE-FCT stabilization methods, positivity preservation

# Contents

<b>Introduction</b>	<b>2</b>
0.1 What is a cross-diffusion system? . . . . .	2
0.2 Thesis outline . . . . .	5
<b>1 On the stabilization of convection-diffusion-reaction equations</b>	<b>8</b>
1.1 Stabilization of steady-state convection- diffusion-reaction equations . . . . .	9
1.2 Stabilization of transient convection- diffusion-reaction equations . . . . .	29
1.3 Application of stabilized methods to cross-diffusion systems . . . .	42
<b>2 Paper I</b>	<b>43</b>
2.1 Global existence of classical solutions and numerical simulations of a cancer invasion model . . . . .	43
<b>3 Papers II and III</b>	<b>72</b>
3.1 Flux-corrected transport stabilization of an evolutionary cross- diffusion cancer invasion model . . . . .	72
3.2 Solvability and numerical solution of a cross-diffusion cancer invasion model . . . . .	93
<b>4 Paper IV</b>	<b>104</b>
4.1 A cross-diffusion system modeling rivaling gangs: global existence of bounded solutions and FCT stabilization for numerical simulation . . . . .	104
<b>5 Paper V</b>	<b>147</b>
5.1 Introduction: Nonstandard finite difference scheme . . . . .	147
5.2 A positive and elementary stable nonstandard explicit scheme for a mathematical model of the influenza disease . . . . .	148
<b>Conclusion</b>	<b>163</b>
<b>Bibliography</b>	<b>166</b>
<b>List of publications</b>	<b>186</b>

# Introduction

## 0.1 What is a cross-diffusion system?

Convection, diffusion, and reaction are physical processes that play an important role in the modeling of many real-life situations. A convection-diffusion-reaction (CDR) equation may simply describe, a substance which undergoes diffusion and spreads out (randomly or/and uniformly) from a higher concentration location, moves in a certain direction due to convection coefficient, or interacts by affecting other particles or influencing each other's dispersal pattern. The situation is more complex for systems of equations, where one considers not only a simple diffuser or random directional mover, but also preferential directional motion of particles, then the aforementioned processes are not able to describe the desirable phenomena under consideration. In these cases, motility of the species is not determined solely by their own characteristic in question but different species are considered to be mutually interfering with each other, in other words the movement does not only depend on the density of  $i^{th}$  species but also on the density of  $j^{th}$  species. The question of how to interpret these motions in the mathematical framework was first addressed in [1] and also in the experiments of [2], where it was suggested that a cross-diffusion term needs to be taken into account in order to describe such a movement. The term cross-diffusion refers to a phenomenon where the gradient of one concentration causes the flux of another concentration in the system. The introduction of cross-diffusion term allows the mathematical models to capture much more features of many phenomena in physics, biology, chemistry, ecology, or engineering sciences. For instance, the Shigesada-Kawasaki-Teramoto (SKT) model [3] in the population dynamics is a special case of a cross-diffusion system which was proposed to investigate a segregation phenomenon of two species that are competing with each other in the same habitat area. A motional type of interaction was considered in [4], which described the movement of predator toward prey and of prey away from predator. The effect of cross-diffusion greatly emerges in the mathematical modeling of cancer invasions and their treatment see, e.g, [5, 6] and [7, 8], respectively. A chemotaxis type cross-diffusion system known as Keller-Segel model was proposed in [9, 10], where cell movement toward or away from a chemical source was investigated. The modeling of population dynamics [11, 12], electrochemistry [13], cell-sorting [14], pattern formation of bacteria [15], tumor invasion [16, 17] are among many other models utilizing cross-diffusion effects.

Many strongly coupled cross-diffusion systems can take the form

$$\partial_t u = \operatorname{div}(A(u)\nabla u) + f(u) \quad \text{in } \Omega \times (0, T], \quad (1)$$

subject to boundary and initial conditions

$$\begin{aligned} (A(u)\nabla u) \cdot \nu &= 0 \quad \text{on } \partial\Omega \times (0, T], \\ u(\cdot, 0) &= u^0 \quad \text{in } \Omega, \end{aligned}$$

where  $\Omega \subset \mathbb{R}^d$  ( $d \geq 1$ ) is an open bounded domain,  $u(x, t) \in \mathbb{R}^n$  is a vector-valued function representing, e.g., population densities or chemical concentration and  $\partial_t u$

is its time derivative,  $A(u) \in \mathbb{R}^{n \times n}$  is the diffusion matrix,  $f(u) \in \mathbb{R}^n$  denotes the reactions and external forces, and  $\nu$  is the outward unit normal vector to  $\partial\Omega$ . Usually, whenever  $u$  models concentration it is expected that these concentration be non-negative, for the cases where  $u$  models mass fraction boundedness and positivity are expected. As mentioned in the preceding, cross-diffusion systems appear in many area of science, however the presence of the cross-diffusion term often complicates analytical and numerical analysis, for which we will give a very brief background in the following.

## 1. Some analytical background

It is known that, if the diffusion matrix  $A(u)$  is a (positive definite) diagonal matrix, then the corresponding equations are quite regular and the global well-posedness can directly follow from applying energy methods. Assuming this condition and that the components are continuous and uniformly bounded with respect to  $u$ , a global weak existence result is proved for a system of the type (1) in [18]. However, this assumption usually does not hold, i.e.,  $A(u)$  is neither symmetric nor positive definite and it is usually non-diagonal as well, which leads to lower regularity in the system in such a way that even questions of local-in-time existence are already quite delicate, thus making such a system very challenging to handle. For the SKT system, an important and well-studied case of (1), the existence of solution is proved for both non-degenerate and degenerate case in [19, 20, 21] and [22, 23], respectively. For this system equipped with a nonlinear reactive term the global existence result is proved in [24]. Sufficient assumptions for the global existence of weak or strong solutions of various nonlinear parabolic equations can be found in, e.g., [25, 26, 27]. A classical approach which provides a very powerful tool for obtaining an uniform estimate of the solutions that can yield the global in time existence is based on entropy structure of the cross-diffusion system. This method is usually used whenever the maximum principle<sup>1</sup> or parabolic regularity theory cannot be applied, and it works by employing a transformation of variables whenever the cross-diffusion system under consideration possesses an entropy. This transformation results in a positive semi-definite diffusion matrix, gradient estimates, and upper and lower bounds of solutions. Moreover, this also leads to suitable a priori estimates which are key step to prove global solvability. This approach has been first introduced in [28], later were employed to analyze several classes of cross-diffusion systems, see, e.g., [28] for a population system in one dimension and [11, 19] in several dimensions, [29] where the global in time existence of bounded weak solutions is proved for a class of strongly-coupled parabolic system, and [30] where the entropy variables allowed  $L^\infty$  bounds without using maximum principle. For more detailed background information regarding this technique we refer the reader to [27]. It is worth noting that there are very few results on uniqueness and this is an open question for many cross-diffusion systems. This motivated the researchers to explore other approaches which can be used in this regard. As a result, a technique was introduced which makes it possible to enhance the regularity of the system

---

<sup>1</sup>The validity of the maximum principle is of great importance since it can play an important role in analyzing the existence, uniqueness, and positivity of solutions to partial differential equations.



without any assumptions about the entropy structure of the cross-diffusion systems or in general whenever the identification of the entropy structure of the system is difficult, in addition, it also does not rely on any assumptions regarding the structure of the diffusion matrix  $A(u)$ . This approach allows to prove the well-posedness of the problem at hand by controlling the ratio of diffusion and cross-diffusion components of diffusion matrix  $A(u)$ . For an example of this approach see [31], where a global existence result of non-negative solution is obtained by applying Schauder's strategy coupled with Meyer regularity result for a system of type (1). Moreover, the question of boundedness of the solutions is also addressed and a weak maximum principle is proved. A very interesting result has been proposed recently in [32], in which the maximum principle for a larger class of cross-diffusion systems has been investigated. It was proved that, in contrast to the conventional believe that the cross-diffusion systems only enjoy maximum principle whenever diffusion matrix  $A(u)$  is diagonal, a new result on the maximum principle was reported considering that  $A(u)$  can also be non-diagonal (provided that  $f(u) > 0$ ). The key idea was to employ a matrix  $B$  to transform the diffusion matrix  $A(u)$  to a lower or upper triangular matrix, where as a result it makes it possible to establish some maximum principles for wider range of cross-diffusion systems, see [32] for more detail. On the other hand, a question that global solutions might not exist at all has drawn much attention in the mathematical analysis of cross-diffusion systems over the past few decades so that many researchers have accepted the low regularity of the systems under consideration and showed that this effect may lead to blow-up of the solution, see, e.g., [33, 34, 35, 36].

## 2. Some numerical background

Compared to the huge amount of the analytical results regarding the cross-diffusion systems obtained over the past few decades, the numerical methods and their respective analysis for this class of problems are far from well-studied. Hence, constructing an appropriate numerical scheme to solve such a problem is of great importance to give predictions for the future and to validate the assumptions that the model is based on, especially when the analytical investigations appear to be very difficult or even impossible in some cases. Various methods can be applied in this regard, very common among them are: finite difference method, finite volume method, and finite element method. An explicit finite difference scheme was examined in [37] for a mechanical model of tumor growth, the considered system was given by a multiphase flow model where the velocity was a regularization of the classical Darcy law. In [38], an efficient nonstandard finite difference approximation was proposed for a strongly coupled system describing cancer migration and invasion. It was shown that the proposed method guarantees the positivity of the numerical solutions for arbitrary mesh size and has explicit and fast performance even with nonlinear reaction terms. A nonlinear and linear discrete-time algorithms for cross-diffusion system were considered in [39], where convergence rates for a time-discretized scheme was derived. A standard two-point finite-volume flux in combination with a nonlinear positivity-preserving approximation of the cross-diffusion coefficients was used for a reaction-diffusion system with cross-diffusion in [40]. The existence and uniqueness of the approx-

imate solution were investigated and a stability analysis for a model of pattern-formation was addressed. An unconditionally positivity-preserving linear finite volume scheme for a class of Keller-Segel systems was addressed in [41], where an upwind technique was employed. A modification of the finite volume method for a class of chemotaxis system was proposed in [42]. For more examples on two-point flux approximation finite volume methods for cross-diffusion systems see [27]. A fully discrete implicit finite element approximation with a regularization technique for the SKT model was considered in [43], where convergence results in several space dimensions were proved. Finite element methods for a class of chemotaxis-driven PDE systems were studied in [44], where different stabilization techniques were used and the most reliable and efficient solvers were investigated. There are many other contributions on the application of different numerical schemes to cross-diffusion systems which can be found in the literature.

As mentioned before, these systems usually contain strongly coupled nonlinear equations, therefore the solution of one equation can severely influence the results of the other equations in the system, which might lead to, e.g., nonphysical (non-positive) numerical solutions, violation of the total mass conservation law, or undershoots and overshoots in the numerical simulations. Hence, the main difficulties in the design of a suitable numerical scheme is to preserve as many essential properties of the considered cross-diffusion system as possible. Most existing standard numerical schemes usually fail to satisfy these properties especially whenever the cross-diffusion term(s) appears to be much stronger in comparison to the other terms in the system, which is usually the case in many applications. Main feature of the solutions in these cases is the appearance of layers which are narrow regions where large gradients of the solutions are present, and standard numerical methods usually lead to heavily oscillating solutions unless these layers are resolved by means of suitable meshes. Therefore, the development of adaptive techniques which are usually based on a priori or a posteriori error estimations to adapt the mesh appropriately in these layers has been the center of attention for many researchers. For instance, an adaptive finite-element/level-set method is employed for a model of tissue invasion in [45], see also [46] and [47] for more examples. However, choosing a suitable mesh resolution is not always feasible and requires high amount of memory and/or CPU expenses, and take a lot of computational time. Therefore, developing appropriate numerical methods which are able to provide sufficiently accurate results even on coarse meshes comparing to the width of the layers is of great importance. In this regard, stabilization methods can be a great choice even though the design of a proper stabilization method may be quite challenging. The application of stabilization methods to cross-diffusion systems has been addressed in several publications, see, e.g., [48, 49, 50, 51, 52]. However, these studies mainly focus on the Keller-Segel system and moreover, their analysis is not available.

## 0.2 Thesis outline

Let us close this introductory part by sketching out the outline of this thesis:

- To begin with, as mentioned above, the behavior of the approximate solu-

tions of cross-diffusion systems are highly under the influence of the cross-diffusion term(s). For comparably large magnitude of this term in the system, i.e, when the cross-diffusion term is dominant, standard numerical systems may become unstable or even lead to blow-up in the numerical simulations, hence they need to be stabilized. This process resembles the convection-dominated regime in the CDR equations in the computational fluid dynamics, for which an extensive amount of stabilization methods has been introduced over the last four decades. Hence, in Chapter 1, we shall recall some of these methods and their improvements with more emphasize on the high-resolution nonlinear finite element flux-correction transport (FE-FCT) method, which will be our method of use.

- Next, a cross-diffusion system of chemotaxis-type modeling the invasion of healthy tissues by cancer cells is considered in Chapter 2. We address the low-regularity of the system which is the result of the cross-diffusion term and also the structure of the other equations in the system. We establish the existence of global classical solutions in two- and three-dimensional bounded domains utilizing the parabolic regularity theory. Then, the numerical stability of the system is investigated by manipulating the respective parameters in the system, it is shown that in the cross-diffusion dominated regime the standard Galerkin method combined with  $\theta$ -scheme gives rise to spurious oscillations and numerical blow-up in the system. Furthermore the theoretical results are supported by numerical simulations in two- and three- dimensions.
- In Chapter 3, we consider the haptotaxis counterpart of the model considered in Chapter 2, for which the techniques presented in the previous chapter is no longer applicable and its solvability is an open problem from the analytical point of view. Though, we address this point by means of the numerical methods, in this regard a high-resolution nonlinear FE-FCT method is employed for space discretization combined with an implicit  $\theta$ -method for time discretization and fixed-point iterations to deal with nonlinearities. Then, making use of Brouwer's fixed point theorem, it is proved that both the nonlinear scheme and the linearized problems used in the fixed-point iteration are solvable. Moreover we prove that they are positivity-preserving. The results are supported with numerical test in two dimensions.
- In Chapter 4, we consider a double cross-diffusion system modeling gang rivaling interactions. The key feature of this problem is that the cross-diffusion term is not only present in one but in two equations in the system which poses even more challenges from analytical and numerical point of view compared to their single cross-diffusion counterparts. We establish the global bounded classical solutions and prove that these solution converge toward homogeneous steady-state for sufficiently small initial data, however, for large data it appears to be very difficult to obtain such results analytically. We utilize FE-FCT scheme once again and prove its positivity preservation, moreover, we investigate the validity of the discrete maximum principle. Making use of the numerical experiments in one- and

two-dimensions, we address not only the asymptotic behavior of large-time solutions but also illustrate the evolution of the gang densities throughout the time.

- So far, we only addressed systems of partial differential equations. In Chapter 5, we consider a system of strongly coupled nonlinear ordinary differential equations describing the influenza disease. We present a nonstandard finite difference scheme and prove that it is positivity-preserving and also elementary-stable. The results are supported by providing some numerical experiments.

# 1. On the stabilization of convection-diffusion-reaction equations

Transport processes play an important role in the distribution of physical quantities in many applications describing simple to very complicated and complex problems, and can be modeled in a simplest form by CDR equations. Since obtaining the solutions of such problems analytically is usually too complicated in many situations, it is necessary to approximate the respective unknown solutions by means of numerical methods at discrete level. However, if the unknown solutions represent concentrations or densities then numerical methods which produce negative solutions and violate discrete maximum principle (DMP) by creating spurious oscillations (over/under-shoots) are not usually useful in practice. On the other hand, preserving the qualitative properties of the method is also of great importance, and hence during the discretization the respective mesh has to be chosen appropriately. Last but not least, the cost of a numerical schemes is of great interest in applications. Therefore, it is desirable that the constructed numerical scheme not only preserves the physical properties of the continuous model in the discrete solutions but also it is highly accurate (with respect to appropriate meshes), robust, and computationally reasonable.

The numerical solution of CDR equations is notoriously difficult if the convection is larger by the order of magnitude in comparison to diffusion or reaction. This usually gives rise to wild oscillations in certain areas (moving fronts, interior and boundary layers), where the quantity of interest changes abruptly at reasonable grid size. In finite difference methods framework, if central-differences are used to approximate the convective term, the obtained solutions are usually polluted by spurious oscillations. However, it was observed that the use of upwind differencing leads to oscillation free solutions but also to a loss of accuracy since these methods are typically only first-order accurate, which results in overly diffuse solutions. It was stated that the classical upwind differences, which are applied only to convection term, produce very poor results in the presence of source terms. Later it was shown that a combination of central and upwind methods works rather better than upwind or central-difference alone in one-dimensional case, but the extension of such a methods to higher dimensions are much more difficult and complicated in practice [53, 54, 55, 56, 57]. If finite element methods are used to solve such problems, the use of the standard Galerkin finite element method should be avoided unless the mesh is severely refined in the problematic regions. It was observed that the use of the Galerkin finite element method is roughly equivalent to central-difference approximation which also inherits similar oscillatory properties in the approximate solutions [58]. The finite element equivalent of upwinding was first presented in the late 1970s in [59], by introducing a Petrov-Galerkin formulation in one-dimension. The basis of this method was to use shape functions that are heavier in the upstream of a node considering the direction of the flow, this approach opposes the Galerkin method in which

test and shape functions are usually selected from the same function space and distributed evenly to each node of an element. Two-dimensional upwind finite element discretization were presented in [60, 61, 62], however, generalization of these methods to higher dimensions seems to introduce diffusion in the crosswind direction. In addition, when applied to more complicated problems, some of these methods were far from adequate. These difficulties have motivated researchers to construct finite element formulations for the CDR equations that are stable, accurate, optimal and applicable to a wide variety of problems. In the following, we will give a brief review of the developments of some of these methods during the past four decades, started from the 1980s through today. These methods are called stabilized finite element methods.

Let us introduce some standard notation which will be used throughout this work: Let  $\Omega \subset \mathbb{R}^d$ ,  $d \geq 1$  be a bounded domain, then  $L^2(\Omega)$  denotes the Lebesgue space,  $W^{k,p}(\Omega)$  is Sobolev spaces which is the space of functions whose distributional derivatives up to order  $k$  are belong to  $L^p(\Omega)$ . The norm (semi-norm) on  $W^{k,p}(\Omega)$  is denoted by  $\|\cdot\|_{k,p,\Omega}$  ( $|\cdot|_{k,p,\Omega}$ ). Note that the spaces  $W^{k,2}(\Omega)$  are Hilbert spaces with convention  $W^{k,2}(\Omega) = H^k(\Omega)$  and  $\|\cdot\|_{k,\Omega} = \|\cdot\|_{k,2,\Omega}$  (similarly  $|\cdot|_{k,\Omega} = |\cdot|_{k,2,\Omega}$ ).

## 1.1 Stabilization of steady-state convection-diffusion-reaction equations

Let  $\Omega \subset \mathbb{R}^d$ ,  $d \in \{1, 2, 3\}$  be a polygonal or polyhedral bounded domain with a Lipschitz-continuous boundary  $\partial\Omega$ . A linear scalar steady-state CDR equation has the form

$$-\varepsilon \Delta u + \mathbf{b} \cdot \nabla u + cu = g \quad \text{in } \Omega, \quad (1.1)$$

$$u = u_b \quad \text{on } \partial\Omega, \quad (1.2)$$

where  $\varepsilon > 0$  is a constant diffusion (or viscosity) coefficient,  $\mathbf{b} \in W^{1,\infty}(\Omega)^d$  is a solenoidal (i.e.,  $\nabla \cdot \mathbf{b} = 0$ ) convection coefficient (or velocity field),  $c \in L^\infty(\Omega)$  is a reaction coefficient,  $g \in L^2(\Omega)$  is describing sources or sink terms,  $u_b \in H^{\frac{1}{2}}(\partial\Omega) \cap C(\partial\Omega)$  is a prescribed boundary condition, and  $u$  is the unknown function. For simplicity of the presentation, Dirichlet boundary condition is considered on the whole boundary  $\partial\Omega$ . In addition, we shall also assume that the following assumption holds

$$\sigma := c - \frac{1}{2} \operatorname{div} \mathbf{b} \geq 0. \quad (1.3)$$

Let  $\tilde{u}_b \in H^1(\Omega)$  be an extension of the boundary condition  $u_b$ . Then, the variational formulation corresponding to (1.1)-(1.2) reads: Find  $u \in H^1(\Omega)$  such that  $u - \tilde{u}_b \in V := H_0^1(\Omega)$  and

$$a(u, v) = (g, v) \quad \forall v \in V, \quad (1.4)$$

$$a(u, v) = \varepsilon (\nabla u, \nabla v) + (\mathbf{b} \cdot \nabla u, v) + (cu, v), \quad (1.5)$$

where  $(\cdot, \cdot)$  indicates the inner product in  $L^2(\Omega)$  or  $L^2(\Omega)^d$ , and  $H_0^1(\Omega)$  consists of functions from  $H^1(\Omega)$  with zero trace on the boundary.

The standard Galerkin finite element discretization is constructed based on the variational formulation (1.4) by replacing the function spaces with finite element subspaces and approximating  $\tilde{u}_b$  by a finite element interpolant  $\tilde{u}_{bh}$ . Let  $\mathcal{T}_h$  be a triangulation of  $\bar{\Omega}$  that belongs to a regular family of triangulations (see [63]) and consists of open set elements. Then we introduce finite element spaces

$$W_h = \left\{ v_h \in C(\bar{\Omega}); v_h|_T \in R(T), \forall T \in \mathcal{T}_h \right\}, \quad V_h = W_h \cap H_0^1(\Omega),$$

consisting of continuous piecewise (multi-)linear functions, and  $R(T) = P_1(T)$  for triangular elements and  $R(T) = Q_1(T)$  for rectangular elements. The usual basis functions  $\phi_1, \dots, \phi_N$  of  $W_h$  are defined by  $\phi_i(x_j) = \delta_{ij}$ ,  $i, j = 1, 2, \dots, N$ , where  $\delta_{ij}$  is the Kronecker symbol and  $x_i, i = 1, \dots, N$  denote the vertices of the triangulation  $\mathcal{T}_h$  such that  $x_1, \dots, x_M \in \Omega$  and  $x_{M+1}, \dots, x_N \in \partial\Omega$ . Clearly, the functions  $\phi_1, \dots, \phi_M$  form a basis for  $V_h$ .

Any function  $u_h \in W_h$  can be written uniquely in the form

$$u_h = \sum_{i=1}^N u_i \phi_i \tag{1.6}$$

and identified with the coefficient vector  $U = (u_1, \dots, u_N)$ .

Now, let  $V_h \subset V$ , then the standard Galerkin method reads: Find  $u_h \in W_h$  such that  $u_h - \tilde{u}_{bh} \in V_h$  and

$$a(u_h, v_h) = (g, v_h) \quad \forall v_h \in V_h. \tag{1.7}$$

As mentioned above, this method leads to solutions that are globally polluted with spurious oscillations in the convection-dominated regime ( $\varepsilon \ll |\mathbf{b}|h$ ) and need to be stabilized. The stabilized finite element methods can be formulated as: Find  $u_h \in W_h$  such that  $u_h - \tilde{u}_{bh} \in V_h$  and

$$a_h(u_h, v_h) = (g, v_h) \quad \forall v_h \in V_h, \tag{1.8}$$

$$a_h(u_h, v_h) = a(u_h, v_h) + S(u_h, v_h), \tag{1.9}$$

where  $S(u_h, v_h)$  indicates the additional terms added to the Galerkin finite element discretization. These terms are added in such a way that the stability is enhanced, consistency is preserved, and accuracy is improved.

## 1. Streamline upwind/Petrov-Galerkin method

Streamline upwind/Petrov-Galerkin (SUPG) method also known as streamline diffusion finite element method (SDFEM) is one of the most popular linear stabilization finite element methods which was originally introduced in [64, 65, 66]. The basic idea behind this method is to add a streamline upwind perturbation acting only in the flow direction to the standard Galerkin formulation (1.7),

which manifests itself as a stabilization term  $S(u_h, v_h)$  in (1.9) in the form

$$S_{SUPG}(u_h, v_h) := \sum_{T \in \mathcal{T}_h} \left( -\varepsilon \Delta u_h + \mathbf{b} \cdot \nabla u_h + c u_h - g, \tau \mathbf{b} \cdot \nabla v_h \right)_T, \quad (1.10)$$

where  $\mathcal{T}_h$  is the same triangulation used for defining the finite element space  $V_h$  as before,  $T$  denotes an arbitrary element of the triangulation,  $\tau$  denotes the non-negative stabilization parameter, and  $(\cdot, \cdot)_T$  denotes integration over  $T$  in  $L^2(T)$  or  $L^2(T)^d$ . This consistent Petrov-Galerkin residual-based stabilization (RBS) method introduces artificial diffusion along streamlines which maintains stability and improves accuracy away from the areas where sharp layers exist.

Now, it remains to define the stabilization parameter  $\tau$ . Although it was noted in [66] that the structure of the stabilization term is far more important than the value of the parameter  $\tau$ , it was shown later in many studies that the appropriate choice of the stabilization parameter is also of great importance, since it determines the amount of the artificial diffusion to be added by the SUPG method to the Galerkin discretization. The stabilization parameter has to be chosen in such a way that it is large enough to suppress the spurious oscillations but also at the same time it is small enough to prevent smearing of the layers. Comparing to the finite difference stencils, a first stability parameter which was limited to the linear interpolation was suggested in [66]. Originated from one-dimensional case, the stabilization parameter on any element  $T \in \mathcal{T}_h$  of linear or bilinear finite elements can be defined by the formula

$$\tau|_T := \frac{h_T}{2|\mathbf{b}|} \left( \coth Pe_T - \frac{1}{Pe_T} \right) \quad \text{with} \quad Pe_T = \frac{|\mathbf{b}| h_T}{2\varepsilon}, \quad (1.11)$$

where  $h_T$  is an approximation of the length of the mesh  $T$  in the direction of the convection vector  $\mathbf{b}$  and  $Pe_T$  is the local mesh cell Péclet number. It was suggested in [67, 68] to set the stabilization parameter  $\tau$  on each element  $T \in \mathcal{T}_h$  as

$$\tau|_T := \frac{\text{diam}(T)}{2|\mathbf{b}|} \left( \coth (Pe_T/2) - \frac{1}{Pe_T/2} \right) \quad \text{with} \quad Pe_T = \frac{|\mathbf{b}| \text{diam}(T)}{2\varepsilon}, \quad (1.12)$$

where  $\text{diam}(T) = \sup\{|x - y|; x, y \in T\}$  denotes the diameter of  $T$ . Another possibility is to set

$$\tau|_T = \begin{cases} \tau_0 h_T & \text{if } Pe_T > 1 \\ \tau_1 h_T^2 / \varepsilon & \text{if } Pe_T \leq 1, \end{cases} \quad (1.13)$$

which was suggested in [56].

Studies in [69, 70] revealed that the choice of the stabilization parameter at interior and boundary layers has negligible influences on reducing the spurious oscillations and the best way to deal with this difficulty is to refine the mesh properly in these regions. Later in [71], the author showed that even though controlling the inadmissible oscillations in the characteristic layers seems to be difficult, however, it is possible to introduce an appropriate stabilization parameter which is able to reduce the oscillation along the outflow boundary layers ( $\partial\Omega^+ = \{x \in \partial\Omega : (\mathbf{b} \cdot \mathbf{n})(x) > 0\}$ ) to a great extent. Thus, it was suggested to keep the stabilization parameter defined in (1.11) as it is away from the outflow



boundary layers, and define  $\tau$  on any other element which intersects with the outflow boundary layers as

$$\tau|_T := \tau_0|_T \left( \coth Pe_T - \frac{1}{Pe_T} \right) \quad \text{with} \quad Pe_T = \frac{|\mathbf{b}_T| h_T}{2 \varepsilon}, \quad (1.14)$$

where  $\mathbf{b}_T$  is the mean value of  $\mathbf{b}$  in  $T$ , and  $\tau_0$  is a positive piecewise constant function satisfying

$$\int_{G_h} \phi_i + \tau_0 \mathbf{b} \cdot \nabla \phi_i dx = 0, \quad i = 1, \dots, M,$$

where  $G_h$  consists of all elements intersecting  $\partial\Omega^+$ . Another possibility to design a proper stabilization parameter was proposed in [72, 73], which relies on a posteriori computation of this parameter and leads to considerable reduction of unwanted nonphysical oscillations in the vicinity of the sharp layers. The analysis regarding several possible choices of the stabilization parameters that can be used in practice were discussed in details in many publications, e.g., [74, 75, 56], see also [76, 77, 78] and references there in for more details and examples.

The stability and error estimate for SUPG method can be achieved with respect to the norm

$$\|v\|_{SUPG} := \left( \varepsilon \|v\|_{1,\Omega}^2 + \|\sigma^{1/2} v\|_{0,\Omega}^2 + \|\tau^{1/2} \mathbf{b} \cdot \nabla v\|_{0,\Omega}^2 \right)^{1/2},$$

see e.g., [79, 80].

A generalization of this method has been applied to various other problems, e.g., coupled multi-dimensional advective-diffusive system [81], Stokes and Navier-Stokes problems [82, 83], compressible flow problems [84, 85], and first order linear hyperbolic system [79]. The SUPG method was further enhanced by incorporating shock capturing operator [86, 87, 77, 88].

The SUPG method does not contain any spurious crosswind diffusion and performs significantly well on many problems in comparison to the Galerkin method, but it is not a monotone method and does not satisfy the DMP, thus it usually yields some oscillations around sharp layers.

## 2. Galerkin/least-squares method

The next popular residual-based stabilization scheme known as Galerkin/least-squares (GLS) method was developed in [89, 90], which as its name suggests adds a least-squares term to the Galerkin method and can be formulated in (1.9) as:

$$S_{GLS}(u_h, v_h) := \sum_{T \in \mathcal{T}_h} \left( -\varepsilon \Delta u_h + \mathbf{b} \cdot \nabla u_h + c u_h - g, \tau (-\varepsilon \Delta v_h + \mathbf{b} \cdot \nabla v_h + c v_h) \right)_T. \quad (1.15)$$

This term is capable of enhancing the stability of the Galerkin discretization without degrading accuracy. Defining a proper choice of the stabilization parameter  $\tau$  can be done similarly as for the SUPG method. Convergence analysis in [89]

indicated that the algorithmic parameter  $\tau$  must behave as

$$\tau|_T = \begin{cases} C \frac{h_T^2}{|\mathbf{b}|} & \text{when } Pe_T \text{ is small,} \\ C' \frac{\bar{\varepsilon}}{|\mathbf{b}|} & \text{when } Pe_T \text{ is large,} \end{cases} \quad (1.16)$$

on each element  $T$ , where  $C$  and  $C'$  are positive constants independent of the mesh size and Péclet number. It was also determined from this analysis when the expression for  $\tau$  changes from one case to the other. The condition (1.16) holds if

$$\tau|_T = \frac{\alpha h_T}{2|\mathbf{b}|}, \quad \alpha = \min\{C_1 Pe, C_2\},$$

where the constants  $C_1$  and  $C_2$  are related to the constant appearing in the interpolation error of the finite element approximation used and also to inverse estimates, see [89] and also [91]. There is also the possibility of using the same parameter  $\tau$  as in the SUPG method for the GLS method.

The GLS method represents a generalization of the SUPG method with additional discontinuity capturing feature and it is applicable to a wide variety of model problems, [92, 93, 94, 95, 96, 97, 98, 99]. The stability and error estimate for GLS method can be achieved with respect to the norm

$$\|v\|_{GLS} := \left( \varepsilon |v|_{1,\Omega}^2 + \|\sigma^{1/2} v\|_{0,\Omega}^2 + \|\tau^{1/2} (-\varepsilon \Delta v + \mathbf{b} \cdot \nabla v + cv)\|_{0,\Omega}^2 \right)^{1/2}.$$

A detailed convergence analysis of the scalar steady-state convection-diffusion equation is presented in [89]. The error analysis and existence results of the GLS method was further investigated in [100] for CDR and Navier-Stokes problems.

Galerkin/gradient-least-squares (GGLS) method was introduced in [101] for the cases that require a strong control over the solution gradients with the stabilization term

$$S_{GGLS}(u_h, v_h) := \sum_{T \in \mathcal{T}_T} \left( \nabla(-\varepsilon \Delta u_h + \mathbf{b} \cdot \nabla u_h + cu_h - g), \tau \nabla(-\varepsilon \Delta v_h + \mathbf{b} \cdot \nabla v_h + cv_h) \right)_T, \quad (1.17)$$

which represents the gradient of a least-square term. In order to have stronger control over the gradient, a combination of these two methods (GLS and GGLS) was introduced in [102].

Similar to the SUPG method, the GLS method is a Petrov-Galerkin method and it is consistent and easily applicable to many problems. However, it is known that the GLS method often presents good stability and accuracy properties if the exact solution is smooth enough, otherwise, the spurious oscillations can still remain in the vicinity of sharp layers.

### 3. Unusual stabilized FE and bubble enriched Galerkin methods

Stabilized finite element methods often enhance the stability and preserve the good accuracy property of the Galerkin method by adding mesh-dependent terms which are evaluated on each element of the triangulation. It was shown that the stability can be achieved also by mesh-dependent stabilization term composed by an adjoint term from the Galerkin discretization as:

$$S_{USFEM}(u_h, v_h) := \sum_{T \in \mathcal{T}_h} \left( -\varepsilon \Delta u_h + \mathbf{b} \cdot \nabla u_h + c u_h - f, \tau (-\varepsilon \Delta v_h - \mathbf{b} \cdot \nabla v_h + c v_h) \right)_T, \quad (1.18)$$

these methods are called unusual stabilized finite element methods (USFEM) which were first introduced in [103, 94] and further developed in [104, 105].

Adding appropriate terms to the Galerkin variational formulation is a well-accepted practice which performs greatly for convection-dominated equations. However, it was found that revisiting Galerkin method using richer subspaces rather than piecewise linear polynomials can also guarantee stability and higher accuracy without the need to manipulate the Galerkin variational form. In other words, the idea is to enlarge the finite element space using bubble functions defined elementwise and then eliminating them using static condensation. Such bubbles can be seen as the addition of stabilization terms, therefore choosing the appropriate (shape and number of) bubble functions is of great importance, in other words, the selection of an optimal stabilization parameter is practically translate into the problem of finding an optimal bubble space.

For brevity we assume that  $c$  is constant and homogeneous Dirichlet boundary condition on  $\partial\Omega$ . Let  $V_h^b := \{v \in H_0^1(\Omega); v|_T \in R(T) \oplus B(T)\}$ , where  $R(T)$  is defined as before and  $B(T)$  denotes the space of bubble functions spanned by the bubble basis function  $\phi_T \in B(T)$  such that:

$$\begin{aligned} \phi_T(x) &> 0, \quad \forall x \in T, \\ \phi_T(x) &= 0, \quad \forall x \in \partial T, \end{aligned} \quad (1.19)$$

and  $\phi_T = 1$  at the barycenter of the triangle (rectangle). Then, the standard Galerkin method enriched with bubble functions is formulated as : Find  $u_h \in V_h^b$  such that

$$a(u_h, v_h) = (g, v_h), \quad \forall v_h \in V_h^b, \quad (1.20)$$

where  $u_h$  is the unknown solution consisting of a linear part  $u_1 \in V_h$  and its part spanned by bubbles, i.e.,

$$u_h = u_1 + \sum_{T \in \mathcal{T}_h} u_b^T \phi_T, \quad (1.21)$$

and  $u_b^T$  is the unknown bubble coefficient. Now, we wish to understand the effect of bubble functions on the linear part of the solution  $u_1$ , for this reason we utilize the static condensation which consists of first taking  $v_h = \phi_T$  on  $T$  (and  $v_h = 0$  elsewhere in  $\Omega$ ) and also substituting (1.21) in (1.20) which leads to:

$$\begin{aligned} (c u_1, \phi_T)_T + (\mathbf{b} \cdot \nabla u_1, \phi_T)_T + (\varepsilon \nabla u_1, \nabla \phi_T)_T \\ + u_b^T (c \phi_T, \phi_T)_T + u_b^T (\mathbf{b} \cdot \nabla \phi_T, \phi_T)_T + u_b^T (\varepsilon \nabla \phi_T, \nabla \phi_T)_T = (g, \phi_T)_T. \end{aligned} \quad (1.22)$$

Solving (1.22) in each element for the bubble coefficient  $u_b^T$  we get:

$$u_b^T = \frac{-1}{c \|\phi_T\|_{0,T}^2 + \varepsilon \|\nabla \phi_T\|_{0,T}^2} (c u_1 + \mathbf{b} \cdot \nabla u_1 - \varepsilon \Delta u_1 - g, \phi_T)_T. \quad (1.23)$$

The second part of the static condensation is to set  $v_h = v_1 \in V_h$  in (1.20) as

$$\begin{aligned} & (c u_1, v_1) + (\mathbf{b} \cdot \nabla u_1, v_1) + (\varepsilon \nabla u_1, \nabla v_1) \\ & + \sum_{T \in \mathcal{T}_h} u_b^T (\phi_T, c v_1)_T - \sum_{T \in \mathcal{T}_h} u_b^T (\phi_T, \mathbf{b} \cdot \nabla v_1)_T - \sum_{T \in \mathcal{T}_h} u_b^T (\phi_T, \varepsilon \Delta v_1)_T = (g, v_1). \end{aligned} \quad (1.24)$$

Substituting the expression for  $u_b^T$  we get:

$$a(u_1, v_1) - \sum_{T \in \mathcal{T}_h} \frac{(c u_1 + \mathbf{b} \cdot \nabla u_1 - \varepsilon \Delta u_1 - g, \phi_T)}{c \|\phi_T\|_{0,T}^2 + \varepsilon \|\nabla \phi_T\|_{0,T}^2} (\phi_T, c v_1 - \mathbf{b} \cdot \nabla v_1 - \varepsilon \Delta v_1) = (g, v_1), \quad (1.25)$$

$\forall v_1 \in V_h$ , bypassing the definition of the bubble shape function  $\phi$ , (1.25) can be simplified as

$$a(u_1, v_1) - \sum_{T \in \mathcal{T}_h} \left( c u_1 + \mathbf{b} \cdot \nabla u_1 - \varepsilon \Delta u_1 - g, \tau (c v_1 - \mathbf{b} \cdot \nabla v_1 - \varepsilon \Delta v_1) \right)_T = (g, v_1), \quad (1.26)$$

where the stability parameter  $\tau$  is given by

$$\tau|_T = \frac{\left( \int_T \phi_T dx \right)^2}{|T| \left[ c \|\phi_T\|_{0,T}^2 + \varepsilon \|\nabla \phi_T\|_{0,T}^2 \right]}.$$

This is the USFEM method mentioned above which is similar to the GLS method. It was shown that this method can reduce to SUPG method see, [106, 105] and references therein for more details. Moreover, the relationship between stabilized methods and Galerkin method enriched with bubble functions is also studied in [107].

As noted in the preceding, choosing bubble functions of correct shape and number is of great importance, taking inappropriate bubbles can lead to Galerkin method which behaves like a stabilized method with a poor selection of the stability parameter. Therefore, another possibility to design special bubbles was introduced by means of residual-free approach where the stabilizing mechanism is considered in the enrichment of the space. In the residual-free bubble method the optimal parameter is determined through the solution of a suitable boundary value problem in each element. This method was suggested and investigated in various studies, see, e.g., [108, 109, 110, 111, 112, 113].

#### 4. Local projection stabilization method

In comparison to the residual-based stabilization methods mentioned above, where the stability was mainly induced into the Galerkin method through stabilization terms that control the derivative of approximate solution in large scales, another scheme known as local projection stabilization (LPS) method was introduced to enforce the stability using only a fluctuation of the derivative of

the approximate solution in small scales. This method was originally proposed for Stokes problems in [114], later extended to stabilization of CDR problems [115, 116, 117], and incompressible flow problems [118, 119, 120, 121].

The idea of the LPS method mainly relies on a projection  $\Pi_h : X_h \rightarrow D_h$  of a finite-dimensional space  $X_h$  into a discontinuous space  $D_h$ , where a stabilization term is added to the Galerkin discretization in such a way that it gives  $L^2$ -control over the fluctuation  $\kappa_h := id - \Pi_h$  of the gradient of the approximate solution. Proving the error estimate and stability results of the LPS scheme is in need of a proper construction of an interpolation operator in the approximation space  $X_h$  that exhibits an orthogonality property with respect to the projection space  $D_h$ , see [122, 123, 124, 125]. It has been proven in [126, 82, 125] that such an interpolation operator exists if both approximation space  $X_h$  and projection space  $D_h$  satisfy local inf-sup conditions. Additionally, the appropriate choice of  $D_h$  is of great importance, on one hand  $D_h$  has to be large enough to satisfy some approximating properties on the other hand it should be small enough to guarantee the inf-sup condition.

It is known that there are two variants of the LPS method: a two-level approach and a one-level approach. LPS method was first introduced as a two-level approach [114, 119, 127, 125, 128], in which the projection space  $D_h \subset L^2(\Omega)$  lies on a coarser grid  $\mathcal{M}_h \subset \Omega$ . This coarse grid  $\mathcal{M}_h$  is constructed by utilizing a basic finer mesh  $\mathcal{T}_h$ , and each of its macro-elements  $M \in \mathcal{M}_h$  is a gathering of neighboring cells  $T \in \mathcal{T}_h$ . Let  $D_h(M) := \left\{ q_h|_M; q_h \in D_h \right\}$  and  $\Pi_M : L^2(M) \rightarrow D_h(M)$  denote the local  $L^2$ -projection on each element  $M \in \mathcal{M}_h$ , which defines the global projection  $\Pi_h : L^2(\Omega) \rightarrow D_h$  by  $(\Pi_h v)|_M := \Pi_M \left( v|_M \right)$ . Furthermore, set  $\kappa_M := id - \Pi_M$  ( $id : L^2(M) \rightarrow L^2(M)$  is the identity operator) to be the so-called fluctuation operator. Then, providing that the local inf-sup condition holds, i.e., there exists a positive constant  $\beta$  independent of  $h$  such that  $\forall M \in \mathcal{M}_h$  :

$$\inf_{q_h \in D_h(M)} \sup_{v_h \in V_h(M)} \frac{(v_h, q_h)_M}{\|v_h\|_{0,M} \|q_h\|_{0,M}} \geq \beta, \quad (1.27)$$

where  $V_h(M) := \{v_h \in V_h; v_h = 0 \text{ in } \bar{\Omega} \setminus M\}$ , the local projection stabilization term  $S(u_h, v_h)$  can be defined as

$$S_{LPS}(u_h, v_h) := \sum_{M \in \mathcal{M}_h} \tau_M \left( \kappa_M (\mathbf{b}_M \cdot \nabla u_h), \kappa_M (\mathbf{b}_M \cdot \nabla v_h) \right)_M, \quad \forall v_h \in V_h(M), \quad (1.28)$$

where  $\tau_M$  is a non-negative stabilization parameter on each macro-element  $M$  which can be defined analogously as for the SUPG method, and  $\mathbf{b}_M$  is the value of  $\mathbf{b}$  in some point inside each  $M$ .

The one-level approach was introduced later in [115, 125, 117, 115] in such a way that the finite-dimensional space  $V_h$  and the projection space  $D_h$  are defined on the same mesh (i.e, set  $\mathcal{M}_h = \mathcal{T}_h$ ). In this case, the approximation space is enriched with bubble functions. Following the same strategy as before : Let  $D_h(T) := \left\{ q_h|_T; q_h \in D_h \right\}$ , and  $\Pi_h : L^2(T) \rightarrow D_h(T)$  be a local projection which

defines the global projection by  $(\Pi_h v)|_T := \Pi_h \left( v|_T \right)$  and  $\kappa_h := id - \Pi_h$  be the fluctuation operator. In addition,  $\forall T \in \mathcal{T}_h$  there is a  $\beta > 0$  independent of  $h$  such that the local inf-sup condition holds:

$$\inf_{q_h \in \mathcal{D}_h(T)} \sup_{v_h \in V_h(T)} \frac{(v_h, q_h)_T}{\|v_h\|_{0,T} \|q_h\|_{0,T}} \geq \beta, \quad (1.29)$$

where  $V_h(T) := \{v_h \in V_h; v_h = 0 \text{ in } \bar{\Omega} \setminus T\}$ , the local projection stabilization term  $S(u_h, v_h)$  reads as:

$$S_{LPS}(u_h, v_h) := \sum_{T \in \mathcal{T}_h} \tau_T \left( \kappa_T (\mathbf{b}_T \cdot \nabla u_h), \kappa_T (\mathbf{b}_T \cdot \nabla v_h) \right)_T, \quad \forall v_h \in V_h(T). \quad (1.30)$$

where  $\tau_T$  is a non-negative stabilization parameter on each element  $T$ .

The LPS method is equipped with the norm

$$\|v\|_{LPS} := \left( \varepsilon |v|_{1,\Omega}^2 + \|\sigma^{1/2} v\|_{0,\Omega}^2 + S_{LPS}(v, v) \right)^{1/2}. \quad (1.31)$$

It was shown in [124] that the LPS method is as stable as the SUPG method in the sense of an inf-sup condition.

A comparison of these two variants, a detailed analysis of computed parameters  $\tau_M$  and  $\tau_T$  based on a priori estimates, and the relation between LPS method and RBS techniques, can be found in [115]. It was mentioned in [115], that there is a close relation between the LPS method and sub-grid modeling [129]. Moreover, these methods can be interpreted as a special class of variational multiscale (VMS) methods [130, 131, 132].

We would like to note that the LPS method has several advantages over the RBS methods, namely: they form a symmetric stabilization term, do not contain second order derivatives, and do not lead to additional coupling between various terms when applied to a system of partial differential equations. However, they have a few drawbacks, first, both variant require more degrees of freedom than residual-based methods, this difficulty has been overcome using overlapping macro-elements in [123, 133], additionally, they are not able to completely remove the wild oscillations in the vicinity of the sharp layers as well.

## 5. Continuous interior penalty and/or edge stabilization method

Next in line to enhance the stability of the Galerkin discretization is continuous interior penalty (CIP) or/and edge (face) stabilization method. The basic idea behind this method consist of adding an stabilization term to (1.8) to control the gradient jumps across element boundaries instead of inside each elements. This stabilization term can be defined as:

$$S_{CIP}(u_h, v_h) := \sum_{E \in \mathcal{E}_h} \tau h_E^2 \left( [\mathbf{b} \cdot \nabla u_h], [\mathbf{b} \cdot \nabla v_h] \right)_E, \quad (1.32)$$

where  $\mathcal{E}_h$  denotes the set of all interior edges (faces),  $h_E$  is the edge length (face measure),  $\tau$  is the stabilization parameter, and  $[\cdot]_E$  denotes the jump of a function across an edge  $E$ . Another possible choice of  $S(u_h, v_h)$  is

$$S_{CIP}(u_h, v_h) := \sum_{E \in \mathcal{E}_h} \tau h_E^2 \left( [\nabla u_h], [\nabla v_h] \right)_E. \quad (1.33)$$

The first attempt to construct such methods was proposed in [134] for CDR equations for which later the inf-sup stability, convergence property and monotonicity of discrete solution was investigated in [135]. The edge stabilization has been successfully applied to the generalized Stokes problem in [136], it was noted however that different results might be observed if the stabilization parameter scales with respect to the mesh size  $h$ . The CIP method was presented and compared with other types of stabilization methods (residual-based / projection-based) in [118] for the Oseen problem. In [137], the authors provided a generalization of the interior penalty method following the work [134] to a model problem of viscoelastic flow and proved an optimal a priori estimate. An extension of CIP for Oseen's equations using equal order interpolation for pressure and velocity was presented in [138]. A CIP  $hp$ -finite element method was introduced and analyzed for advection and advection-diffusion equation in [139].

Compared with other stabilization methods mentioned above, the formulations (1.32) and (1.33) possess a few advantages: they do not require the computation of second-order derivatives, the formulation remains symmetric, no hierarchical meshes are needed. However, the stiffness matrix is denser due to the connection between degrees of freedom on neighboring cells.

## 6. Mizukami-Hughes method

Although the aforementioned methods perform very well on many problems, most of these methods do not satisfy the DMP and are not able to remove the oscillations in the vicinity of sharp layers completely. Mizukami and Hughes suggested that choosing an appropriate upwind direction can improve the possibility of constructing a new method which is able to overcome this difficulty [140]. They proposed one of the first monotone methods for solving (1.1)-(1.2) with  $c = 0$  for linear triangular finite elements which, not only satisfies the DMP but also provides very accurate results, since it does not lead to smearing of the layers.

Assuming that, in addition to the assumptions made at the beginning of this section, the triangulation  $\mathcal{T}_h$  is of weakly acute type, i.e., the magnitude of all angles of elements  $T \in \mathcal{T}_h$  is less than or equal to  $\frac{\pi}{2}$ , then the idea of the Mizukami-Hughes method is to replace the test functions  $\phi_i$  by functions  $\bar{\phi}_i$  defined by

$$\bar{\phi}_i = \phi_i + \sum_{\substack{T \in \mathcal{T}_h, \\ x_i \in T}} C_i^T \chi_T, \quad i = 1, \dots, M, \quad (1.34)$$

where  $C_i^T$ 's are constants and  $\chi_T$  is the characteristic function of  $T$ . Then, choos-

ing the constants  $C_i^T$  for any  $T \in \mathcal{T}_h$  in such a way that

$$C_i^T \geq -\frac{1}{3}, \quad \forall i \in \{1, \dots, N\}, \quad x_i \in \bar{T}, \quad \sum_{\substack{i=1, \\ x_i \in \bar{T}}}^N C_i^T = 0, \quad (1.35)$$

and the local convection matrices are of non-negative type<sup>1</sup>, these fulfill the conditions to prove the DMP (since the triangulation is of weakly acute type, all local diffusion matrices are of non-negative type, thus it is only suffices to prove that convection matrices are of non-negative type). Let  $x_1$ ,  $x_2$ , and  $x_3$  be the vertices of any element  $T$  of the triangulation  $\mathcal{T}_h$  such that for each vertex  $x_i$  it is possible to define a vertex zone and an edge zone [140]. If the convection vector  $\mathbf{b}$  points into a vertex zone then (1.35) holds for

$$C_1^T = \frac{2}{3}, \quad C_2^T = C_3^T = -\frac{1}{3},$$

and the local convection matrices are of non-negative type, however, when  $\mathbf{b}$  points toward an edge zone it is not possible to find appropriate  $C_i^T$ 's which satisfy the above properties. The authors in [140], suggested to change the orientation of the convection vector to deal with this problem, in which  $\mathbf{b}$  is replaced by any function  $\tilde{\mathbf{b}}$  in such a way that  $\tilde{\mathbf{b}} - \mathbf{b}$  is orthogonal to  $\nabla u$ , which is always possible. The definition of  $C_i^T$ 's strongly relies on the orientation of  $\mathbf{b}$  to guarantee correct solutions, for example, it was shown that for  $\mathbf{b}$  pointing to an edge zone if  $T$  lies in the boundary layers the definition of  $C_i^T$ 's are not appropriate and lead to incorrect solutions, hence in [141, 142], the author presented several improvements to overcome such difficulties. Moreover, he showed that it is possible to extend the prescribed method to CDR equations and to the three-dimensional cases. Since the original Mizukami-Hughes method and its improved version are only defined for conforming triangular finite elements, the method was later extended to bilinear quadrilateral finite elements in [143], it was shown that in this case the properties of the method depends on the definitions of four constants on each element of the triangulation and also proved that the method fulfills the DMP. Another improvement of the Mizukami-Hughes method was presented in [144], the accuracy of the discrete solution was improved for different values of the diffusion coefficient or/and whenever the convection is not extremely large, i.e., for small and moderate Péclet number.

The Mizukami-Hughes method possesses many nice properties: it is a monotone method that always satisfies DMP which ensures oscillation-free solution even in the vicinity of sharp layers, it is of upwind type though does not contain any stabilization parameters, it is a Petrov-Galerkin method therefore it is consistent, it is highly accurate in comparison to many other upwind methods or stabilized methods, the construction of the method is simple and clear. However, the method possesses some drawbacks as well: it depends on the discrete solutions and hence it is nonlinear which may cause some difficulties when highly accurate solutions are desired, generalization of the method to more complicated problems is rather difficult, the method may lead to incorrect solution in some cases, and there is no existence, uniqueness and convergence results available for the method.

---

<sup>1</sup>All off-diagonal entries are non-positive, the row sums are non-negative



We would like to note that except Mizukami-Hughes method, most of the upwind approaches introduce too much artificial diffusion which, usually lead to the smearing of the layers and results in poor accuracy see, e.g., [62, 141, 145, 140, 146, 147].

## 7. Spurious oscillations at layers diminishing methods

As mentioned above, the SUPG method

$$a(u_h, v_h) + (R_h(u_h), \tau \mathbf{b} \cdot \nabla v_h) = (g, v_h), \quad \forall v_h \in V_h, \quad (1.36)$$

with the residual  $R_h(u_h) = -\varepsilon \Delta u_h + \mathbf{b} \cdot \nabla u_h + c u_h - g$ , is known to be one of the best and most used stabilization methods in reducing the spurious oscillation arising in the Galerkin solutions, however, since it is not a monotone method it is not able to completely remove or considerably reduce the over- and undershoots in problematic areas, precisely speaking, in small regions where the derivatives of the solution are extremely large. Removing these unwanted oscillations has been the subject of an extensive research over the last few decades, as a result, it was revealed that adding a suitable amount of the artificial diffusion to the left-hand side of (1.36), can lead to a discretization which fulfills the DMP in most model cases, such a procedure is called spurious oscillations at layers diminishing (SOLD) method. The resulting methods are nonlinear since the amount of the artificial diffusion in these methods depends on the unknown discrete solution  $u_h$ . In the following we briefly recall different types of these methods which were developed during the last two decades to eliminate the oscillations in the discrete solutions of the problem (1.1)-(1.2).

- The methods of the first type tend to add isotropic artificial diffusion terms to the left-hand side of the SUPG discretization (1.36) as

$$(\tilde{\varepsilon} \nabla u_h, \nabla v_h), \quad (1.37)$$

which was originally introduced in [86] and later extended in [148, 149]. One of the best choices of  $\tilde{\varepsilon}$  is to set

$$\tilde{\varepsilon} = \max \left\{ 0, \frac{\tau |\mathbf{b}| |R_h(u_h)|}{|\nabla u_h|} - \tau \frac{|R_h(u_h)|^2}{|\nabla u_h|^2} \right\},$$

which was presented in [150]. It was suggested in [151] that the stabilization parameter can be defined as

$$\tilde{\varepsilon}|_T = \max \left\{ 0, \alpha [\text{diam}(T)]^\nu |R_h(u_h)| - \varepsilon \right\}, \quad \forall T \in \mathcal{T}_h,$$

with some constants  $\alpha$  and  $\nu$ . Other possible definitions of  $\tilde{\varepsilon}$  has been suggested in [152, 153, 154]. In [155], an adaptive technique was used along with this type of SOLD method to increase the convergence order and improve the results obtained with the SUPG method. In addition, a priori and a posteriori error estimates for these methods were also investigated in [151, 156].

- Next type is to add an artificial diffusion orthogonal to the streamline direction (i.e., in the crosswind direction) to the (1.36) as

$$(\tilde{\varepsilon} D \nabla u_h, \nabla v_h), \quad (1.38)$$

where  $D$  is the projection onto the line or plane defined by

$$D = \begin{cases} I - \frac{\mathbf{b} \otimes \mathbf{b}}{|\mathbf{b}|^2} & \text{if } \mathbf{b} \neq 0, \\ 0 & \text{if } \mathbf{b} = 0, \end{cases}$$

with  $I$  being the identity tensor. Note that, in two-dimensional case the SOLD term (1.38) can be written in the form

$$\left( \tilde{\varepsilon} \mathbf{b}^\perp \cdot \nabla u_h, \mathbf{b}^\perp \cdot \nabla v_h \right) \quad \text{with} \quad \mathbf{b}^\perp = \frac{-(b_2, b_1)}{|\mathbf{b}|}, \quad (1.39)$$

see [157] for more detail. One of the best choices of  $\tilde{\varepsilon}$  was introduced in [78] following the works of [158] as

$$\tilde{\varepsilon}|_T = \max \left\{ 0, \eta \frac{\text{diam}(T) |R_h(u_h)|}{2 |\nabla u_h|} - \varepsilon \right\},$$

where  $\text{diam}(T)$  is the diameter of  $T$  and  $\eta$  is a suitable constant. In [157] it was suggested to define  $\tilde{\varepsilon}$  by

$$\tilde{\varepsilon}|_T = \max \left\{ 0, |\mathbf{b}| h_T^{3/2} - \varepsilon \right\} \quad \forall T \in \mathcal{T}_h,$$

see also [159, 160, 161, 162] for other variants of  $\tilde{\varepsilon}$ .

- Last type of the SOLD methods is based on edge-stabilization, adding the following term to the formula (1.36)

$$\sum_{T \in \mathcal{T}_h} \int_{\partial T} \tilde{\varepsilon}|_T \text{sign} \left( \frac{\partial u_h}{\partial \mathbf{t}_{\partial T}} \right) \left( \frac{\partial v_h}{\partial \mathbf{t}_{\partial T}} \right) d\sigma, \quad (1.40)$$

where  $\mathbf{t}_{\partial T}$  is a tangent vector to the boundary  $\partial T$  of  $T$ . For these types of methods it was suggested to set  $\tilde{\varepsilon}$  on each  $T \in \mathcal{T}_h$ , as

$$\tilde{\varepsilon}|_T = C \varepsilon |T| \left| R_h(u_h) \Big|_T \right| \quad \forall T \in \mathcal{T}_h,$$

where  $C$  is a non-negative constant [135].

In all aforementioned types of SOLD methods, the parameter  $\tilde{\varepsilon}$  is to be set zero whenever the denominator defining  $\tilde{\varepsilon}$  vanishes. For a great collection, classification, comparison, and numerical experiments of most of the SOLD methods which has been constructed to completely remove or at least considerably reduce the undesirable oscillations without deteriorating the accuracy, we refer to two great reviews [78, 163], and also [71, 77] and references therein.

Similarly to the SUPG method, the LPS scheme is also unable to remove the oscillations completely and some of them still remain along sharp layers. Therefore, in [78], the LPS method was combined with a SOLD term as

$$\sum_{M \in \mathcal{M}_h} \left( \tilde{\varepsilon}|_M \kappa_M (D_M \nabla u_h), \kappa_M (D_M \nabla v_h) \right), \quad (1.41)$$

with

$$\tilde{\varepsilon}|_M = \eta h_M |\mathbf{b}_M| |\kappa_M (D_M \nabla u_h)|, \quad (1.42)$$

and

$$\tilde{\varepsilon}|_M = \eta h_M |\mathbf{b}_M| \frac{h_M^{d/2} |\kappa_M (D_M \nabla u_h)|}{|u_h|_{1,M}}, \quad (1.43)$$

where  $M \in \mathcal{M}_h$  is a macro-element,  $h_M$  is the diameter of  $M$ ,  $\eta > 0$  is a user-chosen constant parameter, and  $D_M : \mathbb{R}^d \rightarrow \mathbb{R}^d$  is the projection onto the line or plane orthogonal to the vector  $\mathbf{b}_M$ . As proposed in [164], it is also possible to add a nonlinear term  $(\tilde{\varepsilon} D \nabla u_h, \nabla v_h)$  of SOLD type to the CIP method.

Although, SOLD methods are considered to be monotone or monotonicity preserving in some model cases and are capable of significantly improve the SUPG solutions, the main difficulty in the application of these approaches is that they contain two user-chosen parameters. Hence, despite a huge amount of literature on these methods, it is still impossible to choose a method with an appropriate parameter which works perfectly in all test cases.

## 8. Flux corrected transport method

Going through a long pass of research to discover a method that guarantees the DMP and as a result computes solutions without spurious oscillations, even in the vicinity of sharp layers, another class of stabilized finite element methods has been introduced which often satisfies the DMP by construction (under certain assumptions on the meshes). Methods of this type are called finite element flux-corrected transport (FE-FCT) methods or algebraic flux correction (AFC) methods and were first introduced in [165, 166], following the pioneer works of [167, 168] in the formulation of the flux correction, and since then they have been intensively developed and improved in many publications, see, e.g., [169, 170, 171, 172, 173, 174, 175]. Unlike previous stabilization techniques, which were mainly based on the modification of a variational formulation of the problem, the AFC stabilization methods are performed on the algebraic level. In other words, they manipulate the matrix and the right hand side of the algebraic system of equations.

The first step towards the design of an AFC scheme is to rewrite the Galerkin formulation (1.7) as:

$$\sum_{j=1}^N a_{ij} u_j = g_i, \quad i = 1, \dots, M, \quad (1.44)$$

$$u_i = u_i^b, \quad i = M + 1, \dots, N, \quad (1.45)$$

where  $a_{ij} = a(\phi_j, \phi_i)$ ,  $i, j = 1, \dots, N$ ,  $g_i = (g, \phi_i)$ ,  $i = 1, \dots, M$  and  $u_i^b = u_b(x_i)$ ,  $i = M + 1, \dots, N$ . Then, modify the algebraic system (1.44) in such a way that the necessary conditions for fulfilling the DMP holds and the layers are not excessively smeared. Let  $\mathbb{A} = (a_{ij})_{i,j=1}^N$  denote the stiffness matrix corresponding to the above-mentioned discretization and  $U = (u_1, \dots, u_N)$  be a solution vector consisting of unknown coefficient  $u_i$ ,  $i = 1, \dots, N$ . Then, the satisfaction of the local DMP is guaranteed if and only if the matrix of the linear discretization above is a matrix of non-negative type with positive diagonal entries, i.e., if and only if

$$a_{ii} > 0, \quad \forall i = 1, \dots, M, \quad (1.46)$$

$$a_{ij} \leq 0, \quad \forall i \neq j, \quad i = 1, \dots, M, \quad j = 1, \dots, N, \quad (1.47)$$

$$\sum_{j=1}^N a_{ij} \geq 0, \quad i = 1, \dots, M. \quad (1.48)$$

If, in addition, the matrix of (1.44)-(1.45) is non-singular, then the global DMP is satisfied as well. However, it is known that for the above discretization in the convection-dominated regime the validity of the conditions above is usually violated and as a results the DMP does not hold. As a remedy, to enforce the DMP, a sufficient amount of artificial diffusion has to be added to (1.44). Let  $\mathbb{D} = (d_{ij})_{i,j=1}^N$  denote a symmetric artificial diffusion matrix with entries

$$d_{ij} = d_{ji} = -\max\{a_{ij}, 0, a_{ji}\}, \quad \forall i \neq j, \quad d_{ii} = -\sum_{j \neq i} d_{ij}, \quad (1.49)$$

then, the matrix  $\tilde{\mathbb{A}} = \mathbb{A} + \mathbb{D}$  has positive diagonal and non-positive off-diagonal entries, and in addition  $\sum_{j=1}^N a_{ij} \geq 0$ ,  $i = 1, \dots, M$  (provided that  $c \geq 0$ ) holds i.e., the necessary conditions for satisfying the DMP are met, see [176]. Now, the system (1.44)-(1.45) can be replaced by:

$$(\tilde{\mathbb{A}}U)_i = g_i, \quad i = 1, \dots, M, \quad (1.50)$$

$$u_i = u_i^b, \quad i = M + 1, \dots, N, \quad (1.51)$$

which is a monotone low-order method. This modified system is a manifestation of a simple artificial diffusion method and results in an excessive smearing of the layers due to the large amount of added artificial diffusion. The smearing can be prevented by restricting the artificial diffusion to the areas where the solutions change abruptly, while the DMP is still respected.

The symmetric matrix  $\mathbb{D}$  is a discrete diffusion operator with zero row and column sums, hence it follows that

$$(\mathbb{D}U)_i := \sum_{i \neq j} f_{ij}, \quad i = 1, \dots, N,$$

with fluxes  $f_{ij} = d_{ij}(u_j - u_i)$ . Obviously,  $f_{ij} = -f_{ji}$  for all  $i, j = 1, \dots, N$ , meaning that the amount of flux added to node  $i$  is subtracted from the node  $j$  and vice-versa. Finally, for the last step of building an AFC scheme, the fluxes

have to be limited appropriately, in other words, so-called solution-dependent flux limiters  $\alpha_{ij} \in [0, 1]$  are introduced in such a way that the DMP is maintained and artificial diffusion is mainly localized around extrema and layer areas without compromising the accuracy. To this end, system (1.50)-(1.51) is replaced by

$$\left(\tilde{\mathbb{A}}U\right)_i = g_i + \sum_{i \neq j} \alpha_{ij} f_{ij}, \quad i = 1, \dots, M, \quad (1.52)$$

$$u_i = u_i^b, \quad i = M + 1, \dots, N, \quad (1.53)$$

or equivalently

$$\sum_{j=1}^N a_{ij} u_j + \sum_{j=1}^N (1 - \alpha_{ij}) d_{ij} (u_j - u_i) = g_i, \quad i = 1, \dots, M, \quad (1.54)$$

$$u_i = u_i^b, \quad i = M + 1, \dots, N. \quad (1.55)$$

This nonlinear scheme makes it possible to easily switch between high- and low-order techniques depending on the smoothness of the solution, for which, clearly the Galerkin finite element method is recovered in the smooth regions and the low-order scheme is used in the vicinity of sharp layers. To maintain the conservation property of the resulting scheme, it is usually assumed that the coefficients  $\alpha_{ij}(u_1, \dots, u_N)$  be symmetric, i.e,

$$\alpha_{ij} = \alpha_{ji}, \quad i, j = 1, \dots, N. \quad (1.56)$$

Let

$$b_{ij} = (1 - \alpha_{ij}) d_{ij}, \quad \forall i \neq j, \quad b_{ii} = - \sum_{i \neq j} b_{ij}, \quad (1.57)$$

then, the system (1.54)-(1.55) can be written in the form

$$\sum_{j=1}^N (a_{ij} + b_{ij}) u_j = g_i, \quad i = 1, \dots, M, \quad (1.58)$$

$$u_i = u_i^b, \quad i = M + 1, \dots, N. \quad (1.59)$$

then, the system (1.58)-(1.59) can be rewritten in the variational form, where the algebraic term is represented by

$$b_h(w; z, v) = \sum_{i,j=1}^N b_{ij}(w) z(x_j) v(x_i), \quad w, z, v \in C(\bar{\Omega}),$$

with  $b_{ij}(w) = b_{ij}(\{w(x_i)\}_{i=1}^N)$ . Then, the stability and error estimate for the AFC scheme can be derived with respect to the solution-dependent norm

$$\|v\|_{AFC} := \left( \varepsilon |v|_{1,\Omega}^2 + \sigma \|v\|_{0,\Omega}^2 + b_h(u_h; v, v) \right)^{1/2}, \quad v \in V_h,$$

see [177] for details.

Over the years, many different variant of the correction limiter  $\alpha_{ij}$  have been suggested in the literature, however, their definition mostly relies on the fluxes  $f_{ij}$ . In the following we shall recall some of these limiters.

### 1. The Kuzmin limiter

This limiter was originally developed in [172].

1. To begin with, one first computes

$$P_i^+ = \sum_{\substack{j=1, \\ a_{ji} \leq a_{ij}}}^N f_{ij}^+, \quad P_i^- = \sum_{\substack{j=1, \\ a_{ji} \leq a_{ij}}}^N f_{ij}^-, \quad Q_i^+ = -\sum_{j=1}^N f_{ij}^-, \quad Q_i^- = -\sum_{j=1}^N f_{ij}^+,$$

for  $i = 1, \dots, M$ , where  $f_{ij} = d_{ij}(u_j - u_i)$ ,  $f_{ij}^+ = \max\{0, f_{ij}\}$ , and  $f_{ij}^- = \min\{0, f_{ij}\}$ .

2. Next, one defines

$$R_i^+ = \min\left\{1, \frac{Q_i^+}{P_i^+}\right\}, \quad R_i^- = \min\left\{1, \frac{Q_i^-}{P_i^-}\right\}, \quad i = 1, \dots, M.$$

If  $P_i^+$  or  $P_i^-$  vanishes, set  $R_i^+ = 1$  or  $R_i^- = 1$ , accordingly. At the Dirichlet nodes, also define  $R_i^+ = R_i^- = 1, i = M + 1, \dots, N$ .

3. Finally, for any  $i, j \in \{1, \dots, N\}$  such that  $a_{ij} \geq a_{ji}$ , the limiter  $\alpha_{ij}$  is defined by

$$\alpha_{ij} = \begin{cases} R_i^+ & \text{if } f_{ij} > 0, \\ 1 & \text{if } f_{ij} = 0, \\ R_i^- & \text{if } f_{ij} < 0. \end{cases}$$

This is an upwind-type limiter which is applicable to  $\mathbb{P}_1$  and  $\mathbb{Q}_1$  elements. The Kuzmin limiter was analyzed thoroughly in [178] for steady-state CDR equations. The existence of a solution of the nonlinear problem and the satisfaction of the DMP was proven under certain assumptions on the mesh, for  $\mathbb{P}_1$  elements.

## 2. The BJK limiter

This limiter was originally derived in [179] for  $\mathbb{P}_1$  elements. The first step is to define

$$u_i^{\max} = \max_{j \in S_i \cup \{i\}} u_j, \quad u_i^{\min} = \min_{j \in S_i \cup \{i\}} u_j, \quad q_i = \gamma_i \sum_{j \in S_i} d_{ij}, \quad i = 1, \dots, M,$$

where  $S_i = \{j \in \{1, \dots, N\} \setminus \{i\} : a_{ij} \neq 0 \text{ or } a_{ji} > 0\}$  and  $\gamma_i$  is a positive constant, see [179] for more details. As next step:

1. Compute for  $i = 1, \dots, M$

$$\begin{aligned} P_i^+ &= \sum_{j \in S_i} f_{ij}^+, & P_i^- &= \sum_{j \in S_i} f_{ij}^-, \\ Q_i^+ &= q_i (u_i - u_i^{\max}), & Q_i^- &= q_i (u_i - u_i^{\min}), \end{aligned} \quad (1.60)$$

2. Then, compute

$$R_i^+ = \min\left\{1, \frac{Q_i^+}{P_i^+}\right\}, \quad R_i^- = \min\left\{1, \frac{Q_i^-}{P_i^-}\right\}, \quad i = 1, \dots, M.$$

If  $P_i^+$  or  $P_i^-$  is zero, set  $R_i^+ = 1$  or  $R_i^- = 1$ , respectively. For Dirichlet nodes, the values of  $R_i^+$  and  $R_i^-$  are set to 1.

3. Next, calculate

$$\tilde{\alpha}_{ij} = \begin{cases} R_i^+ & \text{if } f_{ij} > 0, \\ 1 & \text{if } f_{ij} = 0, \\ R_i^- & \text{if } f_{ij} < 0, \end{cases} \quad i, j = 1, \dots, N,$$

4. Finally, set

$$\alpha_{ij} = \min \{ \tilde{\alpha}_{ij}, \tilde{\alpha}_{ji} \}, \quad i, j = 1, \dots, N.$$

For AFC methods equipped with this limiter, the existence of the nonlinear problem and the satisfactions of local and global DMP on arbitrary simplicial grids was proved in [179]. Moreover, it was shown that the method is linearity-preserving. This property demands that the modification added to the formulation vanishes in case the solution is a polynomial of degree 1, hence leads to improved accuracy in the regions where the solution is smooth.

### 3. Monotone upwind-type algebraically stabilized (MUAS) method

This method was recently proposed in [177]. Let the stabilization term in (1.58)-(1.59) method be as follows:

$$b_{ij} = -\max \left\{ (1 - \alpha_{ij})a_{ij}, 0, (1 - \alpha_{ji})a_{ji} \right\}, \quad i, j = 1, \dots, N, \quad i \neq j, \quad (1.61)$$

$$b_{ii} = -\sum_{j=1}^N b_{ij}, \quad i = 1, \dots, N. \quad (1.62)$$

such that  $b_{ij} = b_{ji}$ ,  $i, j = 1, \dots, N$ . Then, the limiter  $\alpha_{ij}$  is determined as follows

1. First, compute

$$P_i^+ = \sum_{j=1, a_{ij}>0}^N a_{ij}(u_i - u_j)^+, \quad P_i^- = \sum_{j=1, a_{ij}>0}^N a_{ij}(u_i - u_j)^-,$$

and

$$Q_i^+ = \sum_{j=1}^N \max \left\{ |a_{ij}|, a_{ji} \right\} (u_j - u_i)^+, \quad Q_i^- = \sum_{j=1}^N \max \left\{ |a_{ij}|, a_{ji} \right\} (u_j - u_i)^-.$$

2. Then, calculate

$$R_i^+ = \min \left\{ 1, \frac{Q_i^+}{P_i^+} \right\}, \quad R_i^- = \min \left\{ 1, \frac{Q_i^-}{P_i^-} \right\}, \quad i = 1, \dots, M.$$

If  $P_i^+$  or  $P_i^-$  is zero, set  $R_i^+ = 1$  or  $R_i^- = 1$ , respectively. The values of  $R_i^+$  and  $R_i^-$  are 1 at Dirichlet nodes as well.

3. Lastly, define

$$\alpha_{ij} = \begin{cases} R_i^+ & \text{if } u_i > u_j, \\ 1 & \text{if } u_i = u_j, \\ R_i^- & \text{if } u_i < u_j, \end{cases} \quad i, j = 1, \dots, N.$$

The MUAS method has been analyzed in [177]. The solvability of the nonlinear discrete problem and satisfaction of local and global DMP on arbitrary simplicial grids are proven.

4. Symmetrized monotone upwind-type algebraically stabilized method

A new algebraically stabilized method with a new definition of limiter  $\alpha_{ij}$ , known as symmetrized monotone upwind-type algebraically stabilized (SMUAS) method, was recently published in [180]. It was shown that the SMUAS method is linearity-preserving and satisfies the DMP on arbitrary simplicial meshes. This method is defined by (1.61)-(1.62) and computed as follows:

1. Compute

$$\begin{aligned} P_i^+ &= \sum_{j \in S_i} |d_{ij}| \left\{ (u_i - u_j)^+ + (u_i - u_{ij})^+ \right\}, \\ P_i^- &= \sum_{j \in S_i} |d_{ij}| \left\{ (u_i - u_j)^- + (u_i - u_{ij})^- \right\}. \end{aligned} \quad (1.63)$$

2. Compute

$$\begin{aligned} Q_i^+ &= \sum_{j \in S_i} \max \left\{ |a_{ij}|, |a_{ji}| \right\} \left\{ (u_j - u_i)^+ + (u_{ij} - u_i)^+ \right\}, \\ Q_i^- &= \sum_{j \in S_i} \max \left\{ |a_{ij}|, |a_{ji}| \right\} \left\{ (u_j - u_i)^- + (u_{ij} - u_i)^- \right\}. \end{aligned} \quad (1.64)$$

3. Compute

$$R_i^+ = \min \left\{ 1, \frac{Q_i^+}{P_i^+} \right\}, \quad R_i^- = \min \left\{ 1, \frac{Q_i^-}{P_i^-} \right\}, \quad i = 1, \dots, M.$$

If  $P_i^+$  or  $P_i^-$  is zero, set  $R_i^+ = 1$  or  $R_i^- = 1$ , respectively. The values of  $R_i^+$  and  $R_i^-$  are set to 1 for Dirichlet nodes as well.

4. Define

$$\alpha_{ij} = \begin{cases} R_i^+ & \text{if } u_i > u_j, \\ 1 & \text{if } u_i = u_j, \\ R_i^- & \text{if } u_i < u_j, \end{cases} \quad i, j = 1, \dots, N.$$

In the above definition  $u_{ij} = u_i + \nabla u_h|_{T_i^j} \cdot (x_i - x_j)$ . Here  $T_i^j$  is a mesh cell containing  $x_i$  that is intersected by the half line

$$\left\{ x_i + \theta_i(x_i - x_j) : \theta_i > 0 \right\}.$$

Another possible choice of limiter known as BBK limiter is presented in [181]. Numerical studies in [182] show that AFC methods with BJK limiter usually provide more accurate results in comparison to Kuzmin limiter. The results in [183] revealed that using AFC method with Kuzmin limiter leads to solutions



with sharper layers compared with the solution obtained with the BBK limiter. The Kuzmin limiter was replaced by a value that introduces sufficient amount of artificial diffusion and fulfills the conditions to satisfy a local DMP in [184]. A mixture of both Kuzmin and BJK limiter was proposed in [185] to obtain a limiter which is linearity preserving and satisfies a local DMP. A simpler version of the limiter from [186] is considered in [177], which does not use the conventional inter-nodal fluxes as in other AFC methods, it was shown that the resulting method guarantees the DMP. For numerical comparison of the AFC scheme with Kuzmin limiter and the MUAS method see [177]. Another very promising limiting strategy includes the development of a monolithic convex (MC) limiting. This limiter was first proposed in [187] for linear advection equations and nonlinear hyperbolic conservation laws, and its adaptation to other types of equations was investigated very recently in [188]. AFC scheme with Kuzmin, BJK, and MC limiters has been investigated in [189] for a steady-state CDR equation in three dimensions.

As mentioned in the preceding, the discretized algebraically stabilized schemes are nonlinear. With nonlinearity different concern arises, this matter has been recently addressed in [190], where two basic fixed-point iterations and a Newton method were investigated for solving the nonlinear problem arising from the AFC discretization. This investigation was further continued in [182], where comprehensive numerical studies for solving the resulting nonlinear scheme were presented. In this work a mixed fixed-point iteration, a refinement of Newton method, and a regularized formal Newton method were used.

Despite the huge amount of numerical studies, the amount of rigorous mathematical analysis for the algebraically stabilized schemes was considerably low for a long time. It was only very recently that the first numerical analysis for this class of stabilized methods was carried out in [191]. It was shown that the linear problems obtained using the fixed-point iteration are well-posed, however, the nonlinear problem is not solvable in general. In this case, the possible non-existence of solution seems to be the result of the violation of the symmetry condition on the correction limiter, i.e, when  $\alpha_{ij} \neq \alpha_{ji}$ ,  $i, j = 1, \dots, N$ . This shortcoming has been addressed in [178], where for the first time the existence of the solution, existence and uniqueness of a solution of a linearized problem, and an a priori error estimate were proven under rather general assumptions on the limiters  $\alpha_{ij}$ . Moreover, it was shown that under the symmetry condition and certain restrictions on the mesh it is possible to prove the local DMP. A survey on the development and analysis of the AFC schemes has been published in [192] and [183]. The performance of a posteriori error estimates was studied in [193]. Furthermore, the behavior of the AFC method with Kuzmin, and BJK limiters and of the MUAS method on adaptively refined grids, both with conforming meshes and with hanging nodes is studied in [194]. For a survey on the finite element methods that satisfy a local or global DMP for convection-dominated CDR equations see [195]. A numerical study is presented in [196] which investigates finite element methods satisfying the DMP for CDR equations.

Algebraically stabilized methods have several advantages in comparison to

most of the other stabilization methods: they often satisfy DMP by construction and as a result preserve the positivity of the solution, their implementation does not depend on the space dimension, they usually provide sharp approximations of the layers. However, there are also some drawbacks: these methods are usually nonlinear even if the problem at hand is linear, and the application of algebraically stabilizations to higher order finite elements is not developed.

There are even more proposals of stabilized methods which are not included in this section. Among these methods are: consistent approximate upwind (CAU) methods [197], controlled consistent approximate upwind (CCAU) [198], Taylor-Galerkin method [199, 200], entropy viscosity approach [201], variational multiscale (VMS) methods [131, 132, 130], algebraic subgrid-scale stabilization [120, 202], orthogonal subscales methods [203, 204], and many more, see also [56, 205, 206] for reviews. However, despite more than four decade of intensive research, there is still no method that has been proven to be a universal choice, hence the numerical solution of convection-dominated CDR problems is still a challenge. Moreover, for many available discretizations of this simplest model problem the analysis still remains an open problem.

## 1.2 Stabilization of transient convection-diffusion-reaction equations

Time-dependent CDR equations appear in various applications. These equations are not only important on their own but are often part of complex nonlinear systems of equations that are strongly coupled in such a way that inaccuracy in one equation directly effects all other equations in the system. Therefore, over the years, a great deal of effort has been devoted to the development of proper numerical methods for approximating the solution of transient problems involving convection, diffusion, and reaction terms.

Let us consider the evolutionary CDR equation

$$\begin{aligned} u_t - \varepsilon \Delta u + \mathbf{b} \cdot \nabla u + cu &= g && \text{in } (0, T_{max}] \times \Omega, \\ u &= u_b && \text{on } [0, T_{max}] \times \partial\Omega, \\ u(x, 0) &= u_0(x), && \text{in } \Omega, \end{aligned} \quad (1.65)$$

where  $\Omega$  is a polygonal or polyhedral bounded domain in  $\mathbb{R}^d$ ,  $d \in \{1, 2, 3\}$  with a Lipschitz-continuous boundary  $\partial\Omega$ ,  $[0, T_{max}]$  is a time interval,  $\varepsilon > 0$  is a constant diffusion coefficient,  $\mathbf{b}(x, t) \in L^\infty(0, T_{max}; W^{1,\infty}(\Omega)^d)$  is a convection field with  $\nabla \cdot \mathbf{b} = 0$ ,  $c(x, t) \in L^\infty(0, T_{max}; L^\infty(\Omega))$  is a non-negative reaction coefficient,  $g(x, t) \in L^2(0, T_{max}; L^2(\Omega))$  is an outer source of the unknown quantity  $u$ ,  $u_b \in L^2(0, T_{max}; H^{\frac{1}{2}}(\partial\Omega))$  is the boundary condition, and  $u_0(x) \in H_0^1(\Omega)$  is the initial data. Moreover, without loss of generality, it is assumed that there exists a positive constant  $\sigma_0$  such that

$$0 < \sigma_0 \leq \sigma(t, x) = c(t, x) - \frac{1}{2} \operatorname{div} \mathbf{b}(t, x), \quad \forall (t, x) \in [0, T_{max}] \times \Omega, \quad (1.66)$$

which is the standard assumption that guarantees the unique solvability of (1.65), see [56]. Clearly, the numerical solution of an equation of type (1.65) consists of a double discretization process, that is, the temporal discretization and the spatial discretization. We wish to only focus on the finite element methods for spatial discretization in the following, thus we introduce a triangulation  $\mathcal{T}_h$  of  $\Omega$  which possesses the usual compatibility properties and define the finite element spaces as

$$W_h = \left\{ v_h \in C(\bar{\Omega}); v_h|_T \in R(T), \forall T \in \mathcal{T}_h \right\}, \quad V_h = W_h \cap H_0^1(\Omega),$$

consisting of continuous piecewise (multi)linear functions as before. As for the time discretization, the time interval is decomposed by  $0 = t^0 < t^1 < \dots < t^{n+1} = T_{max}$  with  $\Delta t = t^{n+1} - t^n$  being the time-step, afterward any time-integration methods can be employed for discretizing in time. Note that, in the following, we only consider the homogeneous Dirichlet boundary condition for the simplicity of the presentation.

The process of discretization can be done in several ways:

- The time/space discretization, also known as Rothe's method, in which after a temporal discretization the obtained equation has the form of a stationary CDR equation that has to be solved at each time instant. Then, the fully discrete equation can be derived by employing a proper spatial discretization method, see [207] for an example of this case. Now, let us consider  $\theta$ -scheme ( $\theta \in [0, 1]$ ) for discretizing (1.65) in time, this leads at each discrete time  $t^{n+1}$  to

$$\begin{aligned} u^{n+1} + \theta \Delta t \left( -\varepsilon \Delta u^{n+1} + \mathbf{b}^{n+1} \cdot \nabla u^{n+1} + c^{n+1} u^{n+1} \right) = & \quad (1.67) \\ u^n - (1 - \theta) \Delta t \left( -\varepsilon \Delta u^n + \mathbf{b}^n \cdot \nabla u^n + c^n u^n \right) + \theta \Delta t g^{n+1} + (1 - \theta) \Delta t g^n. \end{aligned}$$

Practical cases of interest are backward Euler ( $\theta = 1$ ), Crank–Nicolson ( $\theta = \frac{1}{2}$ ), and forward Euler ( $\theta = 0$ ). Now it remains to apply a finite element method for discretizing in (1.67) in space, in this regard the equation (1.67) can be transformed to a weak formulation by multiplying with a test function from a space  $V = H_0^1(\Omega)$  and applying the integration by parts as usual. Then, after employing a finite-dimensional space  $V_h$  instead of  $V$  ( $V_h \subset V$  for conforming finite element method) and with a suitable approximation of initial data  $u_0(x)$ , the standard Galerkin formulation of (1.67) reads as follow: Find  $u_h^{n+1} \in V_h$  such that

$$\begin{aligned} (u_h^{n+1}, v_h) + \theta \Delta t \left( (\varepsilon \nabla u_h^{n+1}, \nabla v_h) + (\mathbf{b}^{n+1} \cdot \nabla u_h^{n+1} + c^{n+1} u_h^{n+1}, v_h) \right) = & \\ (u_h^n, v_h) - (1 - \theta) \Delta t \left( (\varepsilon \nabla u_h^n, \nabla v_h) + (\mathbf{b}^n \cdot \nabla u_h^n + c^n u_h^n, v_h) \right) + & \quad (1.68) \\ \theta \Delta t (g^{n+1}, v_h) + (1 - \theta) \Delta t (g^n, v_h), \quad \forall v_h \in V_h. \end{aligned}$$

Note that the definition of  $V_h$  is based on the underlying triangulation  $\mathcal{T}_h$  of  $\Omega$  mentioned above.

- The space/time discretization, also known as the method of lines, in which after the discretization in space one obtains a semi-discrete equation, then

the fully discrete version can be derived by applying a suitable time integration method, see [208] for an example of this case. Now, let  $V = H_0^1(\Omega)$ , then the variational formulation of (1.65) reads as follow: Find  $u \in [0, T_{max}] \rightarrow V$  such that

$$(u_t(t), v) + (\varepsilon \nabla u(t), \nabla v) + (\mathbf{b}(t) \cdot \nabla u(t) + c(t) u(t), v) = (g(t), v), \quad (1.69)$$

$$u(0) = u_0 \quad \text{in } \Omega,$$

for  $\forall v \in V$  and  $t \in (0, T_{max}]$ . Replacing  $V$  by a standard finite-dimensional space of piecewise polynomial functions  $V_h$ , the time-continuous Galerkin formulation then aims to find a function  $u_h : [0, T_{max}] \rightarrow V_h$  such that

$$(u_{h,t}(t), v_h) + (\varepsilon \nabla u_h(t), \nabla v_h) + (\mathbf{b}(t) \cdot \nabla u_h(t) + c(t) u_h(t), v_h) = (g(t), v_h), \quad (1.70)$$

$$u_h(0) = u_{0h} \quad \text{in } \Omega,$$

for  $\forall v_h \in V_h$  and  $t \in (0, T_{max}]$ . After this discretization, temporal discretization can follow by employing any suitable time-stepping technique.

- There is also space-time discretization technique, where coupled time and space element are used. As argued in [79], the time derivative and the spatial derivative terms should be combined into a single "material derivative" in this case, see also [209, 90].

It is known that the space-time discretization is the most natural setting to develop finite element methods for problems of type (1.65), however, the number of unknowns for coupled space-time formulation is very high in many applications, which increases the computational cost and that is a rather significant drawback. Hence, separated fully discrete discretization are in much more use for transient CDR problems.

Typically, the size of the convection and/or reaction is much larger by order of magnitude compared to the size of the diffusion, therefore the standard Galerkin finite element method usually fails to deal with spatial discretization in this case. As mentioned in the previous section, a characteristic feature of solutions in convection-dominated regime is the presence of sharp layers, therefore suitable stabilization techniques are in need to be able to approximate these sharp layers properly on one hand and to prevent the occurrence of wild oscillations in these areas on the other hand. Stabilization methods were initially designed for the steady-state CDR equations and eventually found their way to transient problems with time-stepping methods and also to space-time formulations. In the following we shall recall some of these techniques which have been introduced to deal with evolutionary CDR problems in convection-dominated regime over the years.

## **1. SUPG method**

To begin with we start with the SUPG stabilization method which adds a consistent diffusion term in the streamline direction to the original Galerkin formulation of the problem. In case Rothe's method is used for the discretization,

the SUPG finite element method applied to (1.67) consists in finding an approximate solution  $u_h^{n+1} \in V_h$  to  $u^{n+1} \in V$  such that

$$\begin{aligned}
& (u_h^{n+1}, v_h) + \sum_{T \in \mathcal{T}_h} (\tau_T \theta \Delta t) (u_h^{n+1}, \mathbf{b}^{n+1} \cdot \nabla v_h)_T \\
& + \theta \Delta t \left[ (\varepsilon \nabla u_h^{n+1}, \nabla v_h) + (\mathbf{b}^{n+1} \cdot \nabla u_h^{n+1} + c^{n+1} u_h^{n+1} - g^{n+1}, v_h) \right. \\
& + \left. \sum_{T \in \mathcal{T}_h} (\tau_T \theta \Delta t) \left( (-\varepsilon \Delta u_h^{n+1} + \mathbf{b}^{n+1} \cdot \nabla u_h^{n+1} + c^{n+1} u_h^{n+1} - g^{n+1}), \mathbf{b}^{n+1} \cdot \nabla v_h \right)_T \right] \\
& = (u_h^n, v_h) + \sum_{T \in \mathcal{T}_h} (\tau_T \theta \Delta t) (u_h^n, \mathbf{b}^n \cdot \nabla v_h)_T \\
& - (1 - \theta) \Delta t \left[ (\varepsilon \nabla u_h^n, \nabla v_h) + (\mathbf{b}^n \cdot \nabla u_h^n + c^n u_h^n - g^n, v_h) \right. \\
& , + \left. \sum_{T \in \mathcal{T}_h} (\tau_T \theta \Delta t) \left( (-\varepsilon \Delta u_h^n + \mathbf{b}^n \cdot \nabla u_h^n + c^n u_h^n - g^n), \mathbf{b}^n \cdot \nabla v_h \right)_T \right], \tag{1.71}
\end{aligned}$$

for  $\forall v_h \in V_h$ . This is the time discrete stabilized approximation of (1.65) with  $\theta$ -scheme used as the temporal discretization and with a suitable approximation of initial data  $u_0(x)$ , where  $\tau_T$  is the stabilization parameter. Proper choices of the stabilization parameter  $\tau_T$  has been studied in various ways in the literature, most popular among them is based on the convergence analysis of the method. In contrast to the steady-state, where the role of the reaction was usually neglected in the stabilization parameter, this term plays a crucial role in the transient case specially whenever the time-step size is very small, thus, a suitable parameter usually takes this term into account. It was suggested in [158, 207] to chose the stabilization parameter  $\tau_T$  as:

$$\tau_T = \left( \frac{h_T^2}{4\theta \Delta t \varepsilon + 2h_T \theta \Delta t |\mathbf{b}^{n+1}| + h_T^2 (1 + \theta \Delta t c^{n+1})} \right),$$

where  $h_T$  is an appropriate measure for the size of the mesh cell  $T$  which is usually chosen in the direction of the convection field. Another possible proposal was proposed in [159, 210, 207] as

$$\tau_T = \min \left\{ \frac{h_T}{2\theta \Delta t |\mathbf{b}^{n+1}|}, \frac{1}{1 + \theta \Delta t c^{n+1}}, \frac{h_T^2}{\theta \Delta t \varepsilon} \right\}.$$

In [105, 211, 207], it was suggested to set

$$\tau_T = \left( \frac{h_T^2}{(1 + \theta \Delta t c^{n+1}) h_T^2 \xi(Pe_{T,1}) + 6\theta \Delta t \varepsilon \xi(Pe_{T,2})} \right),$$

with

$$Pe_{T,1} = \frac{6\theta \Delta t \varepsilon}{h_T^2 (1 + \theta \Delta t c^{n+1})}, \quad Pe_{T,2} = \frac{h_T \theta \Delta t |\mathbf{b}^{n+1}|}{3\theta \Delta t \varepsilon},$$

and

$$\xi(s) = \begin{cases} 1 & \text{if } 0 < s \leq 1, \\ s & \text{if } s \geq 1. \end{cases}$$

In case method of lines is used for the discretization, the finite element SUPG approximation of (1.70) has the following form (time-continuous case): For all  $t \in (0, T_{max}]$  find  $u_h(t) \in V_h$  such that

$$\begin{aligned}
& \left( u_{h,t}(t), v_h \right) + \sum_{T \in \mathcal{T}_h} \left( u_{h,t}(t), \mathbf{b}(t) \cdot \nabla v_h \right)_T \\
& + \left( \varepsilon \nabla u_h(t), \nabla v_h \right) + \left( \mathbf{b}(t) \cdot \nabla u_h(t) + c(t) u_h(t), v_h \right) \\
& + \sum_{T \in \mathcal{T}_h} \left( -\varepsilon \Delta u_h(t) + \mathbf{b}(t) \cdot \nabla u_h(t) + c(t) u_h(t), \mathbf{b}(t) \cdot \nabla v_h \right)_T \\
& = \left( g(t), v_h \right) + \sum_{T \in \mathcal{T}_h} \left( g(t), \mathbf{b}(t) \cdot \nabla v_h \right), \tag{1.72}
\end{aligned}$$

for  $\forall v_h \in V_h$  and  $t \in (0, T_{max}]$  and a suitable approximation of initial data. This formulation is strongly consistent. Some authors has suggested the use of a non-consistent treatment of the time derivative as: For all  $t \in (0, T_{max}]$  find  $u_h(t) \in V_h$  such that

$$\begin{aligned}
& \left( u_{h,t}(t), v_h \right) + \left( \varepsilon \nabla u_h(t), \nabla v_h \right) + \left( \mathbf{b} \cdot \nabla u_h(t) + c u_h(t), v_h \right) \\
& + \sum_{T \in \mathcal{T}_h} \left( -\varepsilon \Delta u_h(t) + \mathbf{b} \cdot \nabla u_h(t) + c u_h(t), \mathbf{b} \cdot \nabla v_h \right)_T \\
& = \left( g_h(t), v_h \right) + \sum_{T \in \mathcal{T}_h} \left( g_h(t), \mathbf{b} \cdot \nabla v_h \right), \tag{1.73}
\end{aligned}$$

for  $\forall v_h \in V_h$  and  $t \in (0, T_{max}]$ . However, it turned out that this method loses its accuracy, regardless of the choice of the time-discretization.

Lastly, in the case of space-time approach, it was suggested in [79] to use the SUPG method together with the discontinuous Galerkin method in time. The idea behind this approach was to be able to treat the temporal derivative like first order spatial derivative. In this case the space-time SUPG stabilization term is given by

$$S_{SUPG}(v_h) = v_h' + \mathbf{b} \cdot \nabla v_h,$$

see [212] for more details.

Concerning the numerical analysis, a post-processing technique was used in the approximation of linear transient CDR equations in [213]. The technique followed the procedure of approximation with a standard Galerkin method until a certain time instant and then proceeded by using a SUPG stabilization method in a single steady problem. The error bounds were obtained with regard to the SUPG norm in convection-dominated regime and a sub-optimal convergence was proven. Moreover, it was shown that the temporal error of the fully discrete post-processed approximation can be bounded by the temporal error of the Galerkin approximation. An extension of this technique for stabilizing evolutionary convection-diffusion problems was studied in [214]. The stability of the SUPG finite element method was demonstrated analytically in [208] for convection-diffusion equations. In this study, it was pointed out that the combination of the SUPG stabilization method and and implicit time-stepping scheme

can be considered as a safe and effective method regardless of the length of the time-step. It was also mentioned that for a very fine time-step size spurious oscillations might occur in the front layers in the numerical simulation. Later, the first theoretical investigation for the small time-step instability for the SUPG method was reported in [215]. In this work, a pure evolutionary transport equation was considered, where the stability and quasi-optimal convergence in time was established for the SUPG finite element discretization in space together with the backward Euler and Crank–Nicolson finite difference discretization in time. These results were obtained for sufficiently smooth data. However, it was shown that if the data are not sufficiently regular, for small time-step a deterioration might be detected in the stability and convergence analysis which may lead to some oscillations in the layers. The analysis of [215] was further extended in [216] to CDR problems. It was argued that the source of instability for small time-step might be the result of the choice of the stabilization parameter and their dependence on the length of the time-step, whenever Rothe’s method is used. It was pointed out that the stabilization vanishes if the time-step length approaches zero. This idea motivated the authors to search for stabilization parameters that do not depend on the length of the time-step. Baring in mind that the necessity of entering the time derivative in the stabilization parameter secures the consistency, the investigation proceeded by proving error estimates in the  $L^2$ -norm, the norm of the material derivative, and also the norm of the streamline derivative (which were missing in the study of [215]). This allowed the stabilization parameter to be chosen similarly to the steady-state case. The proof was carried out under certain regularity conditions on the data while the backward Euler and Crank–Nicolson scheme were used in the temporal discretization. The method of lines was used for the discretization of the CDR equation in [217], it was shown that the proposed method is adequately stable since due to the procedure of discretizing in space before time, the stabilization parameters are free of the length of the time-step. Another possibility that attracted the attention of some authors was to not only consider adding artificial diffusion to deal with spatial stability but also to consider more proper time-discretization to overcome temporal instability. In this regard, [218] suggested to combine the standard Galerkin method with high-order multi-step explicit method which however led to severe restrictions on the allowable time-steps in either convection- or diffusion-dominated regime. Later, it was suggested in [219, 220] to employ a high-order implicit time-stepping scheme coupled with classical stabilized techniques, and it was shown that the method is spatially stable and highly accurate.

## **2. GLS method**

As for the GLS method, space-time formulation was traditionally the method of use at the start for discretizing the transient CDR equations which was referred to as ST-GLS formulation. In this case the element domain was considered as coupled space and time elements and the formulation was given by

$$S_{ST-GLS}(v_h) = v_h' - \varepsilon \Delta v_h + \mathbf{b} \cdot \nabla v_h + cv_h,$$

see [221, 222] for more details.

By performing the time discretization prior to the spatial one, similarly to the SUPG method the GLS formulation of the problem (1.65) takes the form: Find  $u_h^{n+1} \in V_h$  such that

$$\begin{aligned}
& (u_h^{n+1}, v_h) + \sum_{T \in \mathcal{T}_h} (\tau_T \theta \Delta t) (u_h^{n+1}, -\varepsilon \Delta v_h + \mathbf{b}^{n+1} \cdot \nabla v_h + c^{n+1} v_h)_T \\
& + \theta \Delta t \left[ (\varepsilon \nabla u_h^{n+1}, \nabla v_h) + (\mathbf{b}^{n+1} \cdot \nabla u_h^{n+1} + c^{n+1} u_h^{n+1}, v_h) \right. \\
& + \sum_{T \in \mathcal{T}_h} (\tau_T \theta \Delta t) \\
& \quad \left. \left( -\varepsilon \Delta u_h^{n+1} + \mathbf{b}^{n+1} \cdot \nabla u_h^{n+1} + c^{n+1} u_h^{n+1}, -\varepsilon \Delta v_h + \mathbf{b}^{n+1} \cdot \nabla v_h + c^{n+1} v_h \right)_T \right] \\
& = (u_h^n, v_h) + \sum_{T \in \mathcal{T}_h} (\tau_T \theta \Delta t) (u_h^n, -\varepsilon \Delta v_h + \mathbf{b}^n \cdot \nabla v_h + c^n v_h)_T \\
& - (1 - \theta) \Delta t \left[ (\varepsilon \nabla u_h^n, \nabla v_h) + (\mathbf{b}^n \cdot \nabla u_h^n + c^n u_h^n, v_h) \right. \\
& + \sum_{T \in \mathcal{T}_h} (\tau_T \theta \Delta t) \left( -\varepsilon \Delta u_h^n + \mathbf{b}^n \cdot \nabla u_h^n + c^n u_h^n, -\varepsilon \Delta v_h + \mathbf{b}^n \cdot \nabla v_h + c^n v_h \right)_T \left. \right] \\
& + \theta \Delta t \left[ (g^{n+1}, v_h) + \sum_{T \in \mathcal{T}_h} (\tau_T \theta \Delta t) (g^{n+1}, -\varepsilon \Delta v_h + \mathbf{b}^{n+1} \cdot \nabla v_h + c^{n+1} v_h)_T \right] \\
& + (1 - \theta) \Delta t \left[ (g^n, v_h) + \sum_{T \in \mathcal{T}_h} (\tau_T \theta \Delta t) (g^n, -\varepsilon \Delta v_h + \mathbf{b}^n \cdot \nabla v_h + c^n v_h)_T \right] \quad (1.74)
\end{aligned}$$

for all  $v_h \in V_h$ . The method of lines version of the GLS method can be obtained similarly as the SUPG method as well. From the practical point of view, there are no major differences between SUPG and GLS methods, especially when there is no reaction term present and linear elements are used (the second-order derivatives are zero in the element interior). Several expressions of the stabilization parameter  $\tau_T$  have been proposed for the GLS scheme for transient problems which are mainly based on the minimization of the error, see [90]. Finally, an optimal error bound was derived in [223], where a Galerkin least-squares method was employed for space discretization along with a  $\theta$ -scheme for time discretization.

### 3. LPS method

Performing with local projection into a large scale space, LPS methods add an appropriate stabilization to small scale of the Galerkin finite element solution. Employing two finite element spaces at once, and applying either two-level method or enrichment method, the LPS schemes add a linear stabilization term of the form

$$S_{LPS}(u_h, v_h) := \sum_{M \in \mathcal{M}_h} \tau_M \left( \kappa_M(\nabla u_h), \kappa_M(\nabla v_h) \right)_M, \quad \forall v_h \in V_h, \quad (1.75)$$

where  $\kappa_M$  denotes the fluctuation operator, and  $\mathcal{M}_h$  can be equal to  $\mathcal{T}_h$  in case enrichment method is used, or a coarser triangulation with macro-elements  $M$  (which are a collection of elements  $T$ ) in case of two-level technique. The numerical simulation of the LPS method for time-dependent CDR equations was



first studied in [207], where the enriched LPS method was used and compared with other stabilization methods in convection-dominated regime. In [224], a one-step  $\theta$ -scheme as time discretization was used along with an LPS finite element method containing a nonlinear crosswind-diffusion term for an evolutionary CDR equation. It was shown that both the fully nonlinear and its linearized version are solvable without any restriction on the time-step. Moreover, the uniqueness was established for both approaches and an a priori error estimate with respect to the standard LPS norm was derived. The use of high-order discretization was tackled in [225], where the discontinuous Galerkin method was applied alongside the LPS method for a partial differential equation in a time-dependent domain. The application of the LPS scheme combined with the discontinuous Galerkin method in time was studied for fixed-domain problems in [226]. A monotone local projection scheme for a convection-dominated transport problem was introduced in [227], where the stabilization term was proposed in such a way that it guaranteed the monotonicity and enforced the linearity preservation on each mesh.

#### 4. SOLD methods

SOLD methods have originally been introduced to reduce or even completely remove spurious oscillations at layers from the solutions obtained using SUPG methods for stationary CDR equations. Later, they were utilized with other stabilization methods such as LPS and further extended and were added to stabilization methods for transient CDR problems. These methods are in general nonlinear. Even though there is no general SOLD method known that works perfectly well on every example, it is shown that these methods still work remarkably well for many special cases and are able to reduce the oscillations considerably in the vicinity of sharp layers. An application of various types of SOLD methods for time-dependent CDR problems can be found in [228] and also in [207], among other studies. As mentioned, these methods work in different ways:

- SOLD methods which add isotropic diffusion to the SUPG stabilization method (1.71) as:

$$\left( \tilde{\varepsilon} \nabla u_h^{n+1}, \nabla v_h \right), \quad (1.76)$$

with  $\tilde{\varepsilon} = \sigma \frac{|R_h(u_h^{n+1})|^2}{\|\nabla u_h^{n+1}\|_2^2}$  where  $R_h(u_h^{n+1})$  is the residual of (1.67). The definition of  $\sigma$  has been investigated in many studies. It was shown in [149, 207] that  $\sigma$  can be chosen as

$$\sigma|_T = \max \left\{ 0, \tau_T \left( \theta \Delta t \frac{R_h(u_h^{n+1}) \nabla u_h^{n+1}}{|\nabla u_h^{n+1}|^2} \right) - \tau_T \right\},$$

where  $\tau_T$  is the SUPG stabilization parameter and  $\tau_T(s)$  is this parameter

evaluated in  $s$ . Another idea was suggested in [150, 207] to chose  $\sigma$

$$\sigma|_T = \tau_T \left\{ 0, \frac{\theta \Delta t |\mathbf{b}^{n+1}|}{\left| \theta \Delta t \frac{R_h(u_h^{n+1}) \nabla u_h^{n+1}}{|\nabla u_h^{n+1}|^2} \right|} - 1 \right\}.$$

We refer the reader to [152, 207, 229] for other examples of this parameters and their derivation.

- SOLD methods which add anisotropic diffusion terms as

$$\left( \tilde{\varepsilon} D \nabla u_h^{n+1}, \nabla v_h \right), \quad (1.77)$$

with

$$D = \begin{cases} I - \frac{(\theta \Delta t \mathbf{b}^{n+1}) \otimes (\theta \Delta t \mathbf{b}^{n+1})}{\theta \Delta t |\mathbf{b}^{n+1}|^2} & \text{if } \theta \Delta t \mathbf{b} \neq 0, \\ 0 & \text{else.} \end{cases}$$

for which the parameter  $\tilde{\varepsilon}$  can have the form

$$\tilde{\varepsilon}|_T = \max \left\{ 0, C \frac{\text{diam}(T) |R_h(u_h^{n+1})|}{2 |\nabla u_h^{n+1}|} - \theta \Delta t \varepsilon \right\},$$

as considered in [159, 78, 207], or it can be chosen as

$$\tilde{\varepsilon}|_T = \frac{\tau_T \theta \Delta t |\mathbf{b}^{n+1}|^2 |R_h(u_h^{n+1})|}{\theta \Delta t |\mathbf{b}^{n+1}|^2 |\nabla u_h^{n+1}| + |R_h(u_h^{n+1})|} \quad (1.78)$$

as proposed in [160, 78, 207]. Other choices of the parameter  $\tilde{\varepsilon}$  can be found in [158, 228, 157].

- SOLD methods of edge stabilization type

$$\sum_{T \in \mathcal{T}_h} |T| \int_{\partial T} \Psi_T(u_h^{n+1}) \text{sign} \left( \frac{\partial u_h^{n+1}}{\partial \mathbf{t}_{\partial T}} \right) \left( \frac{\partial v_h}{\partial \mathbf{t}_{\partial T}} \right) d\sigma, \quad (1.79)$$

with the parameter function  $\Psi_T(u_h^{n+1})$  which can be chosen as in [78] given by

$$\Psi_T(u_h^{n+1}) = C \left| R_h(u_h^{n+1}) \right|_T.$$

where  $C$  is a constant. For more examples of this function see [164, 135, 207].

## 5. FCT method

Applying a one-step  $\theta$ -method and using the standard Galerkin finite element discretization of (1.65) leads to an equation of the form (1.68) at each time instant  $t^{n+1}$  which can be rewritten in an algebraic form as follow:

$$\left( \mathbb{M}_C + \theta \Delta t \mathbb{A}^{n+1} \right) \mathbf{u}^{n+1} = \left( \mathbb{M}_C - (1 - \theta) \Delta t \mathbb{A}^n \right) \mathbf{u}^n + \theta \Delta t G^{n+1} + (1 - \theta) \Delta t G^n, \quad (1.80)$$

where  $\mathbb{M}_C = (m_{ij})_{i,j=1,\dots,N}$  is the consistent mass matrix,  $\mathbb{A}^{n+1} = (a_{ij})_{i,j=1,\dots,N}^{n+1}$  is the stiffness matrix consisting of the sum of diffusion, convection, and reaction,  $G^{n+1} = (g_1^{n+1}, \dots, g_N^{n+1})^T$  is the source vector, and  $\mathbf{u}^{n+1} = (u_1^{n+1}, \dots, u_N^{n+1})^T$  denotes the vector of unknowns. The matrix and vector entries are defined by

$$\begin{aligned} m_{ij} &= (\phi_j, \phi_i), \\ a_{ij}^{n+1} &= (\varepsilon \nabla \phi_j, \nabla \phi_i) + (\mathbf{b}^{n+1} \cdot \nabla \phi_j + c^{n+1} \phi_j, \phi_i), \\ g_i^{n+1} &= (g, \phi_i), \end{aligned}$$

where  $N$  is the number of degrees of freedom (the length of the vectors) and  $\phi_i$  are defined as in the previous section.

As already mentioned, the solution of (1.80) exhibits massive spurious oscillations in the convection-dominated regime. Although stabilization methods such as SUPG work quite well in reducing these oscillations, they are not able to completely suppress the under- and over-shoots in various situations specially whenever sharp layers are present. It is known however that the numerical methods that satisfy the DMP such as FE-FCT approaches work remarkably well in this regard. To fulfill the requirements of DMP and derive a FE-FCT scheme, set

$$\begin{aligned} \mathbb{M}_L &= \text{diag}(m_i), \quad m_i = \sum_{j=1}^N m_{i,j}, \quad i = 1, \dots, N, \\ \mathbb{D}^{n+1} &= (d_{ij}^{n+1})_{ij}^{n+1}, \quad d_{ij}^{n+1} = -\max\{a_{ij}^{n+1}, 0, a_{ji}^{n+1}\}, \quad \forall i \neq j, \quad d_{ii}^{n+1} = -\sum_{j \neq i} d_{ij}^{n+1}, \\ \mathbb{L}^{n+1} &= \mathbb{A}^{n+1} + \mathbb{D}^{n+1}. \end{aligned} \tag{1.81}$$

The matrix  $\mathbb{M}_L$  is called the lumped mass matrix. Now, (1.80) can be replaced by

$$(\mathbb{M}_L + \theta \Delta t \mathbb{L}^{n+1}) \mathbf{u}^{n+1} = (\mathbb{M}_L - (1 - \theta) \Delta t \mathbb{L}^n) \mathbf{u}^n + \theta \Delta t G^{n+1} + (1 - \theta) \Delta t G^n. \tag{1.82}$$

Note that  $\mathbb{L}^{n+1}$  does not possess positive off-diagonal entries. This method represents a stable low-order counterpart of (1.80), which does not show any under- or over-shoots, however, it is extremely diffusive and smears the layers to a great extent. Now, the goal is to modify the right-hand side of (1.82) in such a way that the solution becomes less diffusive while at the same time the spurious oscillations are still precluded. First, (1.82) is written in the form

$$\begin{aligned} (\mathbb{M}_L + \theta \Delta t \mathbb{L}^{n+1}) \mathbf{u}^{n+1} &= (\mathbb{M}_L (1 - \theta) \Delta t \mathbb{L}^n) \mathbf{u}^n \\ &+ \theta \Delta t G^{n+1} + (1 - \theta) \Delta t G^n + \mathbf{f}^{n+1}(\mathbf{u}^{n+1}, \mathbf{u}^n), \end{aligned} \tag{1.83}$$

where  $\mathbf{f}^{n+1}(\mathbf{u}^{n+1}, \mathbf{u}^n)$  is defined by subtracting the residual of (1.80) from (1.82). Since the matrix  $\mathbb{D}$  and also the difference of the mass and lumped mass matrices (i.e.,  $\mathbb{M}_C - \mathbb{M}_L$ ) have zero row sums, one can write

$$\mathbf{f}_i^{n+1}(\mathbf{u}^{n+1}, \mathbf{u}^n) = \sum_{j=1}^N f_{ij}^{n+1}, \quad i = 1, \dots, N, \tag{1.84}$$

where  $f_{ij}^{n+1}$  denote the fluxes which are given as

$$\begin{aligned} f_{ij}^{n+1} &= m_{ij}(u_i^{n+1} - u_j^{n+1}) - m_{ij}(u_i^n - u_j^n) \\ &\quad - \theta \Delta t d_{ij}^{n+1}(u_i^{n+1} - u_j^{n+1}) - (1 - \theta) \Delta t d_{ij}^n(u_i^n - u_j^n), \quad i, j = 1, \dots, N. \end{aligned} \quad (1.85)$$

Now, the idea of flux correction in the FE-FCT scheme is to restrict those fluxes  $f_{ij}^{n+1}$  that would otherwise produce spurious oscillations. This can be done by introducing

$$\mathbf{F}_i^{n+1}(\mathbf{u}^{n+1}, \mathbf{u}^n) = \sum_{j=1}^N \alpha_{ij}^{n+1} f_{ij}^{n+1}, \quad i = 1, \dots, N, \quad (1.86)$$

where  $\alpha_{ij}^{n+1} \in [0, 1]$  is a solution-dependent correction factor, which has to be symmetric in order to maintain conservativity, i.e.,  $\alpha_{ij}^{n+1} = \alpha_{ji}^{n+1}$ ,  $i, j = 1, \dots, N$ . Then, the next step is to replace  $\mathbf{f}^{n+1}(\mathbf{u}^{n+1}, \mathbf{u}^n)$  in (1.83) by  $\mathbf{F}^{n+1}(\mathbf{u}^{n+1}, \mathbf{u}^n)$ . The Galerkin method is recovered in the smooth regions for  $\alpha_{ij}^{n+1} = 1$  while  $\alpha_{ij}^{n+1} = 0$  leads to the low-order scheme in the layers.

The FE-FCT scheme is a nonlinear scheme and can be treated in two different ways:

- The nonlinear version: which utilizes an explicit solution  $\bar{\mathbf{u}}$  with forward Euler scheme at the time  $t^{n+1} - \frac{\Delta t}{2}$  as was suggested in [170] which reads

$$\bar{\mathbf{u}} = \mathbf{u}^n - (1 - \theta) \Delta t \mathbb{M}_L^{-1} (\mathbb{L}^n \mathbf{u}^n - G^n), \quad (1.87)$$

where  $\bar{\mathbf{u}}$  is used to define the correction factors  $\alpha_{ij}^{n+1}$  for the nonlinear FE-FCT scheme.

- The linearized version: which utilizes an approximation obtained using an explicit scheme instead of  $\mathbf{u}^{n+1}$  in the fluxes  $f_{ij}^{n+1}$ . Using the idea in [174], and applying  $\mathbf{u}^{n+\frac{1}{2}} = \frac{\mathbf{u}^{n+1} + \mathbf{u}^n}{2}$  in the definition leads to

$$f_{ij}^{n+1} = \Delta t m_{ij} \left( v_i^{n+\frac{1}{2}} - v_j^{n+\frac{1}{2}} \right) + \Delta t d_{ij}^{n+1} \left[ (u_j^n - u_i^n) + \theta \Delta t \left( v_j^{n+\frac{1}{2}} - v_i^{n+\frac{1}{2}} \right) \right],$$

where

$$v_i^{n+\frac{1}{2}} = \mathbb{M}_L^{-1} (G^n - \mathbb{L}^n \mathbf{u}^n)_i.$$

For computing the correction factor  $\alpha_{ij}^{n+1}$  different algorithms has been introduced over the years, among them is Zalesak's limiter suggested in [168] and Monolithic Convex limiting proposed in [187], which are as follows:

### 1. Zalesak's limiter

1. Compute

$$P_i^+ := \sum_{j \in S_i} \max\{0, f_{ij}^{n+1}\}, \quad P_i^- := \sum_{j \in S_i} \min\{0, f_{ij}^{n+1}\},$$

where  $S_i = \{j \in \{1, \dots, N\} \setminus \{i\} : \exists T \in \mathcal{T}_h : x_i, x_j \in \bar{T}\}$ .

2. Compute

$$Q_i^+ := \max_{j \in S_i \cup \{i\}} \{\bar{u}_j^{n+1} - \bar{u}_i^{n+1}\}, \quad Q_i^- := \min_{j \in S_i \cup \{i\}} \{\bar{u}_j^{n+1} - \bar{u}_i^{n+1}\},$$

where  $\bar{\mathbf{u}}^{n+1}$  is the solution of (1.87), which is used to guarantee the positivity of the solution.

3. Compute

$$R_i^+ := \begin{cases} \min\left(1, \frac{m_i Q_i^+}{\Delta t P_i^+}\right) & \text{if } P_i^+ > 0, \\ 1 & \text{if } P_i^+ = 0. \end{cases}, \quad R_i^- := \begin{cases} \min\left(1, \frac{m_i Q_i^-}{\Delta t P_i^-}\right) & \text{if } P_i^- < 0, \\ 1 & \text{if } P_i^- = 0. \end{cases}$$

4.

$$\alpha_{ij} = \begin{cases} \min\{R_i^+, R_j^-\} & \text{if } f_{ij}^{n+1} > 0, \\ \min\{R_i^-, R_j^+\} & \text{otherwise.} \end{cases}$$

This choice of correction factors guarantees that the aforementioned FE-FCT method satisfies the DMP.

### 1. Monolithic Convex limiter

The effect of time-stepping methods on FE-FCT scheme for transient CDR equations appears to be problematic in some situations. In this regard, an alternative was introduced in [187], known as Monolithic Convex (MC) limiting strategy, where suggested that the limited fluxes  $\alpha_{ij}^{n+1} f_{ij}^{n+1} = \bar{f}_{ij}^{n+1}$  could be defined by

$$\bar{f}_{ij}^{n+1} := \begin{cases} \min\left\{f_{ij}^{n+1}, \min\left\{2d_{ij}^{n+1}(\bar{u}_{ij} - u_i^{max}), 2d_{ij}^{n+1}(u_j^{min} - \bar{u}_{ji}), \right\}\right\} & \text{if } f_{ij}^{n+1} > 0, \\ 0 & \text{if } f_{ij}^{n+1} = 0, \\ \max\left\{f_{ij}^{n+1}, \max\left\{2d_{ij}^{n+1}(\bar{u}_{ij} - u_i^{min}), 2d_{ij}^{n+1}(u_j^{max} - \bar{u}_{ji}), \right\}\right\} & \text{if } f_{ij}^{n+1} < 0. \end{cases} \quad (1.88)$$

where

$$2d_{ij}^{n+1} \bar{u}_{ij} = (d_{ij}^{n+1} + d_{ij}^{n+1}) (u_i^{n+1} + u_j^{n+1}),$$

$$u_i^{max} = \max_{j \in S_i \cup \{i\}} u_j^{n+1}, \quad u_i^{min} = \min_{j \in S_i \cup \{i\}} u_j^{n+1}.$$

The linear and nonlinear FE-FCT schemes were considered in [230], in which the accuracy and efficiency of the schemes was investigated. It was shown that

the nonlinear FE-FCT schemes usually provide the most accurate results, however, the explicit schemes seem to be faster and more efficient in practice. A numerical study of FE-FCT methods for solving scalar transient CDR equations was presented in [207]. The method was then compared with other traditional stabilization techniques in convection-dominated regime. The numerical examples were carried out in 2D with homogeneous Dirichlet boundary conditions. Later, the results were extended to 3D problems with inhomogeneous Dirichlet and homogeneous Neumann boundary conditions in [231]. In [232], the FE-FCT scheme was generalized to implicit finite element scheme and nonlinear systems of hyperbolic conservation laws. The Zalesak's limiter was revisited due to its dependence on the time-step length which effects the stability and positivity of the solutions, therefore the use of an iterative limiting strategy was suggested. Moreover, an extension of the FE-FCT scheme to compressible Euler equations was investigated in this study. Three FE-FCT techniques based on the Runge-Kutta, Crank-Nicolson, and backward Euler time-integration were presented in [174], it was shown that the resulted methods are robust and efficient. Additionally, an alternative to FE-FCT schemes was introduced using an intermediate solution of a positivity-preserving low-order scheme. As noted before, the FE-FCT methods are nonlinear, hence the application of different types of solvers were studied in [189] and the question of high accuracy and efficiency was addressed.

Due to the construction of FE-FCT methods which mainly targets the algebraic form of the discretized finite element method compared to other traditional stabilized schemes that usually work by modifying the bilinear form, the numerical analysis of these methods are quite scarce, specially in time-dependent case. First attempt in this regard was reported in [233], where the FCT method was applied to a time-dependent convection-diffusion equation and also to a pure transport equation. Making use of the implicit function theorem for Lipschitz functions, the existence and uniqueness of a solution was proved provided that the time-steps are sufficiently small. In [234], the existence of a solution was established for a nonlinear FE-FCT scheme with arbitrary time-steps. The proof used a consequence of Brouwer's fixed-point theorem. Applying the backward Euler for temporal discretization, the error analysis and stability of the linear and nonlinear FE-FCT schemes for evolutionary CDR equation was reported in [235]. The optimal rate of convergence in  $L^2$  and  $H^1$  norm and sub-optimal convergence rate in FCT norm was proved. The DMP and positivity-preservation for stationary and evolutionary CDR equations was extensively studied in [195].

There are other stabilization techniques for time-dependent convection dominated CDR equations which are not included in this section. Among them are the US-FEM [236], the Bubble stabilization method [237], the subgrid scale method [238], the characteristic Galerkin method [239, 240], and the Taylor-Galerkin method [199]. We refer the reader to [207, 219, 220, 241] for a comparison regarding some of these stabilization approaches. However, we would like to note that to our best knowledge, there seem to be no results on Mizukami-Hughes methods in this regard.

### 1.3 Application of stabilized methods to cross-diffusion systems

As mentioned in the introduction, the very important aspect of a cross-diffusion system is the presence of cross-diffusion term(s), which makes the theoretical and numerical analysis of such a problems much more challenging. It is known that for comparably large magnitude of this term(s) in the system, standard numerical schemes such as Galerkin finite element method usually become unstable and simply fail to produce desirable results, therefore a proper stabilization technique need to be applied. In this regard, we considered several types of cross-diffusion systems (which will be thoroughly studied in the following chapters). To enhance the stability of the standard Galerkin method in use, we tried different stabilization methods as mentioned in the previous section, we noticed that traditional stabilized techniques such as SUPG, GLS and USFEM fail to overcome the instability resulting from the Galerkin formulation and huge amount of under- and over-shoots was observed that led to blow-up in the numerical simulations. However, approaches which were based on the algebraic stabilization worked remarkably well by reducing the spurious oscillations and also preserving the positivity of solutions through the time. Therefore, due to the obtained desirable results using algebraic stabilization techniques only these type of methods will be considered in the following chapters.

## 2. Paper I

This chapter is based on the paper entitled "Global existence of classical solutions and numerical simulations of a cancer invasion model", published in Mathematical Modeling and Numerical Analysis (ESAIM: M2AN).

### 2.1 Global existence of classical solutions and numerical simulations of a cancer invasion model

In this paper we considered a model of cancer invasion which is described by a system of nonlinear PDEs consisting of a cross-diffusion-reaction equation and two additional nonlinear ordinary differential equations as:

$$\begin{cases} u_t = \frac{1}{\alpha}\Delta u - \chi\nabla \cdot (u\nabla c) + \mu u(1 - u) & \text{in } \Omega \times (0, \infty), \\ c_t = -pc & \text{in } \Omega \times (0, \infty), \\ p_t = \frac{1}{\epsilon}(uc - p) & \text{in } \Omega \times (0, \infty), \\ \frac{1}{\alpha}\partial_\nu u = \chi_\nu \partial_\nu c & \text{on } \partial\Omega \times (0, \infty), \\ (u, c, p)(\cdot, 0) = (u_0, c_0, p_0) & \text{in } \Omega, \end{cases} \quad (2.1)$$

where  $\mu, \chi, \alpha, \epsilon$  are positive constants,  $u_0, c_0, p_0$  are initial conditions, and  $u, c, p$  denote the concentrations of the invasive cancer cells, extracellular matrix and protease, respectively. Let us summarize the idea behind the definition of the system:

- The cancer cells spread isotropically inside the domain  $\Omega$ : term  $\frac{1}{\alpha}\Delta u$ ,
- The cancer cells grow with a proliferation rate  $\mu$ : term  $\mu u(1 - u)$ ,
- The cancer cells move spatially toward the higher concentration of the extracellular matrix : term  $-\chi\nabla \cdot (u\nabla c)$ ,
- The degradation of the extracellular matrix: term  $-pc$
- The protease production, which is a result of interaction between invading tumor and the connecting tissue: term  $\frac{1}{\epsilon}uc$ ,
- The natural decay of the extracellular matrix: term  $-\frac{1}{\epsilon}p$ .

This system is a variant of a cancer invasion model developed in [242] for the malignant invasion of tumor, where we added the extra chemotaxis movement  $\frac{1}{\alpha}\Delta u$  which allowed us to study the properties of the model analytically.

In the first part of the paper, we proved under several transformations of the system, which allowed us to control the problematic taxis term  $-\chi\nabla \cdot (u\nabla c)$ , that the system (2.1) possesses global classical solutions for widely arbitrary initial data in two- and three-dimensional space. This led to our main analytical result:



**Theorem.** *Suppose that  $\alpha, \chi, \mu, \epsilon$  are positive constants, that*

*$\Omega$  is a smooth bounded domain in  $\mathbb{R}^d$ ,  $d \in \{1, 2, 3\}$ ,*

*and that  $u_0, c_0, p_0 \in \bigcup_{\gamma \in (0,1)} C^{2,\gamma}(\overline{\Omega})$  are non-negative functions satisfying*

$$\frac{1}{\alpha} \partial_\nu u_0 = \chi_\nu u_0 \partial_\nu c_0 \quad \text{on } \partial\Omega.$$

*Then, there exists a unique global classical solution  $(u, c, p)$  of (2.1) with regularity  $(u, c, p) \in (C^{2,1}(\overline{\Omega} \times (0, T]) \cap C^1(\overline{\Omega} \times [0, T]))^3$ , which, moreover, is non-negative.*

In the next part of the paper, we analyzed the behavior of these solutions numerically. For this purpose, we utilized Galerkin method for spatial discretization,  $\theta$ -method for time discretization, and fixed-point iteration to treat the nonlinear terms. We carried out several numerical experiments in two and three spatial dimensions which support our analytical results. Moreover, we addressed the numerical stability of the system and showed that it heavily relies on the choices of the haptotactic coefficient  $\chi$ . Fixing the proliferation rate  $\mu$  and varying the taxis coefficient  $\chi$  one can easily make the diffusion or the transport of the cancer cells dominant. The latter usually gives rise to spurious oscillations or numerical blow up in the system, in which case a stabilization method is required to prevent these difficulties. This observation led to our results to be presented in the next chapter.

## GLOBAL EXISTENCE OF CLASSICAL SOLUTIONS AND NUMERICAL SIMULATIONS OF A CANCER INVASION MODEL

MARIO FUEST<sup>1,\*</sup>, SHAHIN HEYDARI<sup>2</sup>, PETR KNOBLOCH<sup>2</sup>,  
JOHANNES LANKEIT<sup>1</sup> AND THOMAS WICK<sup>1</sup>

**Abstract.** In this paper, we study a cancer invasion model both theoretically and numerically. The model is a nonstationary, nonlinear system of three coupled partial differential equations modeling the motion of cancer cells, degradation of the extracellular matrix, and certain enzymes. We first establish existence of global classical solutions in both two- and three-dimensional bounded domains, despite the lack of diffusion of the matrix-degrading enzymes and corresponding regularizing effects in the analytical treatment. Next, we give a weak formulation and apply finite differences in time and a Galerkin finite element scheme for spatial discretization. The overall algorithm is based on a fixed-point iteration scheme. Our theory and numerical developments are accompanied by some simulations in two and three spatial dimensions.

**Mathematics Subject Classification.** 35A01, 35K57, 35Q92, 65M22, 65M60, 92C17.

Received May 17, 2022. Accepted April 28, 2023.

### 1. INTRODUCTION

#### The model

One of the defining characteristics of a malignant tumour is its capability to invade adjacent tissues [21]; accordingly the mathematical literature directed at understanding underlying mechanisms is vast (see *e.g.* the surveys [29, 36]).

In this paper we focus on the following variant of a cancer invasion model developed by Perumpanani *et al.* [34] for the malignant invasion of tumours and investigate

$$\begin{cases} u_t = \frac{1}{\alpha} \Delta u - \chi \nabla \cdot (u \nabla c) + \mu u(1 - u) & \text{in } \Omega \times (0, \infty), \\ c_t = -pc & \text{in } \Omega \times (0, \infty), \\ p_t = \frac{1}{\varepsilon} (uc - p) & \text{in } \Omega \times (0, \infty), \\ \frac{1}{\alpha} \partial_\nu u = \chi u \partial_\nu c & \text{on } \partial\Omega \times (0, \infty), \\ (u, c, p)(\cdot, 0) = (u_0, c_0, p_0) & \text{in } \Omega. \end{cases} \quad (1.1)$$

---

*Keywords and phrases.* Haptotaxis, Tumour invasion, Global existence, Fixed-point scheme, Numerical simulations.

<sup>1</sup> Leibniz University Hannover, Institute of Applied Mathematics, Welfengarten 1, 30167 Hannover, Germany.

<sup>2</sup> Charles University, Faculty of Mathematics and Physics, Sokolovská 83, 18675 Praha 8, Czech Republic.

\*Corresponding author: [fuest@ifam.uni-hannover.de](mailto:fuest@ifam.uni-hannover.de)

We aim for a rigorous existence proof for global solutions and the development of a reliable numerical scheme with an implementation in a modern open-source finite element library.

In (1.1), the motion of cancer cells (density denoted by  $u$ ) mainly takes place by means of haptotaxis, *i.e.* directed motion toward higher concentrations of extracellular matrix (density  $c$ ), of strength  $\chi \geq 0$ . Motivated by experiments of Aznavoorian *et al.* [7], who reported only “a minor chemokinetic component” of the cell motion, the original model of [34] does not include a term for random (chemokinetic) cell motility at all. Acknowledging that “minor” does not mean “none at all”, we deviate from [34] in this aspect and incorporate this motility term in (1.1) ( $\alpha \in (0, \infty)$ , with the formal limit  $\alpha \rightarrow \infty$  corresponding to the model of [34]). Additional growth of the population of tumour cells is described by a logistic term (with  $\mu$  being a positive parameter). The extracellular matrix is degraded upon contact with certain enzymes (proteases, concentration  $p$ ), which, in turn, are produced where cancer cells and matrix meet and decay over time. The reaction speed of these protein dynamics can be adjusted via the parameter  $\varepsilon > 0$ . As many proteases remain bound to the cellular membrane – or are only activated when on the cell surface (*cf.* the model derivation in [34]), no diffusion for  $p$  is incorporated in the model. This last point is in contrast to the otherwise similar popular models in the tradition of [4, 33] or [9], the latter of which additionally included a chemotactic component of the motion of cancer cells.

## Global solvability

In order to construct global classical solutions of (1.1), it is necessary to control the haptotaxis term  $-\chi \nabla \cdot (u \nabla c)$  in the first equation and thus in particular to gain information on the spatial derivative of the second solution component. For relatives of (1.1) including a diffusion term  $\Delta p$  in the third equation, this has already been achieved in [40] and [28] by applying parabolic regularity theory to the equation for  $p$ , first yielding estimates for the spatial derivative of  $p$  and then also on  $c$ ; the results of [28] even cover the long-term asymptotics of solutions. Moreover, the presence of diffusion for the produced quantities has also been made use of to obtain global existence results for different cancer invasion models, see for instance [42].

However, the absence of any spatial regularization in both the second and third equation makes the corresponding analysis much more challenging. Up to now, global classical solutions have only been constructed for a rather limited set of initial data: already in [34], where (1.1) has been proposed for  $\alpha = \infty$ , it has been shown that the model formally obtained by taking the limit  $\varepsilon \searrow 0$  admits a family of travelling wave solutions. Corresponding results for positive  $\varepsilon$  have then been achieved in [31]. Moreover, if  $\varepsilon = 0$ , travelling wave solutions may contain shocks [30] and solutions of related systems without a logistic source may even blow up in finite time [35]. In general, the destabilizing effect of taxis terms such as  $-\chi \nabla \cdot (u \nabla c)$  may not only make it challenging but even impossible to obtain global existence results for certain problems. We refer to the survey [25] for further discussion regarding the consequences of low regularity in chemotaxis systems.

Despite these challenges, in the first part of the present paper we are able to give an affirmative answer to the question whether (1.1) also possesses global classical solutions for widely arbitrary initial data in the two and three-dimensional setting. Our analytical main result is the following

**Theorem 1.1.** *Suppose that  $\alpha, \chi, \mu, \varepsilon$  are positive constants, that*

$$\Omega \text{ is a smooth bounded domain in } \mathbb{R}^n, n \in \{1, 2, 3\},$$

*and that  $u_0, c_0, p_0 \in \bigcup_{\gamma \in (0, 1)} C^{2+\gamma}(\overline{\Omega})$  are nonnegative and such that  $\frac{1}{\alpha} \partial_\nu u_0 = \chi u_0 \partial_\nu c_0$  on  $\partial\Omega$ . Then there exists a unique global classical solution  $(u, c, p)$  of (1.1) with regularity*

$$(u, c, p) \in (C^{2,1}(\overline{\Omega} \times (0, \infty)) \cap C^1(\overline{\Omega} \times [0, \infty)))^3,$$

*which, moreover, is nonnegative.*

### Numerical modeling

In the second part of our paper we then analyze the behavior of these solutions numerically with an implementation in the modern open-source finite element library deal.II [5, 6]. Related numerical studies in various software libraries using different numerical schemes are briefly described in the following. The traditional method of lines has been widely used for simulations of the cancer invasion process [15, 19]. In addition, finite difference methods [23] have been considered and in [10, 22], the authors proposed a nonstandard finite difference method which satisfies the positivity-preservation of the solution, that is an important property in the stability of the model. Moreover, the finite volume method [11], spectral element methods [41], algebraically stabilized finite element method [37], the discontinuous Galerkin method [16], combinations of level-set/adaptive finite elements [1, 44], and a hybrid finite volume/finite element method [2, 8] have also been proposed in the literature for some cancer invasion models and chemotaxis. Finally, we mention that in [38] the authors illustrate their theoretical results for a related cancer model employing discontinuous Galerkin finite elements implemented as well in deal.II.

The main objective in the numerical part is the design of reliable algorithms for (1.1) and their corresponding implementation in deal.II. First, we discretize in time using a  $\theta$ -method, which allows for implicit  $A$ -stable time discretizations. Then, a Galerkin finite element scheme is employed for spatial discretization. The nonlinear discrete system of equations is decoupled by designing a fixed-point algorithm. This algorithm is newly designed and then implemented and debugged in deal.II.

These developments then allow to link our theoretical part and the numerical sections in order to carry out various numerical simulations to complement Theorem 1.1. Specifically, several parameter variations of the proliferation coefficient  $\mu$  and the haptotactic coefficient  $\chi$  will be studied in two- and three spatial dimensions. These studies are non-trivial due to the nonlinearities and the high sensitivity of (1.1) with respect to such parameter variations.

### Plan of the paper

The outline of this paper is as follows. In Section 2, we study the global existence of classical solutions. Next, in Section 3, we introduce the discretization in time and space using finite differences in time and a Galerkin finite element scheme in space. We also describe the solution algorithm. In Section 4, we carry out several numerical simulations demonstrating the properties of our model and the corresponding theoretical results. Therein, we specifically study parameter variations. Finally, our work is summarized in Section 5.

### Notation

Let  $\Omega \subset \mathbb{R}^n$ ,  $n \in \mathbb{N}$  be a bounded domain. By  $L^p(\Omega)$  and  $W^{1,p}(\Omega)$ , we denote the usual Lebesgue and Sobolev spaces, respectively, and we abbreviate  $H^1(\Omega) := W^{1,2}(\Omega)$ . Furthermore,  $\langle \cdot, \cdot \rangle$  denotes the duality product between  $(H^1)^*$  and  $H^1$ .

For  $m \in \mathbb{N}_0$  and  $\gamma \in (0, 1)$ , we denote by  $C^{m+\gamma}(\bar{\Omega})$  the space of functions  $\varphi \in C^m(\bar{\Omega})$  with finite norm

$$\|\varphi\|_{C^{m+\gamma}(\bar{\Omega})} := \|\varphi\|_{C^m(\bar{\Omega})} + \sup_{x,y \in \bar{\Omega}, x \neq y} \frac{|\varphi(x) - \varphi(y)|}{|x - y|^\gamma}.$$

Moreover, for  $m_1, m_2 \in \mathbb{N}_0$ ,  $\gamma_1, \gamma_2 \in [0, 1)$  and  $T > 0$ , we denote by  $C^{m_1+\gamma_1, m_2+\gamma_2}(\bar{\Omega} \times [0, T])$  the space of all functions  $\varphi$  whose derivatives  $D_x^\alpha D_t^\beta \varphi$ ,  $|\alpha| \leq m_1$ ,  $0 \leq \beta \leq m_2$ , (exist and) are continuous, and which have finite norm

$$\|\varphi\|_{C^{m_1+\gamma_1, m_2+\gamma_2}(\bar{\Omega} \times [0, T])} := \sum_{\substack{|\alpha| \leq m_1, \\ 0 \leq \beta \leq m_2}} \left\| D_x^\alpha D_t^\beta \varphi \right\|_{C^0(\bar{\Omega} \times [0, T])}$$

$$\begin{aligned}
 & + \sum_{\substack{|\alpha|=m_1, \\ 0 \leq \beta \leq m_2}} \sup_{\substack{x, y \in \bar{\Omega}, x \neq y, \\ t \in [0, T]}} \frac{|D_x^\alpha D_t^\beta \varphi(x, t) - D_x^\alpha D_t^\beta \varphi(y, t)|}{|x - y|^{\gamma_1}} \\
 & + \sum_{\substack{|\alpha| \leq m_1, \\ \beta = m_2}} \sup_{\substack{x \in \bar{\Omega}, \\ s, t \in [0, T], s \neq t}} \frac{|D_x^\alpha D_t^\beta \varphi(x, t) - D_x^\alpha D_t^\beta \varphi(x, s)|}{|t - s|^{\gamma_2}}.
 \end{aligned}$$

Notationally, we do not distinguish between spaces of scalar- and vector-valued functions.

## 2. GLOBAL EXISTENCE OF CLASSICAL SOLUTIONS

As a first step in the proof of Theorem 1.1, we apply two transformations in Subsection 2.1; the first one allows us to get rid of some parameters in (1.1), the second one changes the first equation to a more convenient form. We then employ a fixed point argument to obtain a local existence result for the transformed system in Lemma 2.5.

The proof that these solutions are global in time consists of two key parts, both relying on the fact that the second and third equation in (1.1) at least regularize in time (which allows us to prove Lemma 2.7 and Lemma 2.10). First, in order to prove boundedness in  $L^\infty$ , the comparison principle allows us to conclude boundedness in small time intervals (cf. Lemma 2.8). We then iteratively apply this bound to obtain the result also for larger times (cf. Lemma 2.9). As to bounds for the spatial derivatives, we secondly apply a testing procedure to derive estimates valid on small time intervals (cf. Lemma 2.11), which then is again complemented by an iteration procedure (cf. Lemma 2.12). Finally, we are able to make use of parabolic regularity theory (inter alia in the form of maximal Sobolev regularity) to conclude in Lemma 2.14 that the solutions exist globally.

### 2.1. Two transformations

We first note that with regards to Theorem 1.1 we may without loss of generality assume  $\chi = 1$  and  $\varepsilon = 1$ . Indeed, suppose that Theorem 1.1 holds for this special case. Then, assuming the conditions of Theorem 1.1 to hold, we set

$$\tilde{\alpha} := \frac{\alpha\chi}{\varepsilon}, \quad \tilde{\chi} := 1, \quad \tilde{\mu} := \varepsilon\mu, \quad \tilde{\varepsilon} := 1$$

and further

$$\tilde{u}_0(\tilde{x}) = u_0(\sqrt{\chi}\tilde{x}), \quad \tilde{c}_0(\tilde{x}) = \varepsilon c_0(\sqrt{\chi}\tilde{x}), \quad \tilde{p}_0(\tilde{x}) = \varepsilon p_0(\sqrt{\chi}\tilde{x})$$

for  $\tilde{x} \in \tilde{\Omega} := \frac{1}{\sqrt{\chi}}\Omega$ . By Theorem 1.1, there then exists a global classical solution of (1.1) (with all parameters and initial data replaced by their pendants with tildes)  $(\tilde{u}, \tilde{c}, \tilde{p})$ . Then

$$(u, c, p)(x, t) := \left( \tilde{u}\left(\frac{x}{\sqrt{\chi}}, \frac{t}{\varepsilon}\right), \frac{1}{\varepsilon}\tilde{c}\left(\frac{x}{\sqrt{\chi}}, \frac{t}{\varepsilon}\right), \frac{1}{\varepsilon}\tilde{p}\left(\frac{x}{\sqrt{\chi}}, \frac{t}{\varepsilon}\right) \right), \quad (x, t) \in (\bar{\Omega} \times [0, \infty)), \tag{2.1}$$

fulfills

$$\begin{cases} u_t = \frac{\chi}{\tilde{\alpha}\tilde{\varepsilon}}\Delta u - \frac{\varepsilon\chi}{\tilde{\varepsilon}}\nabla \cdot (u\nabla c) + \frac{\tilde{\mu}}{\tilde{\varepsilon}}u(1 - u) & \text{in } \Omega \times (0, \infty), \\ c_t = -\frac{\tilde{\varepsilon}}{\varepsilon^2}cp & \text{in } \Omega \times (0, \infty), \\ p_t = \frac{1}{\varepsilon^2}(\varepsilon uc - \varepsilon p) & \text{in } \Omega \times (0, \infty), \\ \partial_\nu u = \frac{\tilde{\alpha}\varepsilon\sqrt{\chi}}{\sqrt{\chi}}u\partial_\nu c & \text{on } \partial\Omega \times (0, \infty), \\ (u, c, p)(x, 0) = (\tilde{u}_0, \frac{1}{\varepsilon}\tilde{c}_0, \frac{1}{\varepsilon}\tilde{p}_0)\left(\frac{x}{\sqrt{\chi}}\right) & \text{for } x \in \Omega \end{cases}$$

and thus (1.1). Moreover, following precedents from, e.g. [13, p.19], we set

$$w(x, t) := u(x, t)e^{-\alpha c(x, t)}, \quad x \in \bar{\Omega}, \quad t \geq 0. \tag{2.2}$$

Then  $\nabla w = e^{-\alpha c} \nabla u - \alpha e^{-\alpha c} u \nabla c$ , so that

$$\frac{1}{\alpha} \Delta w + \nabla c \cdot \nabla w = e^{-\alpha c} \nabla \cdot \left( \frac{1}{\alpha} e^{\alpha c} \nabla w \right) = e^{-\alpha c} \left[ \frac{1}{\alpha} \Delta u - \nabla \cdot (u \nabla c) \right] \tag{2.3}$$

and (1.1) (with  $\chi = \varepsilon = 1$ ) is equivalent to

$$\begin{cases} w_t = \frac{1}{\alpha} \Delta w + \nabla c \cdot \nabla w + \alpha p c w + \mu w - \mu e^{\alpha c} w^2 & \text{in } \Omega \times (0, \infty), \\ c_t = -p c & \text{in } \Omega \times (0, \infty), \\ p_t = w e^{\alpha c} c - p & \text{in } \Omega \times (0, \infty), \\ \partial_\nu w = 0 & \text{on } \partial\Omega \times (0, \infty), \\ (w, c, p)(\cdot, 0) = (w_0, c_0, p_0) & \text{in } \Omega \end{cases} \tag{2.4}$$

for  $w_0 := u_0 e^{-\alpha c_0}$ . Expanding  $\nabla \cdot (u \nabla c)$  to  $\nabla u \cdot \nabla c + u \Delta c$  shows that this transformation allows us to get rid of a term involving  $\Delta c$  in the first equation at the price of adding several zeroth order terms. In particular as the second equation does not regularize in space, (2.4) turns out to be a more convenient form for the following analysis.

### 2.2. Local existence

In this subsection, we construct maximal classical solutions of (2.4) in  $\bar{\Omega} \times [0, T_{\max})$  for some  $T_{\max} \in [0, \infty)$  by means of a fixed point argument. Moreover, we provide a criterion for when these solutions are global in time (that is, when  $T_{\max} = \infty$  holds), which then will finally be seen to hold true in Lemma 2.14.

As a preparation, we first collect results on (Hölder) continuous dependency of solutions to ODEs on the data.

**Lemma 2.1.** *Let  $\Omega \subset \mathbb{R}^n$ ,  $n \in \mathbb{N}$ , be a bounded domain,  $T > 0$ ,  $d \in \mathbb{N}$ ,  $\gamma_1, \gamma_2 \in [0, 1)$ ,  $v_0 \in C^{\gamma_1}(\bar{\Omega})$  and assume that  $f: \bar{\Omega} \times [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$  is  $(\gamma_1, \gamma_2)$ -Hölder continuous with respect to its first two arguments and locally Lipschitz continuous w.r.t. the third variable, in the sense that for every compact  $K \subset \mathbb{R}^d$  there is  $L > 0$  such that*

$$\begin{aligned} \sup_{t \in [0, T], v \in K} \|f(\cdot, t, v)\|_{C^{\gamma_1}(\bar{\Omega})} &\leq L, \\ \sup_{x \in \bar{\Omega}, v \in K} \|f(x, \cdot, v)\|_{C^{\gamma_2}([0, T])} &\leq L, \\ \sup_{(x, t) \in \bar{\Omega} \times [0, T]} |f(x, t, v) - f(x, t, w)| &\leq L|v - w| \quad \text{for all } v, w \in K. \end{aligned}$$

Then for any compact  $K \subset \mathbb{R}^d$  there is  $C > 0$  such that whenever  $v: \bar{\Omega} \times [0, T] \rightarrow \mathbb{R}^d$  is such that  $v(x, \cdot) \in C^0([0, T]) \cap C^1((0, T))$  for all  $x \in \bar{\Omega}$ ,  $v$  solves

$$v_t(x, t) = f(x, t, v(x, t)) \quad \text{for all } x \in \bar{\Omega}, t \in (0, T); \quad v(x, 0) = v_0(x) \quad \text{for all } x \in \bar{\Omega} \tag{2.5}$$

and satisfies  $v(\bar{\Omega} \times [0, T]) \subset K$ , then  $v, v_t \in C^{\gamma_1, \gamma_2}(\bar{\Omega} \times [0, T])$  and

$$\|v\|_{C^{\gamma_1, 1+\gamma_2}(\bar{\Omega} \times [0, T])} \leq C.$$

*Proof.* First, we let  $K$  be a compact superset of  $v(\bar{\Omega} \times [0, T])$  and let  $L$  be as in the assumptions on  $f$ . We introduce  $\omega_1 \in C^0([0, \infty))$  such that  $|f(x, t, w) - f(y, t, w)| \leq \omega_1(|x - y|)$  for all  $x \in \bar{\Omega}$ ,  $y \in \bar{\Omega}$ ,  $t \in [0, T]$  and  $w \in K$ ,  $|v_0(x) - v_0(y)| \leq \omega_1(|x - y|)$  for all  $x \in \bar{\Omega}$ ,  $y \in \bar{\Omega}$  and such that  $\omega_1(0) = 0$  and  $\sup_{r>0} r^{-\gamma_1} \omega_1(r) < \infty$ . We then fix  $x \in \bar{\Omega}$ ,  $y \in \bar{\Omega} \setminus \{x\}$  and let  $\tilde{v}(t) = (v(x, t) - v(y, t)) \cdot \frac{1}{\omega_1(|x - y|)}$ . Then

$$\tilde{v}_t(t) = \frac{1}{\omega_1(|x - y|)} (f(x, t, v(x, t)) - f(y, t, v(x, t))) + \frac{1}{\omega_1(|x - y|)} (f(y, t, v(x, t)) - f(y, t, v(y, t)))$$

$$\leq 1 + \frac{L}{\omega_1(|x-y|)} |v(x,t) - v(y,t)| \quad \text{for all } t \in (0, T),$$

so that  $\tilde{v}_t \leq 1 + L|\tilde{v}|$  and, analogously,  $\tilde{v}_t \geq -1 - L|\tilde{v}|$ , so that boundedness of  $|\tilde{v}|$  results from Grönwall's inequality and hence  $v \in C^{\gamma_1, 0}(\bar{\Omega} \times [0, T])$ , and  $\sup_{t \in [0, T]} \|v(\cdot, t)\|_{C^{\gamma_1}(\bar{\Omega})}$  is bounded due to the choice of  $\omega_1$ .

For  $\tau > 0$ , we treat  $\bar{v}(x, t) = (v(x, t) - v(x, t + \tau))/\omega_2(\tau)$  with some  $\omega_2$  such that  $|f(x, t, v) - f(x, t + \tau, v)| \leq \omega_2(\tau)$  in the same way. This ensures the claimed regularity of  $v$ , whereupon that of  $v_t$  follows from  $v_t(x, t) = f(x, t, v(x, t))$ ,  $(x, t) \in \bar{\Omega} \times (0, T)$ , and continuity of the right-hand side up to  $t = 0$  and  $t = T$ .  $\square$

**Lemma 2.2.** *In addition to the assumptions of Lemma 2.1, let  $m \in \mathbb{N}$  and  $v_0 \in C^{m+\gamma_1}(\bar{\Omega})$ . If all derivatives of  $f$  w.r.t.  $x$  and  $v$  up to order  $m$  satisfy the conditions Lemma 2.1 poses on  $f$ , then any solution  $v$  as in Lemma 2.1 belongs to  $C^{m+\gamma_1, 1+\gamma_2}(\bar{\Omega} \times [0, T])$ .*

*Proof.* For  $i \in \{1, \dots, n\}$ ,  $\tilde{v} = \partial_{x_i} v$  satisfies

$$\tilde{v}_t = f_{x_i}(x, t, v(x, t)) + f_v(x, t, v(x, t))\tilde{v} \quad \text{in } \Omega \times (0, T), \quad \tilde{v}(\cdot, 0) = (v_0)_{x_i} \quad \text{in } \Omega$$

and Lemma 2.1 can be applied to  $\tilde{v}$ . An inductive argument takes care of higher derivatives.  $\square$

By applying these results to the ODEs appearing in (2.4), we obtain

**Lemma 2.3.** *Let  $\Omega \subset \mathbb{R}^n$ ,  $n \in \mathbb{N}$ , be a bounded domain,  $\alpha \geq 0$  and  $T \in (0, \infty)$ . Let  $0 \leq w \in C^0(\bar{\Omega} \times [0, T])$  and  $c_0, p_0 \in C^0(\bar{\Omega}; [0, \infty))$ .*

(a) *Then*

$$\begin{cases} c_t = -pc \\ p_t = we^{\alpha c}c - p \end{cases} \quad (2.6)$$

*has a unique solution  $(c, p) \in (C^{0,1}(\bar{\Omega} \times [0, T]))^2$ .*

(b) *For any  $M > 0$  there is  $C = C(M) > 0$  such that whenever  $\tilde{T} \in (0, T]$ ,  $w \in C^{1,0}(\bar{\Omega} \times [0, \tilde{T}])$ ,  $c_0, p_0 \in C^1(\bar{\Omega})$  with*

$$\sup_{t \in [0, \tilde{T}]} \|w(\cdot, t)\|_{C^1(\bar{\Omega})} \leq M, \quad \|c_0\|_{C^1(\bar{\Omega})} \leq M, \quad \|p_0\|_{C^1(\bar{\Omega})} \leq M,$$

*then*

$$\|(c, p)\|_{C^1(\bar{\Omega} \times [0, \tilde{T}])} \leq C.$$

(c) *If, for some  $k \in \mathbb{N}_0$  and  $\gamma_1, \gamma_2 \in [0, 1)$ ,  $w \in C^{k+\gamma_1, \gamma_2}(\bar{\Omega} \times [0, T])$  and  $c_0, p_0 \in C^{k+\gamma_1}(\bar{\Omega})$ , then  $c, p \in C^{k+\gamma_1, 1+\gamma_2}(\bar{\Omega} \times [0, T])$ .*

*Proof.* (a) Given  $w \in C^0(\bar{\Omega} \times [0, T])$ , for every  $x \in \bar{\Omega}$  the existence and uniqueness of a solution  $(c, p)(x, \cdot) \in C^0([0, T_{\max}(x)]) \cap C^1((0, T_{\max}(x)))$  of (2.6), with some  $T_{\max}(x) \in (0, T]$  such that  $\limsup_{t \nearrow T_{\max}(x)} (|c(x, t)| + |p(x, t)|) = \infty$  or  $T_{\max}(x) = T$ , follows from Picard–Lindelöf's theorem. By an ODE comparison argument, nonnegativity of  $c$  follows from that of  $c_0$ , and nonnegativity of  $p$  from that of  $p_0, w$  and  $c$ . Therefore,  $0 \leq c(x, t) \leq c_0(x)$  for all  $x \in \bar{\Omega}$  and  $t \in (0, T_{\max}(x))$  due to the sign of  $c_t = -pc$ . Consequently,

$$\begin{aligned} p(x, t) &= e^{-t} p_0(x) + \int_0^t e^{-(t-s)} w(x, s) e^{\alpha c(x, s)} c(x, s) \, ds \\ &\leq \max \left\{ p_0(x), c_0(x) e^{\alpha c_0(x)} \max_{s \in [0, T]} w(x, s) \right\} \quad \text{for all } x \in \bar{\Omega}, t \in [0, T_{\max}(x)), \end{aligned}$$

which also shows that  $T_{\max}(x) = T$  for all  $x \in \bar{\Omega}$ . The remainder of part (a) follows from an application of Lemma 2.1 with  $\gamma_1 = \gamma_2 = 0$ .

(b) We apply Lemma 2.1 to  $v = (\nabla c, \nabla p)$ , the solution of

$$v_t = \begin{pmatrix} -pI_{n \times n} & -cI_{n \times n} \\ (\alpha w e^{\alpha c} c + w e^{\alpha c})I_{n \times n} & -I_{n \times n} \end{pmatrix} v + \begin{pmatrix} 0 \\ e^{\alpha c} c \nabla w \end{pmatrix}, \quad v(0) = \begin{pmatrix} \nabla c_0 \\ \nabla p_0 \end{pmatrix} \in \mathbb{R}^{2n},$$

noting that sufficient regularity is given by the assumption on  $w$  and part (a).

(c) Follows from Lemma 2.2. □

Both as an ingredient to the proof of Lemma 2.5 below and also for its own interest, we note that classical solutions of (2.4) are unique.

**Lemma 2.4.** *Let  $\Omega \subset \mathbb{R}^n$ ,  $n \in \mathbb{N}$ , a bounded domain and let  $\alpha, \mu > 0$  as well as  $\chi = \varepsilon = 1$ . Suppose  $(u_1, c_1, p_1)$  and  $(u_2, c_2, p_2)$  are two solutions of (1.1) with the same initial data  $(u_0, c_0, p_0) \in C^0(\bar{\Omega}) \times C^1(\bar{\Omega}) \times C^1(\bar{\Omega})$  and assume that they belong to*

$$(C^{2,1}(\bar{\Omega} \times (0, T)) \cap C^1(\bar{\Omega} \times [0, T]))^3$$

for some  $T > 0$ . Then  $(u_1, c_1, p_1) = (u_2, c_2, p_2)$  in  $\bar{\Omega} \times [0, T]$ .

*Proof.* We let  $C > 0$  be such that

$$\max \left\{ \|u_i(\cdot, t)\|_{L^\infty(\Omega)}, \|\nabla u_i(\cdot, t)\|_{L^\infty(\Omega)}, \|c_i(\cdot, t)\|_{L^\infty(\Omega)}, \|p_i(\cdot, t)\|_{L^\infty(\Omega)}, \|\nabla c_i(\cdot, t)\|_{L^\infty(\Omega)}, \|\nabla p_i(\cdot, t)\|_{L^\infty(\Omega)} \right\} \leq C$$

for all  $t \in [0, T]$  and  $i \in \{1, 2\}$ . Then computing

$$\begin{aligned} & \frac{1}{2} \frac{d}{dt} \left( \int_{\Omega} (u_1 - u_2)^2 + \int_{\Omega} (c_1 - c_2)^2 + \int_{\Omega} (p_1 - p_2)^2 + \int_{\Omega} |\nabla(c_1 - c_2)|^2 + \int_{\Omega} |\nabla(p_1 - p_2)|^2 \right) \\ &= -\frac{1}{\alpha} \int_{\Omega} |\nabla(u_1 - u_2)|^2 + \int_{\Omega} \nabla(u_1 - u_2)[(u_1 - u_2)\nabla c_1 + u_2\nabla(c_1 - c_2)] + \mu \int_{\Omega} (u_1 - u_2)^2 \\ & \quad - \mu \int_{\Omega} (u_1 - u_2)^2(u_1 + u_2) - \int_{\Omega} (c_1 - c_2)(p_1 - p_2)c_1 - \int_{\Omega} (c_1 - c_2)^2 p_2 - \int_{\Omega} (p_1 - p_2)^2 \\ & \quad + \int_{\Omega} (p_1 - p_2)u_1(c_1 - c_2) + \int_{\Omega} (p_1 - p_2)(u_1 - u_2)c_2 - \int_{\Omega} \nabla(c_1 - c_2)\nabla(p_1 - p_2)c_1 \\ & \quad - \int_{\Omega} \nabla(c_1 - c_2)(p_1 - p_2)\nabla c_1 - \int_{\Omega} p_2|\nabla(c_1 - c_2)|^2 - \int_{\Omega} \nabla(c_1 - c_2)(c_1 - c_2)\nabla p_2 \\ & \quad - \int_{\Omega} |\nabla(p_1 - p_2)|^2 + \int_{\Omega} \nabla(p_1 - p_2)(c_1 - c_2)\nabla u_1 + \int_{\Omega} \nabla(p_1 - p_2)u_1\nabla(c_1 - c_2) \\ & \quad + \int_{\Omega} \nabla(p_1 - p_2)\nabla(u_1 - u_2)c_2 + \int_{\Omega} \nabla(p_1 - p_2)(u_1 - u_2)\nabla c_2 \\ & \leq ((2 + \alpha C)C + \mu + 2C\mu) \int_{\Omega} (u_1 - u_2)^2 + 5C \int_{\Omega} (c_1 - c_2)^2 + (4C - 1) \int_{\Omega} (p_1 - p_2)^2 \\ & \quad + (5 + \alpha C)C \int_{\Omega} |\nabla(c_1 - c_2)|^2 + ((4 + \alpha C)C - 1) \int_{\Omega} |\nabla(p_1 - p_2)|^2 \quad \text{in } (0, T), \end{aligned}$$

we obtain  $(u_1, c_1, p_1) = (u_2, c_2, p_2)$  from Grönwall's inequality. □

Making use of Schauder's fixed point theorem and applying Lemmas 2.1–2.4, we now obtain a local existence result for the system (2.4).



**Lemma 2.5.** *Assume that*

$$\Omega \text{ is a smooth bounded domain in } \mathbb{R}^n, n \in \{1, 2, 3\}, \alpha, \mu > 0 \quad (2.7)$$

and that

$$w_0, c_0, p_0 \in C^{2+\gamma}(\bar{\Omega}) \text{ for some } \gamma \in (0, 1) \text{ are nonnegative and fulfill } \partial_\nu w_0 = 0 \text{ on } \partial\Omega. \quad (2.8)$$

Then (2.4) has a nonnegative unique solution

$$(w, c, p) \in \left( C^{2+\gamma, 1+\frac{\gamma}{2}}(\bar{\Omega} \times (0, T_{\max})) \cap C^1(\bar{\Omega} \times [0, T_{\max})) \right)^3$$

for some  $T_{\max} > 0$ , which can be chosen such that

$$T_{\max} = \infty \text{ or } \limsup_{t \nearrow T_{\max}} \|w(\cdot, t)\|_{C^{1+\gamma}(\bar{\Omega})} = \infty. \quad (2.9)$$

*Proof.* For  $T > 0$  and  $M > 0$  we introduce the set

$$S_{M,T} = \left\{ w \in C^0([0, T]; C^1(\bar{\Omega})) \mid 0 \leq w, \sup_{t \in [0, T]} \|w(\cdot, t)\|_{C^1(\bar{\Omega})} \leq M \right\}$$

and given  $w \in S_{M,T}$  we let  $(c, p)$  be the solution of (2.6) (cf. Lemma 2.3(a)). We then let  $v$  be the unique (weak) solution (see [24, Thm. III.5.1], [27, Thm. 6.39]) of

$$v_t = \frac{1}{\alpha} \Delta v + f \cdot \nabla v + gv \text{ in } \Omega \times (0, T), \quad \partial_\nu v = 0 \text{ on } \partial\Omega \times (0, T), \quad v(\cdot, 0) = w_0 \text{ in } \Omega, \quad (2.10)$$

where  $f = \nabla c$  and  $g = \alpha pc + \mu - \mu e^{\alpha c} w$  belong to  $L^\infty(\Omega \times (0, T))$  according to Lemma 2.3(a) and (b), and define  $\Phi(w) = v$  in  $\bar{\Omega} \times [0, T]$ . We choose  $M > 0$  such that

$$M > \|c_0\|_{C^1(\bar{\Omega})}, \quad M > \|p_0\|_{C^1(\bar{\Omega})}, \quad M > \|w_0\|_{C^{1+\gamma}(\bar{\Omega})}, \quad M > \|w_0\|_{C^1(\bar{\Omega})} + 1 \quad (2.11)$$

and introduce the constants  $c_1 = c_1(M)$  from Lemma 2.3(b) (for  $T = 1$ ),  $c_2 > 0$  from [27, Thm. 6.40] such that all solutions  $v$  of (2.10) with  $\|f\|_{L^\infty(\Omega \times (0, T))} \leq c_1$ ,  $\|g\|_{L^\infty(\Omega \times (0, T))} \leq \alpha c_1^2 + \mu + \mu e^{\alpha c_1} M$  and  $\|w_0\|_{L^\infty(\Omega)} \leq M$  satisfy  $\|v\|_{L^\infty(\Omega \times (0, T))} \leq c_2$ , and  $c_3 > 0$  such that by [26, Thms. 1.1 and 1.2], all solutions  $v$  of (2.10) with  $\partial_\nu w_0 = 0$  on  $\partial\Omega$ ,  $\|w_0\|_{C^{1+\gamma}(\bar{\Omega})} \leq M$ ,  $\|f\|_{L^\infty(\Omega \times (0, T))} \leq c_1$ ,  $\|g\|_{L^\infty(\Omega \times (0, T))} \leq \alpha c_1^2 + \mu e^{\alpha c_1} + \mu e^{2\alpha c_1} M$  and  $\|v\|_{L^\infty(\Omega \times (0, T))} \leq c_2$  also fulfil  $\|v\|_{C^{1+\gamma, \frac{1+\gamma}{2}}(\bar{\Omega} \times [0, T])} \leq c_3$ . Finally, we fix  $T \in (0, 1]$  such that  $c_3 \sqrt{T} \leq 1$ .

Successive applications of Lemma 2.3 [27, Thm. 6.40] and [26, Thms. 1.1 and 1.2] then ensure that

$$\|\Phi(w)\|_{C^{1+\gamma, \frac{\gamma}{2}}(\bar{\Omega} \times [0, T])} \leq c_3. \quad (2.12)$$

In particular,

$$\|\Phi(w)(\cdot, t)\|_{C^1(\bar{\Omega})} \leq \|\Phi(w)(\cdot, 0)\|_{C^1(\bar{\Omega})} + \|\Phi(w)(\cdot, t) - \Phi(w)(\cdot, 0)\|_{C^1(\bar{\Omega})} \leq \|w_0\|_{C^1(\bar{\Omega})} + c_3 t^{\frac{1}{2}} \leq M$$

for every  $t \in [0, T]$ . As  $v$  is moreover nonnegative by the maximum principle,  $\Phi$  maps  $S_{M,T}$  to itself and, according to (2.12) has a compact image. Schauder's fixed point theorem provides a fixed point  $w$  of  $\Phi$ , that is, (together with  $c$  and  $p$ ) a solution of (2.4) in  $\Omega \times [0, T]$ .

As  $w \in C^{1+\gamma, 0}(\bar{\Omega} \times [0, T_{\max}))$ ,  $p, c, \nabla c$  belong to  $C^{\gamma, \frac{\gamma}{2}}(\bar{\Omega} \times [0, T_{\max}))$  by Lemma 2.3(c) with  $\gamma_2 = 0$ . Since moreover  $w \in C^{\gamma, \frac{\gamma}{2}}(\bar{\Omega} \times [0, T_{\max}))$ , also  $f$  and  $g$  in (2.10) belong to this space. Then [24, Thm. IV.5.3] (if

combined with the uniqueness statement of [24, Thm. III.5.1] and  $C^{2+\gamma}$  regularity of  $w_0$ ) makes  $w$  a classical solution with  $w_t, D_x^2 w \in C^{\gamma, \frac{\gamma}{2}}(\bar{\Omega} \times [0, T_{\max}))$ . An invocation of Lemma 2.3(c) yields  $D_x^2 c, D_x^2 p, \nabla p \in C^{\gamma, 1+\frac{\gamma}{2}}(\bar{\Omega} \times [0, T_{\max}))$ . In order to prove  $\nabla w_t \in C^{\gamma, \frac{\gamma}{2}}(\bar{\Omega} \times (0, T_{\max}))$ , we fix  $t_0 \in (0, T_{\max})$  and a cutoff function  $\zeta \in C^\infty(\mathbb{R})$  with  $\zeta = 0$  on  $(-\infty, \frac{t_0}{2}]$  and  $\zeta = 1$  on  $[t_0, \infty)$ . As  $\zeta w$  then fulfills

$$\begin{cases} (\zeta w)_t = \frac{1}{\alpha} \Delta(\zeta w) + \nabla c \cdot \nabla(\zeta w) + \alpha p c(\zeta w) + \mu(\zeta w) - \mu e^{\alpha c} w(\zeta w) + \zeta' w & \text{in } \Omega \times (0, T_{\max}), \\ \partial_\nu(\zeta w) = 0 & \text{on } \partial\Omega \times (0, T_{\max}), \\ (\zeta w)(\cdot, 0) = 0 & \text{in } \Omega, \end{cases}$$

an application of [24, Thm. IV.5.3] to  $\zeta w$  ensures that  $\nabla(\zeta w)_t \in C^{\gamma, \frac{\gamma}{2}}(\bar{\Omega} \times [0, T_{\max}))$ , which entails  $\nabla w_t \in C^{\gamma, \frac{\gamma}{2}}(\bar{\Omega} \times [t_0, T_{\max}))$ . Moreover, nonnegativity of  $w$  as well as of  $c$  and  $p$  follows from the inclusion  $w \in S_{M,T}$  and the comparison principle for ordinary differential equations, respectively.

Since  $M$  and  $T$  only depend on the quantities in (2.11) and as solutions are unique by Lemma 2.4; the above reasoning can therefore be applied to extend the solution until some maximal existence time characterized by

$$T_{\max} = \infty \quad \text{or} \quad \limsup_{t \nearrow T_{\max}} \left( \|c(\cdot, t)\|_{C^1(\bar{\Omega})} + \|p(\cdot, t)\|_{C^1(\bar{\Omega})} + \|w(\cdot, t)\|_{C^{1+\gamma}(\bar{\Omega})} \right) = \infty. \quad (2.13)$$

According to Lemma 2.3(b), boundedness of the norm of  $w$  in this expression already implies that of the norms of  $c$  and  $p$ , therefore (2.13) can be reduced to (2.9).

Finally, uniqueness of this solution has already been asserted in Lemma 2.4.  $\square$

By fixing initial data as in (2.8), we henceforth implicitly also fix the unique classical solution  $(w, c, p)$  of (2.4) given by Lemma 2.5 and denote its maximal existence time by  $T_{\max}$ .

### 2.3. $L^\infty$ bounds

The results in the previous subsection show that Theorem 1.1 follows once we have shown that for the solutions constructed in Lemma 2.5 the second alternative in (2.9) cannot hold. That is, we need to derive sufficiently strong a priori estimates. In this subsection, we begin with bounds in  $L^\infty$ .

For the second solution component, such a bound directly follows from the comparison principle.

**Lemma 2.6.** *Suppose (2.7) and that  $(w_0, c_0, p_0)$  satisfies (2.8). Then the solution  $(w, c, p)$  constructed in Lemma 2.5 fulfills*

$$\|c(\cdot, t)\|_{L^\infty(\Omega)} \leq \|c_0\|_{L^\infty(\Omega)} \quad \text{for all } t \in (0, T_{\max}).$$

*Proof.* The function  $\bar{c} := \|c_0\|_{L^\infty(\Omega)}$  is a supersolution of the second equation in (2.4) and  $c \geq 0$  by Lemma 2.5.  $\square$

As a preparation for obtaining  $L^\infty$  estimates also for the other two solution components, we note that the time regularization in the third equation in (2.4) implies that we can bound  $p$  by a quantity including an *arbitrarily small* contribution of the  $L^\infty$  norm of  $w$  – at least if we are willing to shrink the time interval on which this estimates holds accordingly.

**Lemma 2.7.** *Suppose (2.7) and let  $M > 0$ . Then there exists  $T^* > 0$  such that for all  $(w_0, c_0, p_0)$  satisfying (2.8) and*

$$\|c_0\|_{L^\infty(\Omega)} \leq M,$$

*the solution  $(w, c, p)$  of (2.4) fulfills*

$$\|p\|_{L^\infty(\Omega \times (0, T))} \leq \|p_0\|_{L^\infty(\Omega)} + \min\left\{\frac{\mu}{2M\alpha}, 1\right\} \|w\|_{L^\infty(\Omega \times (0, T))} \quad \text{for all } T \in (0, T^*] \cap (0, T_{\max}).$$

*Proof.* We choose  $T^* > 0$  so small that  $Me^{\alpha M}(1 - e^{-T^*}) \leq \min\{\frac{\mu}{2M\alpha}, 1\}$  and fix  $T \in (0, T^*) \cap (0, T_{\max})$ . By the variation-of-constants formula and Lemma 2.6, we have

$$\begin{aligned} \|p(\cdot, t)\|_{L^\infty(\Omega)} &\leq e^{-t}\|p_0\|_{L^\infty(\Omega)} + \int_0^t e^{-(t-s)}\|we^{\alpha c}\|_{L^\infty(\Omega)}(\cdot, s) \, ds \\ &\leq \|p_0\|_{L^\infty(\Omega)} + \|c_0\|_{L^\infty(\Omega)}e^{\alpha\|c_0\|_{L^\infty(\Omega)}}\|w\|_{L^\infty(\Omega \times (0, T))} \int_0^{T^*} e^{-s} \, ds \\ &\leq \|p_0\|_{L^\infty(\Omega)} + Me^{\alpha M}(1 - e^{-T^*})\|w\|_{L^\infty(\Omega \times (0, T))} \quad \text{for all } t \in (0, T), \end{aligned}$$

which implies the statement due to the definition of  $T^*$ . □

We now turn our attention to  $L^\infty$  estimates of  $w$ . The fact that the transformed quantity  $w$  fulfills an equation whose first- and second-order terms reduce to  $e^{-\alpha c}\nabla \cdot (\frac{1}{\alpha}e^{\alpha c}\nabla w)$  (cf. (2.3)), an expression without any explicit  $\nabla c$ , opens the door for certain testing procedures. In related works, these have been used to first derive boundedness in  $L^p$  and then, after an iteration argument, also in  $L^\infty$  (see for instance [42, Prop. 5.1], [40, Lemma 3.5] and [39, Lemma 3.10]). However, here we are able to employ a slightly faster method: another advantage of the transformation  $w = e^{-\alpha c}u$  is that sufficiently large constant functions are supersolutions of the equation for  $w$  in (2.4), at least as long both  $c$  and  $p$  are bounded. Therefore, the previous two lemmata allow us to prove boundedness for  $w$  on small timescales.

We also emphasize that the following proof crucially relies on the presence of a logistic source in the first equation, that is, on positivity of  $\mu$ . (The same would be true for testing procedures similar to those performed in the works referenced above.) In fact, this is essentially the only place where we directly make use of the assumption  $\mu > 0$ .

**Lemma 2.8.** *Suppose (2.7) and let  $M > 0$ . Let  $T^* > 0$  be as given by Lemma 2.7. Then there is  $K > 0$  with the following property: for all  $L > 0$  and all  $(w_0, c_0, p_0)$  satisfying (2.8) and*

$$\|w_0\|_{L^\infty(\Omega)} \leq L, \quad \|c_0\|_{L^\infty(\Omega)} \leq M \quad \text{as well as} \quad \|p_0\|_{L^\infty(\Omega)} \leq L,$$

*the solution  $(w, c, p)$  of (2.4) fulfills*

$$\|w(\cdot, t)\|_{L^\infty(\Omega)} + \|p(\cdot, t)\|_{L^\infty(\Omega)} \leq K(L + 1) \quad \text{for all } t \in (0, T^*) \cap (0, T_{\max}).$$

*Proof.* We fix  $T \in (0, T^*) \cap (0, T_{\max})$ . By Lemmas 2.6 and 2.7, we may estimate

$$\begin{aligned} w_t - \frac{1}{\alpha}\Delta w - \nabla c \cdot \nabla w &= w(\alpha pc + \mu - \mu e^{\alpha c}w) \leq w(M\alpha\|p\|_{L^\infty(\Omega \times (0, T))} + \mu - \mu w) \\ &\leq w\left(LM\alpha + \frac{\mu}{2}\|w\|_{L^\infty(\Omega \times (0, T))} + \mu - \mu w\right) \end{aligned}$$

in  $\Omega \times (0, T)$ . Therefore, the comparison principle, applied with the constant supersolution

$$\bar{w} := \max\left\{L, \frac{LM\alpha}{\mu} + 1 + \frac{1}{2}\|w\|_{L^\infty(\Omega \times (0, T))}\right\},$$

asserts

$$\|w\|_{L^\infty(\Omega \times (0, T))} \leq \bar{w} \leq \max\{L, LM\alpha\mu^{-1} + 1\} + \frac{1}{2}\|w\|_{L^\infty(\Omega \times (0, T))}$$

and thus

$$\|w\|_{L^\infty(\Omega \times (0, T))} \leq 2 \max\{L, LM\alpha\mu^{-1} + 1\}.$$

Since  $\|p\|_{L^\infty(\Omega \times (0, T))} \leq L + \|w\|_{L^\infty(\Omega \times (0, T))}$  by Lemma 2.7, this implies the statement for  $K := 5 \max\{1, M\alpha\mu^{-1}\}$ . □

Next, iteratively applying Lemma 2.8 allows us to derive boundedness for all solution components on all bounded time intervals. We note that a prerequisite for such an iteration procedure is that the time  $T^*$  given by Lemma 2.7 (and to a lesser extent also the constant  $K$  given by Lemma 2.8) only depends on the data in a manageable way. This justifies why we have kept track of the dependencies of the constants in the previous lemmata.

**Lemma 2.9.** *Suppose (2.7) and let  $M > 0$ . Then there are  $C_1, C_2 > 0$  with the following property: for all  $L > 0$  and all  $(w_0, c_0, p_0)$  satisfying (2.8) and*

$$\|w_0\|_{L^\infty(\Omega)} \leq L, \quad \|c_0\|_{L^\infty(\Omega)} \leq M \quad \text{as well as} \quad \|p_0\|_{L^\infty(\Omega)} \leq L,$$

the solution  $(w, c, p)$  of (2.4) fulfills

$$\|p(\cdot, t)\|_{L^\infty(\Omega)} + \|w(\cdot, t)\|_{L^\infty(\Omega)} \leq e^{C_1 t} C_2 (L + 1) \quad \text{for all } t \in (0, T_{\max}). \quad (2.14)$$

*Proof.* We set  $w(\cdot, t) = w_0$  and  $p(\cdot, t) = p_0$  for  $t < 0$ , and let  $T^* > 0$  and  $K > 1$  be as given by Lemmas 2.7 and 2.8, respectively. Moreover, setting

$$I_j := ((j - 1)T^*, jT^*] \cap (-\infty, T_{\max}) \quad \text{for } j \in \mathbb{N}_0$$

and applying Lemma 2.7 (which is applicable for the same  $M$  by Lemma 2.6) and Lemma 2.8 to initial data  $(w, c, p)(\cdot, (j - 1)T^*)$ , we can estimate

$$\|p\|_{L^\infty(\Omega \times I_j)} \leq \|p\|_{L^\infty(\Omega \times I_{j-1})} + \|w\|_{L^\infty(\Omega \times I_j)}$$

and

$$\|w\|_{L^\infty(\Omega \times I_j)} \leq K(\|w\|_{L^\infty(\Omega \times I_{j-1})} + \|p\|_{L^\infty(\Omega \times I_{j-1})} + 1)$$

for all  $j \in \mathbb{N}$  with  $(j - 1)T^* < T_{\max}$ . However, the same estimates hold trivially also in the case of  $(j - 1)T^* \geq T_{\max}$ , as then  $I_j = \emptyset$ . Thus,

$$A_j := \|p\|_{L^\infty(\Omega \times I_j)} + \|w\|_{L^\infty(\Omega \times I_j)} + 1, \quad j \in \mathbb{N},$$

fulfills

$$A_j \leq \|p\|_{L^\infty(\Omega \times I_{j-1})} + 2K(\|w\|_{L^\infty(\Omega \times I_{j-1})} + \|p\|_{L^\infty(\Omega \times I_{j-1})} + 1) + 1 \leq (2K + 1)A_{j-1} \quad \text{for all } j \in \mathbb{N}.$$

A straightforward induction then yields  $A_j \leq (2K + 1)^j A_0 \leq e^{j \ln(2K+1)}(2L + 1)$  for all  $j \in \mathbb{N}$ . If  $t \in I_j$  for some  $j \in \mathbb{N}_0$  and thus  $(j - 1)T^* < t$ , that is  $j < \frac{t}{T^*} + 1$ , then

$$\|w(\cdot, t)\|_{L^\infty(\Omega)} + \|p(\cdot, t)\|_{L^\infty(\Omega)} \leq A_j \leq e^{j \ln(2K+1)}(2L + 1) \leq e^{t(T^*)^{-1} \ln(2K+1)} 2(2K + 1)(L + 1).$$

Since  $(0, T_{\max}) \subset \bigcup_{j \in \mathbb{N}_0} I_j$ , this implies (2.14) for  $C_1 = \frac{\ln(2K+1)}{T^*}$  and  $C_2 = 2(2K + 1)$ . □

### 2.4. Gradient bounds in $L^4$

While the  $L^\infty$  estimates proven in the previous subsection surely form an important step towards proving global existence, the extensibility criterion (2.9) also requires boundedness of the gradients – which will be the topic of the present and the following subsection.

As a first step, we again make use of the time regularization in the second and third equation in (2.4) to obtain

**Lemma 2.10.** *Suppose (2.7) and let  $T_0 \in (0, \infty)$  as well as  $q \in (1, \infty)$ . For all  $M > 0$ , there exists  $C > 0$  such that if  $(w_0, c_0, p_0)$  satisfying (2.8) are such that the corresponding solution  $(w, c, p)$  of (2.4) fulfills*

$$w \leq M, \quad c \leq M \quad \text{and} \quad p \leq M \quad \text{in } \bar{\Omega} \times [0, T], \tag{2.15}$$

where  $T := \min\{T_0, T_{\max}\}$ , then

$$\left( \int_{\Omega} |\nabla c(\cdot, t)|^q + \int_{\Omega} |\nabla p(\cdot, t)|^q \right) \leq C \left( \int_{\Omega} |\nabla c_0|^q + \int_{\Omega} |\nabla p_0|^q + \int_0^t \int_{\Omega} |\nabla w|^q \right) \quad \text{for all } t \in [0, T]. \tag{2.16}$$

*Proof.* According to (2.4),

$$(\nabla c)_t = -p \nabla c - c \nabla p \quad \text{and} \quad (\nabla p)_t = -\nabla p + w(\alpha c + 1) e^{\alpha c} \nabla c + e^{\alpha c} c \nabla w$$

hold in  $\Omega \times (0, T)$ . By testing these equations with  $q|\nabla c|^{q-2} \nabla c$  and  $q|\nabla p|^{q-2} \nabla p$ , respectively, and applying Young's inequality, we obtain

$$\frac{d}{dt} \int_{\Omega} |\nabla c|^q = -q \int_{\Omega} p |\nabla c|^q - q \int_{\Omega} c |\nabla c|^{q-2} \nabla c \cdot \nabla p \leq Mq \left( \int_{\Omega} |\nabla c|^q + \int_{\Omega} |\nabla p|^q \right)$$

and

$$\begin{aligned} \frac{d}{dt} \int_{\Omega} |\nabla p|^q &= -q \int_{\Omega} |\nabla p|^q + q \int_{\Omega} w(\alpha c + 1) e^{\alpha c} |\nabla p|^{q-2} \nabla p \cdot \nabla c + q \int_{\Omega} e^{\alpha c} c |\nabla p|^{q-2} \nabla p \cdot \nabla w \\ &\leq M(\alpha M + 1) e^{\alpha M} q \left( \int_{\Omega} |\nabla c|^q + \int_{\Omega} |\nabla p|^q \right) + M e^{\alpha M} q \left( \int_{\Omega} |\nabla p|^q + \int_{\Omega} |\nabla w|^q \right) \end{aligned}$$

in  $(0, T)$ . Thus, setting  $c_1 := Mq + M(\alpha M + 1) e^{\alpha M} q + M e^{\alpha M} q$ , we can conclude

$$\frac{d}{dt} \left( \int_{\Omega} |\nabla c|^q + \int_{\Omega} |\nabla p|^q \right) \leq c_1 \left( \int_{\Omega} |\nabla c|^q + \int_{\Omega} |\nabla p|^q \right) + c_1 \int_{\Omega} |\nabla w|^q \quad \text{in } (0, T),$$

which after an application of Grönwall's inequality turns into

$$\begin{aligned} \left( \int_{\Omega} |\nabla c(\cdot, t)|^q + \int_{\Omega} |\nabla p(\cdot, t)|^q \right) &\leq e^{c_1 t} \left( \int_{\Omega} |\nabla c_0|^q + \int_{\Omega} |\nabla p_0|^q \right) + c_1 \int_0^t e^{c_1(t-s)} \int_{\Omega} |\nabla w(\cdot, s)|^q ds \\ &\leq e^{c_1 T} \left( \int_{\Omega} |\nabla c_0|^q + \int_{\Omega} |\nabla p_0|^q \right) + c_1 e^{c_1 T} \int_0^t \int_{\Omega} |\nabla w|^q, \quad t \in (0, T), \end{aligned}$$

and thus asserts (2.16) for  $C := \max\{c_1, 1\} e^{c_1 T}$ . □

Next, we follow [39, Lemma 3.14] and test the first equation in (2.4) with  $-\Delta w$ , which when combined with Lemma 2.10 allows us to obtain certain gradient bounds first on small time scales and then by means of an iteration argument also on each finite time interval.

**Lemma 2.11.** *Suppose (2.7) and let  $M > 0$ . There exists  $T_1 \in (0, \infty)$  such that if  $(w_0, c_0, p_0)$  satisfying (2.8) are such that the corresponding solution  $(w, c, p)$  of (2.4) fulfills (2.15), we can find  $C > 0$  such that*

$$\int_{\Omega} |\nabla c(\cdot, t)|^4 \leq C \quad \text{for all } t \in [0, T_1] \cap [0, T_{\max}]. \tag{2.17}$$

*Proof.* By  $c_1$ , we denote the constant appearing in (2.16) given by Lemma 2.10 applied to  $T_0 = 1$  and  $q = 4$ . Moreover, the Gagliardo–Nirenberg inequality (cf. [32] or [18, (A.2)] for this form) asserts that there is  $c_2 > 0$  such that

$$\int_{\Omega} |\nabla\varphi|^4 \leq c_2 \int_{\Omega} |\Delta\varphi|^2 + c_2 \quad \text{for all } \varphi \in C^2(\bar{\Omega}) \quad \text{with } \partial_\nu\varphi = 0 \quad \text{in } \partial\Omega \quad \text{and } \|\varphi\|_{L^\infty(\Omega)} \leq M. \quad (2.18)$$

We then choose  $T_1 \in (0, 1)$  so small that

$$T_1 c_1 c_2 \alpha^3 \leq \frac{1}{8c_2\alpha} \quad (2.19)$$

holds and fix a solution  $(w, c, p)$  of (2.4) with maximal existence time  $T_{\max}$  fulfilling (2.15).

These choices now allow us to infer from Lemma 2.10 that

$$\begin{aligned} \int_0^T \int_{\Omega} |\nabla c|^4 &\leq c_1 \left( \int_0^T \int_{\Omega} |\nabla c_0|^4 + \int_0^T \int_{\Omega} |\nabla p_0|^4 + \int_0^T \int_0^t \int_{\Omega} |\nabla w(\cdot, \tau)|^4 \, d\tau \, dt \right) \\ &\leq c_1 \left( \int_0^T \int_{\Omega} |\nabla c_0|^4 + \int_0^T \int_{\Omega} |\nabla p_0|^4 + \int_0^T \int_0^T \int_{\Omega} |\nabla w(\cdot, \tau)|^4 \, d\tau \, dt \right) \\ &\leq T_1 c_1 \left( \int_{\Omega} |\nabla c_0|^4 + \int_{\Omega} |\nabla p_0|^4 + \int_0^T \int_{\Omega} |\nabla w|^4 \right) \end{aligned}$$

holds for all  $T \in (0, T_1] \cap (0, T_{\max})$ . Next, we test the first equation in (2.4) with  $-\Delta w$  to obtain

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \int_{\Omega} |\nabla w|^2 &= -\frac{1}{\alpha} \int_{\Omega} |\Delta w|^2 - \int_{\Omega} (\nabla c \cdot \nabla w) \Delta w - \int_{\Omega} (\alpha p c w + \mu w - \mu e^{\alpha c} w^2) \Delta w \\ &\leq -\frac{1}{2\alpha} \int_{\Omega} |\Delta w|^2 + \alpha \int_{\Omega} |\nabla c \cdot \nabla w|^2 + \underbrace{\alpha |\Omega| (\alpha M^3 + \mu M + \mu e^{\alpha M} M^2)}_{=: c_3} \end{aligned}$$

and thus

$$\begin{aligned} \frac{1}{2} \int_{\Omega} |\nabla w(\cdot, T)|^2 - \frac{1}{2} \int_{\Omega} |\nabla w_0|^2 + \frac{1}{2\alpha} \int_0^T \int_{\Omega} |\Delta w|^2 \\ \leq \alpha \int_0^T \int_{\Omega} |\nabla c \cdot \nabla w|^2 + T c_3 \\ \leq \frac{1}{4c_2\alpha} \int_0^T \int_{\Omega} |\nabla w|^4 + c_2 \alpha^3 \int_0^T \int_{\Omega} |\nabla c|^4 + T_1 c_3 \\ \leq \left( \frac{1}{4c_2\alpha} + T_1 c_1 c_2 \alpha^3 \right) \int_0^T \int_{\Omega} |\nabla w|^4 + T_1 c_1 c_2 \alpha^3 \left( \int_{\Omega} |\nabla c_0|^4 + \int_{\Omega} |\nabla p_0|^4 \right) + T_1 c_3 \quad (2.20) \end{aligned}$$

for all  $T \in (0, T_1] \cap (0, T_{\max})$ . Since (2.19) and (2.18) imply

$$\begin{aligned} \left( \frac{1}{4c_2\alpha} + T_1 c_1 c_2 \alpha^3 \right) \int_0^T \int_{\Omega} |\nabla w|^4 &\leq \left( \frac{1}{2c_2\alpha} - \frac{1}{8c_2\alpha} \right) \int_0^T \int_{\Omega} |\nabla w|^4 \\ &\leq \frac{1}{2\alpha} \int_0^T \int_{\Omega} |\Delta w|^2 + \frac{T}{2\alpha} - \frac{1}{8c_2\alpha} \int_0^T \int_{\Omega} |\nabla w|^4 \quad (2.21) \end{aligned}$$

for all  $T \in (0, T_1] \cap (0, T_{\max})$ , we conclude from (2.20) and (2.21) that

$$\frac{1}{8c_2\alpha} \int_0^T \int_{\Omega} |\nabla w|^4 \leq \frac{1}{2} \int_{\Omega} |\nabla w_0|^2 + T_1 c_1 c_2 \alpha^3 \left( \int_{\Omega} |\nabla c_0|^4 + \int_{\Omega} |\nabla p_0|^4 \right) + T_1 c_3 + \frac{T}{2\alpha}$$

for all  $T \in (0, T_1] \cap (0, T_{\max})$ . Again applying Lemma 2.10, we finally see that (2.17) holds for some  $C > 0$  (depending on  $w_0$ ,  $c_0$  and  $p_0$ ).  $\square$

**Lemma 2.12.** *Suppose (2.7) and that  $(w_0, c_0, p_0)$  satisfies (2.8). For all  $T \in (0, T_{\max}] \cap (0, \infty)$ , there exists  $C > 0$  such that the solution of (2.4) fulfills*

$$\int_{\Omega} |\nabla c(\cdot, t)|^4 \leq C \quad \text{for all } t \in [0, T]. \quad (2.22)$$

*Proof.* Lemma 2.6 and Lemma 2.9 assert that (2.15) holds for some  $M > 0$ . We fix  $T_1 \in (0, \infty)$  be as given by Lemma 2.11 for this  $M$ . If  $j \in \mathbb{N}_0$  is such that  $T_1 j < T$ , an application of Lemma 2.11 to the solution with initial data  $(w, c, p)(\cdot, T_1 j)$  shows that there is  $c_j > 0$  such that (2.22) holds with  $C$  replaced by  $c_j$  for all  $t \in [T_1 j, T_1(j+1)]$ . Thus, the statement follows for  $C := \max\{c_j : j \in \mathbb{N}_0, j < \frac{T}{T_1}\}$ .  $\square$

## 2.5. Hölder estimates for the gradients

Lemmas 2.6, 2.9 and 2.12 provide several bounds for the right-hand side of the first equation in (2.4), which allow us to make use of parabolic regularity theory to iteratively improve our bounds. In particular, we adapt the techniques developed in [39, pp. 791–792], where only planar domains are considered, to the three-dimensional setting.

As it is used multiple times in the proof of Lemma 2.14 below, we first state the following consequence of maximal Sobolev regularity results.

**Lemma 2.13.** *Suppose that  $\Omega \subset \mathbb{R}^n$ ,  $n \in \mathbb{N}$ , is a smooth, bounded domain. Let  $T > 0$ ,  $\alpha > 0$ ,  $s \in (0, \infty)$  and  $q \in (n, \infty)$ . For any  $M > 0$ , there is  $C > 0$  such that if  $w_0 \in C^2(\bar{\Omega})$  with  $\partial_\nu w_0 = 0$  on  $\partial\Omega$ ,  $f \in L^\infty((0, T); L^q(\Omega))$  and  $g \in L^s((0, T); L^q(\Omega))$  are such that*

$$\|w_0\|_{C^2(\bar{\Omega})} \leq M, \quad \|f\|_{L^\infty((0, T); L^q(\Omega))} \leq M \quad \text{and} \quad \|g\|_{L^s((0, T); L^q(\Omega))} \leq M, \quad (2.23)$$

then every solution  $w \in C^{2,1}(\bar{\Omega} \times (0, T)) \cap C^1(\bar{\Omega} \times [0, T])$  of

$$\begin{cases} w_t = \frac{1}{\alpha} \Delta w + f \cdot \nabla w + g & \text{in } \Omega \times (0, T), \\ \partial_\nu w = 0 & \text{on } \partial\Omega \times (0, T), \\ w(\cdot, 0) = w_0 & \text{in } \Omega \end{cases}$$

with  $|w| \leq M$  in  $\Omega \times (0, T)$  fulfills

$$\|w_t\|_{L^s((0, T); L^q(\Omega))} + \|\Delta w\|_{L^s((0, T); L^q(\Omega))} + \|\nabla w\|_{L^s((0, T); L^\infty(\Omega))} \leq C. \quad (2.24)$$

*Proof.* We fix the data  $w_0$ ,  $f$  and  $g$  and a solution  $w$  but emphasize that the constants  $c_1$  and  $c_2$  below only depend on  $M$  (and  $\Omega$ ,  $T$ ,  $\alpha$ ,  $s$  and  $q$ ). Since  $f \cdot \nabla w + g \in L_{\text{loc}}^s([0, T]; L^q(\Omega))$  by assumption, [20, Thm. 2.3] asserts that  $w$  is also the unique solution of [20, (2.6)] and thus that the estimate [20, (2.7)] holds. From [20, (2.7)] in conjunction with (2.23), we hence obtain  $c_1 > 0$  such that

$$\|w_t\|_{L^s((0, T); L^q(\Omega))} + \frac{1}{\alpha} \|\Delta w\|_{L^s((0, T); L^q(\Omega))} \leq c_1 \|f \cdot \nabla w\|_{L^s((0, T); L^q(\Omega))} + c_1. \quad (2.25)$$

Since  $q > n$ , the embedding  $W^{2,q}(\Omega) \hookrightarrow W^{1,\infty}(\Omega)$  is compact, so that an application of Ehrling's lemma combined with elliptic regularity (cf. [17, Thm. 19.1]) shows that there is  $c_2 > 0$  such that

$$\|\nabla \varphi\|_{L^\infty(\Omega)} \leq \frac{1}{2Mc_1\alpha} \|\Delta \varphi\|_{L^q(\Omega)} + c_2 \|\varphi\|_{L^\infty(\Omega)} \quad \text{for all } \varphi \in C^2(\bar{\Omega}) \text{ with } \partial_\nu \varphi = 0 \text{ on } \partial\Omega. \quad (2.26)$$

Additionally relying on Minkowski's inequality, we thus obtain

$$\begin{aligned} \|f \cdot \nabla w\|_{L^s((0,T);L^q(\Omega))} &\leq \|f\|_{L^\infty((0,T);L^q(\Omega;\mathbb{R}^n))} \|\nabla w\|_{L^s((0,T);L^\infty(\Omega))} \\ &\leq M \left( \frac{1}{2Mc_1\alpha} \|\Delta w\|_{L^s((0,T);L^q(\Omega))} + c_2 \|w\|_{L^s((0,T);L^\infty(\Omega))} \right) \\ &\leq \frac{1}{2c_1\alpha} \|\Delta w\|_{L^s((0,T);L^q(\Omega))} + M^2 T^{\frac{1}{s}} c_2. \end{aligned}$$

In combination with (2.25), this yields

$$\|w_t\|_{L^s((0,T);L^q(\Omega))} + \frac{1}{2\alpha} \|\Delta w\|_{L^s((0,T);L^q(\Omega))} \leq c_1 + M^2 T^{\frac{1}{s}} c_1 c_2,$$

upon which another application of (2.26) implies (2.24) for some  $C > 0$ . □

**Lemma 2.14.** *Suppose (2.7) and that  $(w_0, c_0, p_0)$  satisfies (2.8). Then the maximal classical solution  $(w, c, p)$  of (2.4) given by Lemma 2.5 is global in time.*

*Proof.* We may without loss of generality assume that  $\gamma$  in (2.8) satisfies  $\gamma < \frac{7}{12}$ , let  $(w, c, p)$  the solution of (2.4) provided by Lemma 2.5 and suppose that on the contrary the maximal existence time  $T_{\max}$  is finite. By Lemmas 2.6 and 2.9,

$$w_t = \frac{1}{\alpha} \Delta w + \nabla c \cdot \nabla w + g \quad \text{in } \bar{\Omega} \times [0, T_{\max})$$

holds for  $g = \alpha p c w + \mu w - \mu e^{\alpha c} w^2 \in L^\infty(\Omega \times (0, T_{\max}))$ . Moreover, the initial data fulfill (2.8),  $\nabla c$  belongs to  $L^\infty((0, T_{\max}); L^4(\Omega))$  by Lemma 2.12 (applied to  $T = T_{\max} < \infty$ ) and  $w$  is bounded by Lemma 2.9, hence an application of Lemma 2.13 with  $s = 12$  and  $q = 4$  shows that (2.24) holds, which entails that

$$\|\nabla w\|_{L^{12}(\Omega \times (0, T_{\max}))} \leq c_1$$

for some  $c_1 > 0$ . Therefore, Lemma 2.10 (applied to  $T_0 = T_{\max}$ ) asserts that  $\nabla c \in L^\infty((0, T_{\max}); L^{12}(\Omega))$ , so that we may again apply Lemma 2.13, this time with  $s = 12$  and  $q = 12$ , to obtain  $c_2 > 0$  such that

$$\|w_t\|_{L^{12}(\Omega \times (0, T_{\max}))} + \|\Delta w\|_{L^{12}(\Omega \times (0, T_{\max}))} \leq c_2.$$

This in turn renders [24, Lemma II.3.3] applicable, which asserts finiteness of  $\|w\|_{C^{\frac{19}{12}, \frac{19}{24}}(\bar{\Omega} \times [0, T_{\max}])}$ , contradicting the extensibility criterion in Lemma 2.5. Thus our assumption that  $T_{\max}$  is finite must be false. □

### 2.6. Proof of Theorem 1.1

The proof of Theorem 1.1 has now been reduced to referencing some of the lemmata above.

*Proof of Theorem 1.1.* Lemma 2.5 asserts the local existence of a unique maximal classical solution of (2.4), which is global in time by Lemma 2.14. Therefore, the statement follows by transforming back to the original variables; that is, first setting  $u(x, t) := w(x, t)e^{\alpha c(x, t)}$  for  $x \in \bar{\Omega}$  and  $t \in [0, \infty)$  and then applying the transformation in (2.1). □

## 3. WEAK FORMULATION, DISCRETIZATION AND NUMERICAL SOLUTION

In this section, we address the numerical realization of (1.1). To this end, we first derive a weak formulation and then apply the Rothe method, namely, first temporal discretization using finite differences, and afterward spatial discretization based on a Galerkin finite element scheme. Due to the highly nonlinear behavior, we then propose and implement a fixed-point algorithm to solve all three equations sequentially. Similar algorithms and implementations are available in the deal.II library [5, 6], and we have former experiences in solving highly nonlinear coupled PDE systems, e.g., [43], but the algorithmic design, implementation and code verification of (1.1) in deal.II is novel to the best of our knowledge.



### 3.1. Weak formulation

Using integration by parts and the homogeneous boundary conditions, the variational formulation for the system (1.1) reads: find  $u, c, p \in L^2(0, T, H^1(\Omega))$  with  $u_t, c_t, p_t \in L^2(0, T, (H^1(\Omega))^*)$  and the initial conditions  $u^0 = u(0) \in H^1(\Omega), c^0 = c(0) \in L^2(\Omega), p^0 = p(0) \in L^2(\Omega)$  such that for almost all times  $t \in (0, T)$ , we have

$$\begin{aligned} \langle u_t, \phi^u \rangle + \frac{1}{\alpha} \int_{\Omega} \nabla u \cdot \nabla \phi^u \, dx - \chi \int_{\Omega} u \nabla c \cdot \nabla \phi^u \, dx - \mu \int_{\Omega} u(1-u) \phi^u \, dx &= 0 \quad \forall \phi^u \in C^\infty(\bar{\Omega}), \\ \langle c_t, \phi^c \rangle + \int_{\Omega} p c \phi^c \, dx &= 0 \quad \forall \phi^c \in C^\infty(\bar{\Omega}), \\ \langle p_t, \phi^p \rangle - \varepsilon^{-1} \int_{\Omega} (u c - p) \phi^p \, dx &= 0 \quad \forall \phi^p \in C^\infty(\bar{\Omega}). \end{aligned} \quad (3.1)$$

### 3.2. Temporal discretization and fixed point scheme

Let us now proceed and subdivide the time interval  $[0, T]$  into  $N$  subintervals  $[0, T] = \cup_{n=0}^{N-1} [t^n, t^{n+1}]$  with the uniform time steps  $\Delta t = t^{n+1} - t^n, n = 0, 1, 2, \dots, N-1$ . We use  $c^{n+1}(x) := c(x, t^{n+1}), p^{n+1}(x) := p(x, t^{n+1})$  and  $u^{n+1}(x) := u(x, t^{n+1})$  to denote the approximation of the solutions at time  $t^{n+1}$ . Specifically, for time discretization we employ the well-known  $\theta$  method allowing us to work with implicit  $A$ -stable time-stepping schemes for choice  $\theta \in [0.5, 1]$  in each equation. Further, a fixed-point scheme is used to decouple the previous system and to treat the nonlinear and coupled terms.

Then, a semi-discrete and linearized form of the system (3.1) in the interval  $[t^n, t^{n+1}]$  reads: for given  $u_0^{n+1} = u^n, c_0^{n+1} = c^n$  and  $p_0^{n+1} = p^n$  find  $u_k^{n+1} \in H^1(\Omega), c_k^{n+1} \in H^1(\Omega)$  and  $p_k^{n+1} \in H^1(\Omega)$  such that

$$\begin{aligned} \int_{\Omega} u_k^{n+1} \phi^u \, dx + \theta \Delta t \left( \frac{1}{\alpha} \int_{\Omega} \nabla u_k^{n+1} \cdot \nabla \phi^u \, dx - \chi \int_{\Omega} u_k^{n+1} \nabla c_{k-1}^{n+1} \cdot \nabla \phi^u \, dx - \mu \int_{\Omega} u_k^{n+1} (1 - u_{k-1}^{n+1}) \phi^u \, dx \right) \\ = \int_{\Omega} u^n \phi^u \, dx - (1 - \theta) \Delta t \left( \frac{1}{\alpha} \int_{\Omega} \nabla u^n \cdot \nabla \phi^u \, dx \right. \\ \left. - \chi \int_{\Omega} u^n \nabla c^n \cdot \nabla \phi^u \, dx - \mu \int_{\Omega} u^n (1 - u^n) \phi^u \, dx \right) \quad \forall \phi^u \in C^\infty(\bar{\Omega}) \end{aligned}$$

and

$$\int_{\Omega} c_k^{n+1} \phi^c \, dx + \theta \Delta t \int_{\Omega} p_{k-1}^{n+1} c_k^{n+1} \phi^c \, dx = \int_{\Omega} c^n \phi^c \, dx - (1 - \theta) \Delta t \int_{\Omega} p^n c^n \phi^c \, dx \quad \forall \phi^c \in C^\infty(\bar{\Omega})$$

and

$$\begin{aligned} (1 + \varepsilon^{-1} \theta \Delta t) \int_{\Omega} p_k^{n+1} \phi^p \, dx \\ = (1 - \varepsilon^{-1} (1 - \theta) \Delta t) \int_{\Omega} p^n \phi^p \, dx + \varepsilon^{-1} \theta \Delta t \int_{\Omega} u_k^{n+1} c_k^{n+1} \phi^p \, dx + \varepsilon^{-1} (1 - \theta) \Delta t \int_{\Omega} u^n c^n \phi^p \, dx \quad \forall \phi^p \in C^\infty(\bar{\Omega}) \end{aligned}$$

for  $k = 1, 2, \dots, k^*$ , where  $k^*$  is the iteration index where some stopping criterion is met, and for  $n = 0, 1, \dots, N-1$ . For details on the specific steps and stopping tolerances we refer the reader to Section 3.4.

### 3.3. Spatial Galerkin discretization with finite elements

Our spatial discretization is based on a Galerkin finite element scheme using conforming finite elements (bilinear in two dimensions and trilinear in three dimensions). To this end,  $\Omega$  is decomposed into quadrilaterals

or hexahedra making up a mesh  $\mathcal{T}_h$ . Then, a conforming subspace  $V_h \subset H^1(\Omega)$  for approximating  $u_h^{n+1}, c_h^{n+1}$  and  $p_h^{n+1}$  is designed, which is composed of  $Q_1^c$  functions. In detail, we define

$$V_h = \{v \in C^0(\bar{\Omega}); v|_K \in Q_1(K) \text{ for } K \in \mathcal{T}_h\}.$$

Denoting by  $Q_1(\hat{K})$  the space of polynomials on the reference cell  $\hat{K}$  (square in two dimensions and cube in three dimensions) which are linear in each variable, the shape functions from  $Q_1(K)$  are obtained using  $Q_1(\hat{K})$  transformations of functions in  $Q_1(\hat{K})$  onto  $K$ , so-called isoparametric finite elements. We refer the reader to the classical textbook [12] for more details.

Moreover, we denote by  $(\cdot, \cdot)$  the scalar product in  $L^2(\Omega)$ .

The discrete solutions  $u_h^{n+1}, c_h^{n+1}$  and  $p_h^{n+1}$  are written as linear combinations of standard basis functions of  $V_h$ :

$$u_h^{n+1}(x) = \sum_{i=1}^M u_i^{n+1} \phi_i(x), \quad c_h^{n+1}(x) = \sum_{i=1}^M c_i^{n+1} \phi_i(x), \quad p_h^{n+1}(x) = \sum_{i=1}^M p_i^{n+1} \phi_i(x), \quad (3.2)$$

where  $M$  denotes the number of spatial degrees of freedom, *i.e.*,  $\dim V_h = M$ . The fully discrete system then reads as follows:

$$\begin{aligned} & \sum_{i=1}^M \left[ (\phi_i, \phi_j) + \theta \Delta t \left( \frac{1}{\alpha} (\nabla \phi_i, \nabla \phi_j) - \chi \left( \phi_i \nabla c_{h,k-1}^{n+1}, \nabla \phi_j \right) - \mu \left( \phi_i (1 - u_{h,k-1}^{n+1}), \phi_j \right) \right) \right] u_{i,k}^{n+1} \\ &= \sum_{i=1}^M \left[ (\phi_i, \phi_j) - (1 - \theta) \Delta t \left( \frac{1}{\alpha} (\nabla \phi_i, \nabla \phi_j) - \chi \left( \phi_i \nabla c_h^n, \nabla \phi_j \right) - \mu \left( \phi_i (1 - u_h^n), \phi_j \right) \right) \right] u_i^n, \end{aligned} \quad (3.3)$$

and

$$\sum_{i=1}^M \left[ (\phi_i, \phi_j) + \theta \Delta t \left( p_{h,k-1}^{n+1} \phi_i, \phi_j \right) \right] c_{i,k}^{n+1} = \sum_{i=1}^M \left[ (\phi_i, \phi_j) - (1 - \theta) \Delta t \left( p_h^n \phi_i, \phi_j \right) \right] c_i^n, \quad (3.4)$$

and

$$\begin{aligned} & \sum_{i=1}^M \left[ (1 + \varepsilon^{-1} \theta \Delta t) (\phi_i, \phi_j) \right] p_{i,k}^{n+1} \\ &= \sum_{i=1}^M \left[ (1 - \varepsilon^{-1} (1 - \theta) \Delta t) (\phi_i, \phi_j) \right] p_i^n + \varepsilon^{-1} \theta \Delta t \left( u_{h,k}^{n+1} c_{h,k}^{n+1}, \phi_j \right) + \varepsilon^{-1} (1 - \theta) \Delta t \left( u_h^n c_h^n, \phi_j \right), \end{aligned} \quad (3.5)$$

where  $j = 1, \dots, M$  and the unknown solution coefficients  $\left\{ u_{i,k}^{n+1} \right\}_{i=1}^M \in \mathbb{R}^M$ ,  $\left\{ c_{i,k}^{n+1} \right\}_{i=1}^M \in \mathbb{R}^M$  and  $\left\{ p_{i,k}^{n+1} \right\}_{i=1}^M \in \mathbb{R}^M$  at each fixed-point iteration  $k$  and each time step  $n + 1$  define the corresponding finite element functions  $u_{h,k}^{n+1} \in V_h$ ,  $c_{h,k}^{n+1} \in V_h$  and  $p_{h,k}^{n+1} \in V_h$ , respectively, analogously as in (3.2). Each linear system is solved with a sparse direct solver.

### 3.4. Algorithm

Collecting all pieces from the previous subsections, we arrive at the following final algorithm.

**Algorithm 3.1** (Fixed-point iterative scheme).

Let the fully discrete form (3.3)–(3.5) be given.

Step 1: initialize at time  $t = 0$  for  $n = 0$  with  $u_h^0 = i_h u_0$ ,  $c_h^0 = i_h c_0$  and  $p_h^0 = i_h p_0$ , where  $i_h$  is the standard Lagrange interpolation operator,

Step 2: for  $n \geq 0$  (time step number index)

set  $u_{h,0}^{n+1} = u_h^n$ ,  $c_{h,0}^{n+1} = c_h^n$  and  $p_{h,0}^{n+1} = p_h^n$

for  $k \geq 1$  (fixed-point iteration index)

(a) Given  $u_h^n$ ,  $c_h^n$  and  $u_{h,k-1}^{n+1}$ ,  $c_{h,k-1}^{n+1}$ . Determine  $u_{h,k}^{n+1}$  with (3.3).

(b) Given  $p_h^n$  and  $p_{h,k-1}^{n+1}$ . Determine  $c_{h,k}^{n+1}$  with (3.4).

(c) Given  $u_h^n$ ,  $c_h^n$ ,  $p_h^n$  and  $u_{h,k}^{n+1}$ ,  $c_{h,k}^{n+1}$ . Determine  $p_{h,k}^{n+1}$  with (3.4).

(d) if  $\left\{ \left\| u_{h,k}^{n+1} - u_{h,k-1}^{n+1} \right\|_{l^2}, \left\| c_{h,k}^{n+1} - c_{h,k-1}^{n+1} \right\|_{l^2}, \left\| p_{h,k}^{n+1} - p_{h,k-1}^{n+1} \right\|_{l^2} \right\} < Tol = 10^{-8}$  stop and set

$$u_h^{n+1} = u_{h,k}^{n+1}, \quad c_h^{n+1} = c_{h,k}^{n+1}, \quad p_h^{n+1} = p_{h,k}^{n+1},$$

increment  $n \mapsto n + 1$  and go back to step 2 (proceed to next time point)

(e) else set

$$u_{h,k}^{n+1} = \beta u_{h,k}^{n+1} + (1 - \beta) u_{h,k-1}^{n+1},$$

$$c_{h,k}^{n+1} = \beta c_{h,k}^{n+1} + (1 - \beta) c_{h,k-1}^{n+1},$$

$$p_{h,k}^{n+1} = \beta p_{h,k}^{n+1} + (1 - \beta) p_{h,k-1}^{n+1},$$

for some  $\beta \in [0, 1]$  and go to (a) and increment  $k \mapsto k + 1$  (next fixed-point iteration); here we set  $\beta = 0.5$ .

**Remark 3.2.** The system of algebraic equations of each equation at each step is solved using the sparse direct solver UMFPACK [14].

**Remark 3.3.** We notice that the relaxation parameter  $\beta$  can also be obtained via a backtracking procedure starting with  $\beta = 1$  and constructing a sequence with  $\beta \rightarrow 0$  for  $k \rightarrow \infty$  until convergence is achieved.

**Remark 3.4.** A rigorous numerical convergence analysis in weak function spaces of the discretized equations for  $\Delta t \rightarrow 0$  and  $h \rightarrow 0$  towards their continuous limits exceed the purpose of this paper and is left for future studies. However, there is hope for convergence in light of the classical solutions obtained in Section 2.

## 4. NUMERICAL SIMULATIONS

In order to illuminate the evolution of solutions and show their qualitative behavior, beyond the mere existence assertion of Theorem 1.1, in this section, we perform several numerical simulations in two and three spatial dimensions. The main objective are investigations of the influence of variations in the proliferation coefficient  $\mu$  and the haptotactic coefficient  $\chi$ , whose size did not matter for Theorem 1.1. The specific values are chosen for illustrative purposes, not due to their biological relevance. For a discussion of realistic diffusion and taxis coefficients of tumor cells see, for instance, [3].

### 4.1. Geometry, final time, parameters, and initial conditions

For all the experiments except those in Subsection 4.5, the computations are performed on the square domain  $\Omega = (0, 20)^2$ , discretized uniformly using quadrilateral elements. This mesh is uniformly refined 5 times at the beginning of the computation resulting into 1089 degrees of freedom. The final time is  $T = 50$ , we set  $\theta = 0.5$  and use as time step size  $\Delta t = 1$ . As initial conditions, we use

$$u_0(x) = \exp(-x^2), \quad c_0(x) = 1 - 0.5 \exp(-x^2), \quad p_0(x) = 0.5 \exp(-x^2),$$

in all computations, unless otherwise mentioned. As parameters, we use the fixed values  $\alpha = 10$  and  $\varepsilon = 0.2$ , while  $\mu$  and  $\chi$  are varied and specified in each respective subsection below. We notice that the smoothness conditions on the domain and satisfaction of the boundary conditions at the initial time  $t = 0$  are violated in this section in comparison to our theory established in Section 2.

Upfront, concerning the computational cost of the fixed-point scheme, in a computational analysis for all numerical examples, we observed the iteration numbers displayed in Table 1.

TABLE 1. Fixed-point iteration numbers in the numerical simulation in Section 4.

Section 4.2	$\mu = 10^{-10}$	$\mu = 0.5$	$\mu = 1.0$
# of iteration at $t = 1$	31	31	31
# of iteration at $t = 50$	24	22	22
Section 4.3	$\chi = 0.25$	$\chi = 0.75$	$\chi = 1.25$
# of iteration at $t = 1$	30	30	31
# of iteration at $t = 50$	26	26	-
Sections 4.4 and 4.5	Section 4.4: $\mu = \chi = 1$	Section 4.5(3d): $\mu = \chi = 1.0$	
# of iteration at $t = 1$	32	31	
# of iteration at $t = 50$	22	24	

### 4.2. Simulations for different proliferation coefficients $\mu$

First, we study the influence of the cancer cell proliferation coefficient on the cancer invasion for  $\mu = 10^{-10}$ , 0.5, 1.0 with small haptotactic rate  $\chi = 0.01$ . We notice that our theory in Section 2 requires  $\mu > 0$  and for this reason we made the previous choice  $\mu = 10^{-10}$ . Numerically we are interested in a value being close to zero in order to study the behavior of the cancer invasion model. Proliferation shows the ability of a cancer cell to copy its DNA and divide into 2 cells, therefore an increase in the proliferation rate of tumours causes an accelerated invasion of cancer cells into connective tissues domain. In all the computations we use  $\alpha^{-1} = 0.1$ ,  $\varepsilon = 0.2$ .

The results obtained with the standard Galerkin discretization of the system (1.1) are displayed in Figures 1–6, at time instances 5, 15, 25 and 35. The snapshots of cancer cell invasion, connective tissue and protease are plotted in Figures 1, 3 and 5. We start with  $\mu = 10^{-10}$ , that is, almost no growth in the cancer cell density. As we can see from Figure 1, there is no growth in the cancer during the initial stage, and despite a small amount of concentration at the initial period, the cancer cell density and also protease (which is produced by cancer cells upon contact with connective tissues) are decreased and spread slowly due to diffusion effect and the invasion does not continue after time  $t = 15$ . Now, let us consider  $\mu = 0.5$ . As we can see from Figure 3, in this case, an increase of the concentration of cancer cells becomes visible, and it continues during the time. The cancer invasion gradually increases and degrades nearly half of the connective tissue by the time  $t = 25$ . For  $\mu = 1.0$ , Figure 5 shows the growth effect. Due to high proliferation rate, cancer cells produce more protease, which helps them to invade the connective tissues area rapidly. In particular, cancer cells complete invasion in three-quarters of the connective tissue domain at  $t = 25$  when  $\mu = 1.0$  is used. The snapshots of cancer cell invasion for different values of proliferation rate are given in Figures 2, 4 and 6. As explained, by increasing the value of  $\mu$ , cancer cells increase and the invasion happens more rapidly for all the considered time intervals.

### 4.3. Effects of the haptotactic coefficient $\chi$

In this subsection, we consider the effect of the haptotactic coefficient on the connective cells degeneration by varying  $\chi$ . We choose  $\chi = 0.25, 0.75, 1.25$  with small proliferation rate  $\mu = 0.01$ , diffusion coefficient  $\alpha^{-1} = 0.1$  and  $\varepsilon = 0.2$ . The effects of haptotactic coefficient at different time instances are depicted in Figures 7–12. Starting with  $\chi = 0.25$ , the snapshot at  $t = 5$  shows that the cancer cells reduce at the origin and start migrating towards the direction of the gradient of connective tissue. The migration of the cancer cells becomes more clear and the effect of haptotaxis can be clearly seen at  $t = 5$  in Figures 9 and 10, where the small cluster of cancer cells is created and spreads further by time. Increasing the amount of  $\chi$  accelerates the cancer cells

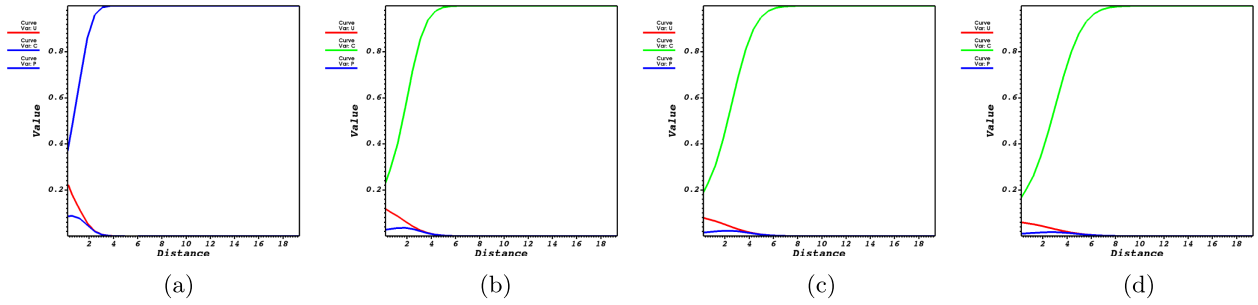


FIGURE 1. The effect of the proliferation rate on cancer cell invasion  $u$ , connective tissue  $c$  and protease  $p$  at different time instants,  $t = 5, 15, 25, 35$  for  $\mu = 10^{-10}$ . The functions are plotted along the line  $y = x$ . (a)  $t = 5$ , (b)  $t = 15$ , (c)  $t = 25$  and (d)  $t = 35$ .

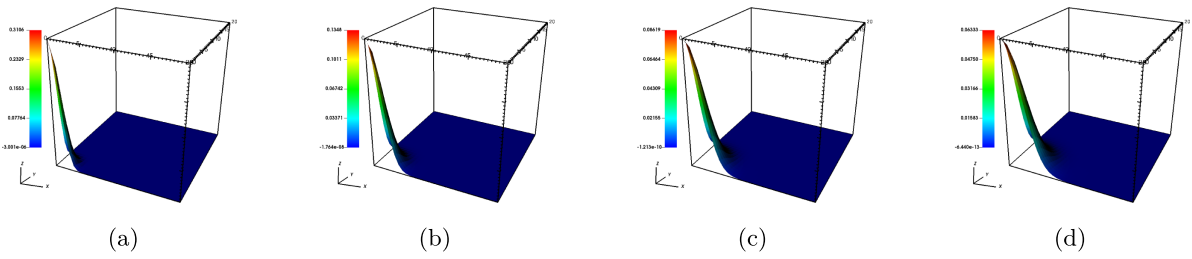


FIGURE 2. The snapshots of cancer cell invasion  $u$  for  $\mu = 10^{-10}$ , the maximum amount of cancer cells decreasing from left to right is 0.3106, 0.1348, 0.08619, and 0.06333. The color scale in the legend is not fixed in order to display better the current shape. (a)  $t = 5$ , (b)  $t = 15$ , (c)  $t = 25$  and (d)  $t = 35$ .

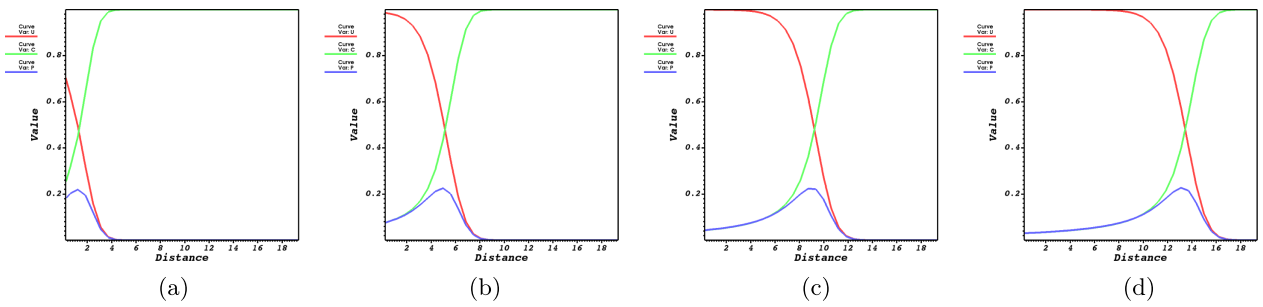


FIGURE 3. The effect of proliferation rate on cancer cell invasion  $u$ , connective tissue  $c$  and protease  $p$  at different time instants,  $t = 5, 15, 25, 35$  for  $\mu = 0.5$ . The functions are plotted along the line  $y = x$ . (a)  $t = 5$ , (b)  $t = 15$ , (c)  $t = 25$  and (d)  $t = 35$ .

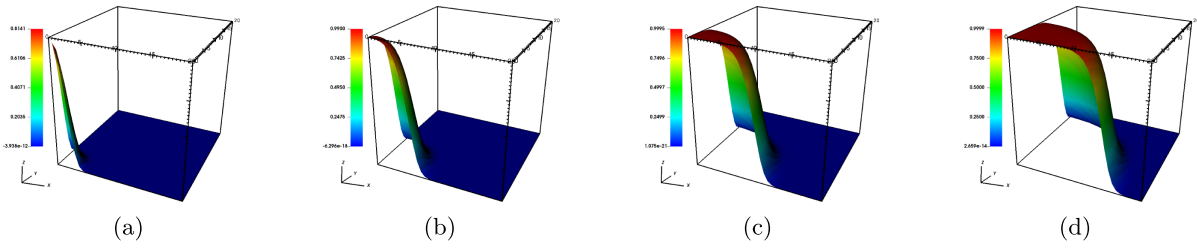


FIGURE 4. The snapshots of cancer cell invasion  $u$  for  $\mu = 0.5$ . The color scale in the legend is not fixed. (a)  $t = 5$ , (b)  $t = 15$ , (c)  $t = 25$  and (d)  $t = 35$ .

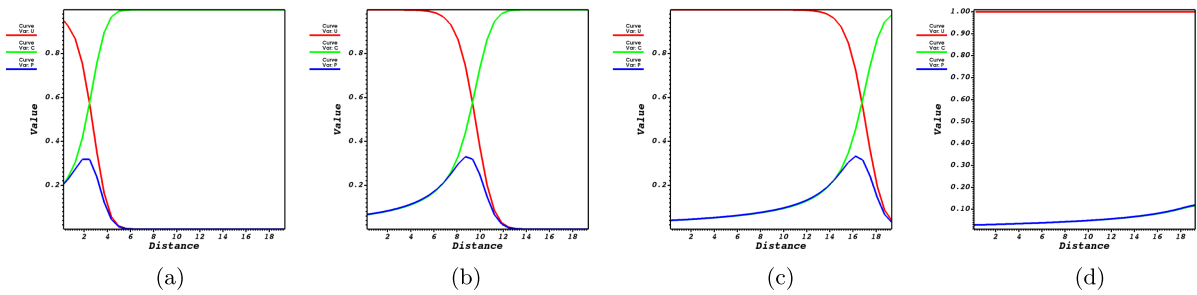


FIGURE 5. The effect of proliferation rate on cancer cell invasion, connective tissue and protease at different time instants,  $t = 5, 15, 25, 35$  for  $\mu = 1.0$ . The functions are plotted along the line  $y = x$ . (a)  $t = 5$ , (b)  $t = 15$ , (c)  $t = 25$  and (d)  $t = 35$ .

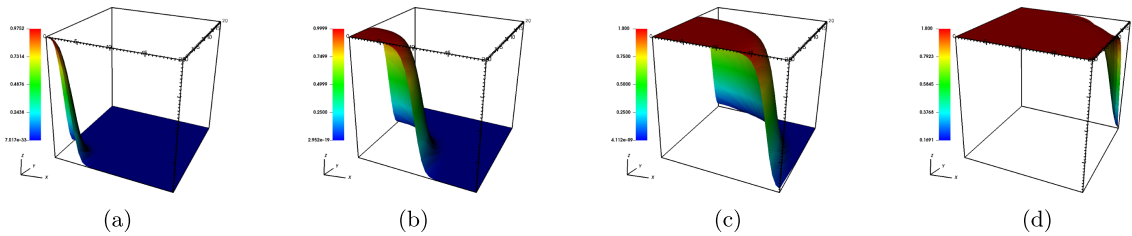


FIGURE 6. The snapshots of cancer cell invasion  $u$  for  $\mu = 1.0$ . The color scale in the legend is not fixed. (a)  $t = 5$ , (b)  $t = 15$ , (c)  $t = 25$  and (d)  $t = 35$ .

migration and the cancer cells should move toward the boundary of the domain quickly, but as we can see from Figures 11 to 12, oscillations start at  $t = 5$  and the numerical simulation breaks down for  $\chi = 1.25$ .

#### 4.4. Identical proliferation and haptotactic coefficients

In this subsection, we consider the case when the proliferation rate is equal to haptotactic coefficient, *i.e.*,  $\mu = \chi = 1$ , and all other parameters are the same as in the previous subsections. As it can be seen from Figures 13 and 14, due to the proliferation rate, the concentration of cancer grows quickly even from the beginning resulting from a high amount of haptotaxis, therefore the tumour migrates rapidly inside the domain and degrades the connective tissue in a much shorter amount of time.

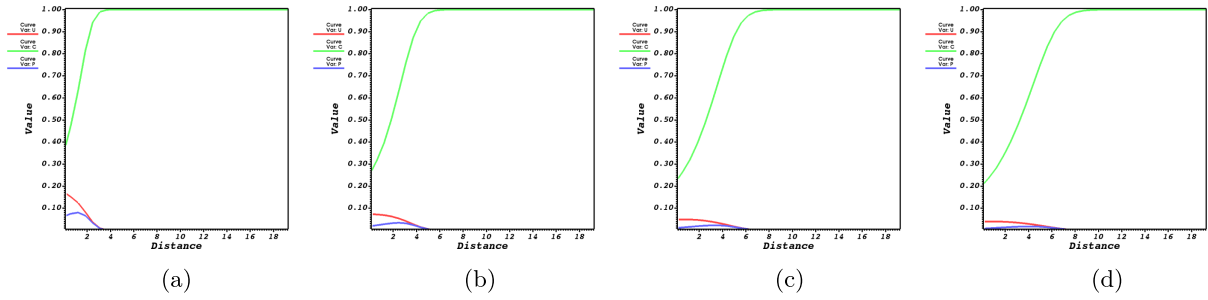


FIGURE 7. Haptotactic effect on cancer cell invasion, connective tissue and protease at different time instants,  $t = 5, 15, 25, 35$  for  $\chi = 0.25$ . (a)  $t = 5$ , (b)  $t = 15$ , (c)  $t = 25$  and (d)  $t = 35$ .

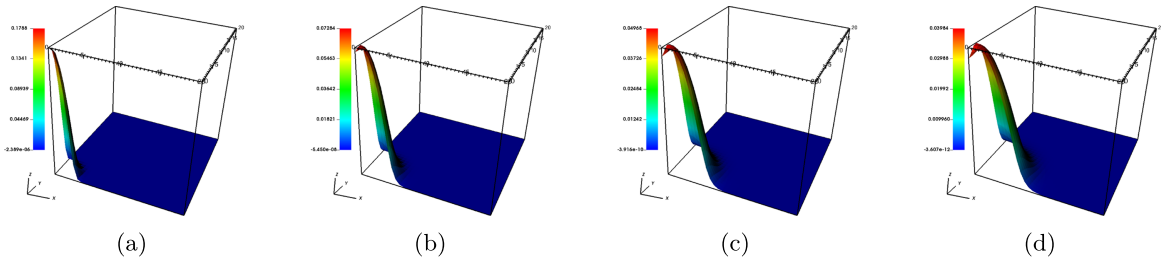


FIGURE 8. The snapshots of cancer cell invasion  $u$  for  $\chi = 0.25$ , the maximum amount of cancer cells decreasing from left to right is 0.1788, 0.07284, 0.04968, and 0.03984. The color scale in the legend is not fixed in order to display better the current shape. (a)  $t = 5$ , (b)  $t = 15$ , (c)  $t = 25$  and (d)  $t = 35$ .

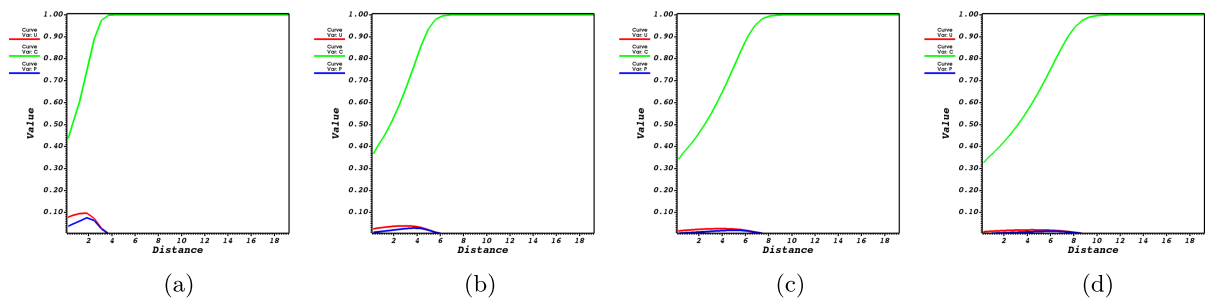


FIGURE 9. Haptotactic effect on cancer cell invasion, connective tissue and protease at different time instants,  $t = 5, 15, 25, 35$  for  $\chi = 0.75$ . (a)  $t = 5$ , (b)  $t = 15$ , (c)  $t = 25$  and (d)  $t = 35$ .

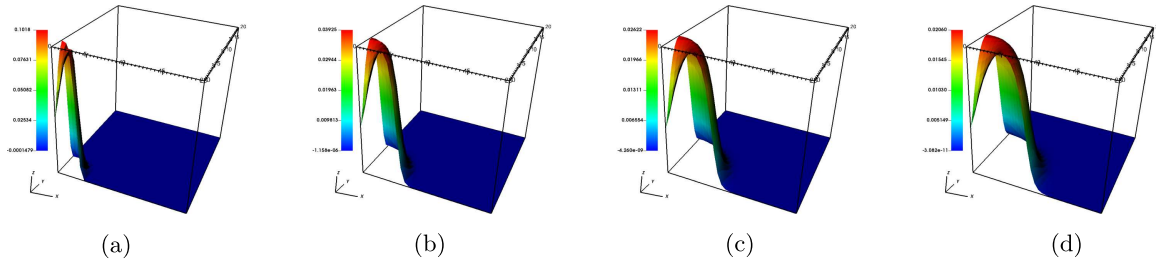


FIGURE 10. The snapshots of cancer cell invasion  $u$  for  $\chi = 0.75$ , the maximum amount of cancer cells decreasing from left to right is 0.1018, 0.03925, 0.02622, and 0.02060. The color scale in the legend is not fixed in order to display better the current shape. (a)  $t = 5$ , (b)  $t = 15$ , (c)  $t = 25$  and (d)  $t = 35$ .

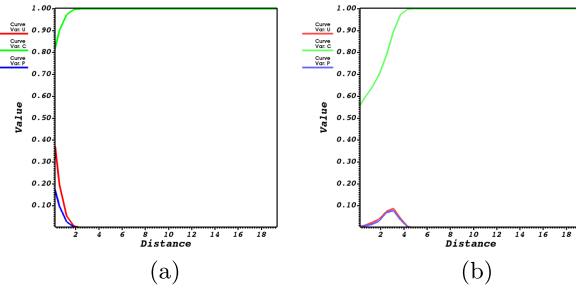


FIGURE 11. Haptotactic effect on cancer cell invasion, connective tissue and protease at different time instants,  $t = 0, 5$  for  $\chi = 1.25$ . (a)  $t = 0$  and (b)  $t = 5$ .

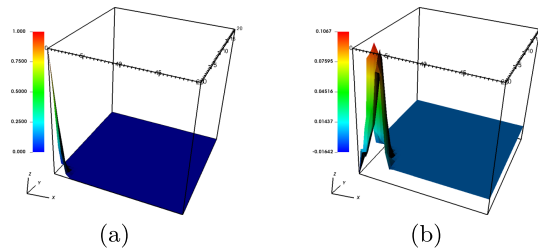


FIGURE 12. The snapshots of cancer cell invasion  $u$  for  $\chi = 1.25$ , with maximal values 1.0 and 0.1067, respectively. The color scale in the legend is not fixed in order to display better the current shape. (a)  $t = 0$  and (b)  $t = 5$ .



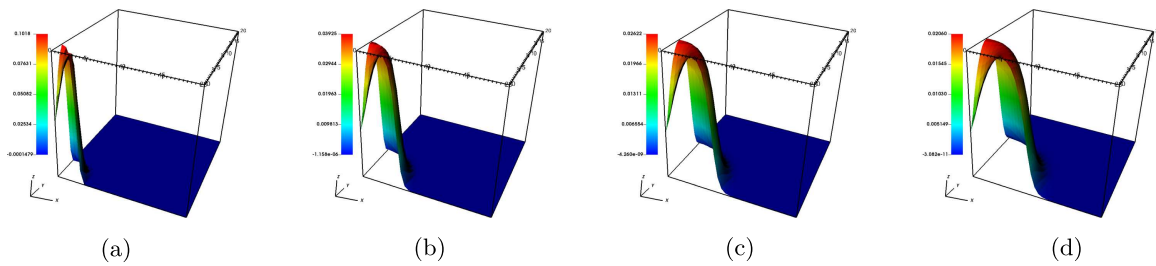


FIGURE 13. Degradation of connective tissue  $c$  for  $\chi = 1.0, \mu = 1.0$  at different time instants  $t = 0, 10, 20$  and  $30$ . The color scale in the legend is not fixed. (a)  $t = 0$ , (b)  $t = 10$ , (c)  $t = 20$  and (d)  $t = 30$ .

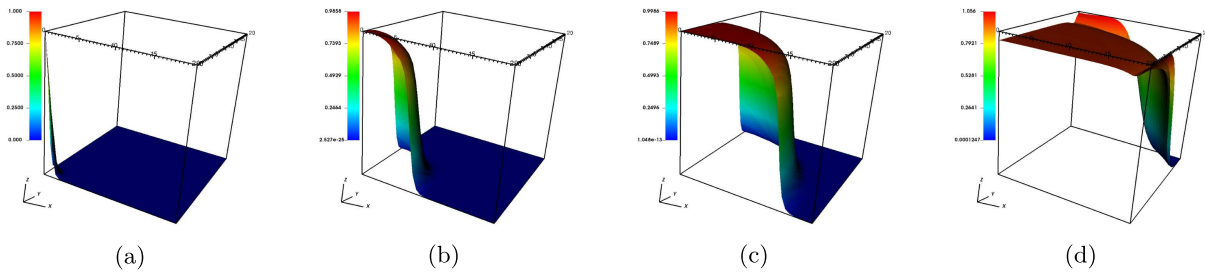


FIGURE 14. Invasion of cancer cells  $u$  for  $\chi = 1.0, \mu = 1.0$  at different time instants  $t = 0, 10, 20$  and  $30$ . The color scale in the legend is not fixed in order to display better the current shape. (a)  $t = 0$ , (b)  $t = 10$ , (c)  $t = 20$  and (d)  $t = 30$ .

#### 4.5. Three dimensional simulations

In this final subsection, we perform numerical simulations in three spatial dimensions to consider some more realistic movement. Here, the experiments are performed on a mesh with 32 768 hexahedral elements covering the domain  $\Omega$ . Figures 15 and 16 show the snapshots of cancer cells and connective tissues for growth rate  $\mu = 1$  and haptotactic coefficient  $\chi = 1$ . Further, we use the parameters  $\alpha^{-1} = 0.1$  and  $\varepsilon = 0.2$ . As it can be seen, at  $t = 5$  the connective tissue covers the entire domain and only a small amount of cancer cells exists at the corner, by the time cancer cells growth and invade the domain of connective tissue quickly and by  $t = 35$  almost all the domain is occupied by cancer cells.

### 5. CONCLUSIONS

In this paper, we established theoretical proofs, numerical algorithms, implementations and numerical simulations for a cancer invasion model. In our theoretical part, existence of global classical solutions in both two- and three-dimensional bounded domains was established. In the proofs, we employed the fact that the second and third equation in (1.1) at least regularize in time. For showing boundedness in  $L^\infty$ , the comparison principle allowed us to conclude boundedness in small time intervals, which then was iteratively applied to obtain the result also for larger times. For the spatial derivatives, we secondly applied a testing procedure for deriving estimates valid on small time intervals, again followed by an iteration procedure. Parabolic regularity theory yielded global existence of the solutions.

The numerical stability of the system heavily depends on the haptotactic coefficient  $\chi$ . By fixing proliferation rate  $\mu$  and varying the  $\chi$  one can make either the diffusion or transport of the cells dominant. The later usually gives rise to spurious oscillations or numerical blow up in the system. In order to study such properties, (1.1)

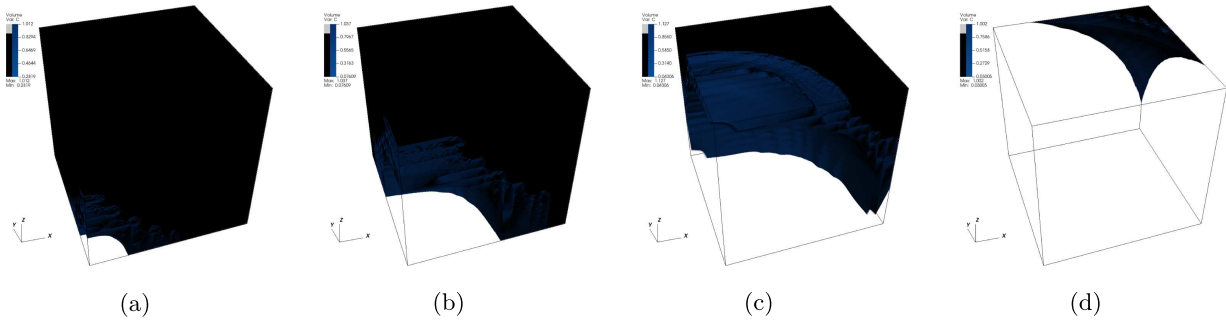


FIGURE 15. Degradation of connective tissue  $c$  for  $\chi = 1.0, \mu = 1.0$  at different time instants  $t = 5, 15, 25$  and  $35$ . The color scale in the legend is not fixed in order to display better the current shape. (a)  $t = 5$ , (b)  $t = 15$ , (c)  $t = 25$  and (d)  $t = 35$ .

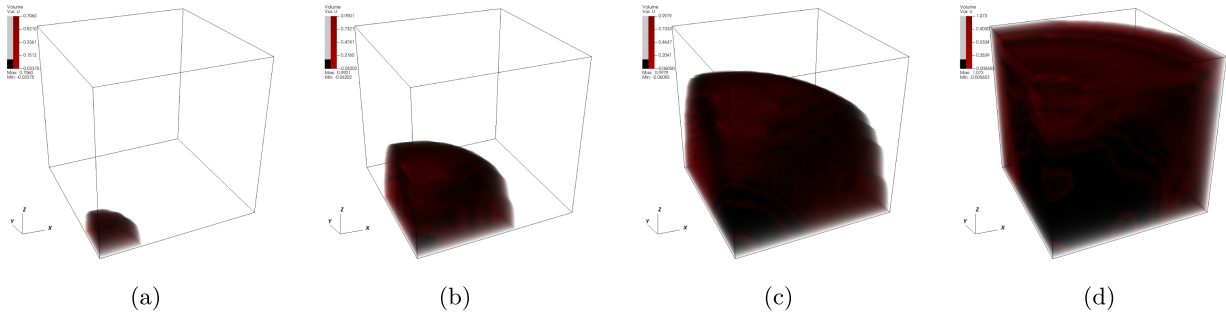


FIGURE 16. Invasion of cancer cells  $u$  for  $\chi = 1.0, \mu = 1.0$  at different time instants  $t = 5, 15, 25$  and  $35$ . The color scale in the legend is not fixed in order to display better the current shape. (a)  $t = 5$ , (b)  $t = 15$ , (c)  $t = 25$  and (d)  $t = 35$ .

was discretized using finite differences in time and Galerkin finite elements in space. A fixed-point scheme was designed to decouple the three equations, yielding a robust nonlinear procedure. These developments and their implementation allowed us to study numerically variations in  $\mu$  and  $\chi$  in two and three spatial dimensions and to illustrate our theoretical results.

Compared to other models, the system in this article did not feature any spatial regularizing effects in the third (or second) equation. This was based on the modelling in [34], where it was argued that there should be no diffusion term for the protease equation. Key challenges both in the analytical and numerical part are precisely caused by this biologically motivated choice. Related works treating systems including a diffusion term also for the third equation crucially make use of the corresponding smoothing effects – a direct adaptation of their methods would evidently have been insufficient for the system at hand.

As to future work, we notice that higher parameter variations resulting into convection-dominated regimes, require the design and implementation of stabilization methods such as streamline upwind Petrov–Galerkin stabilizing formulations or algebraic flux corrected transport. This would introduce additional terms in the equation (3.3) of our nonlinear fixed-point scheme. In case of an algebraic stabilization, an additional nonlinearity would be possibly created since it involves limiters that often depend on the unknown discrete solution. Nevertheless, this nonlinearity can be treated in the framework of the considered fixed-point iterations so that it does not increase the computational cost significantly.

*Acknowledgements.* The work of Shahin Heydari has been supported through the grant No. 396921 of the Charles University Grant Agency and Charles University Mobility Fund No. 2-068. She also would like to thank Institute of Applied Mathematics and the Leibniz University Hannover for their hospitality during the six months stay from November 2021 to April 2022. The work of Petr Knobloch was supported through the grant No. 22-01591S of the Czech Science Foundation.

## REFERENCES

- [1] A. Amoddeo, Adaptive grid modelling for cancer cells in the early stage of invasion. *Comput. Math. Appl.* **69** (2015) 610–619.
- [2] A.R. Anderson, A hybrid mathematical model of solid tumour invasion: the importance of cell adhesion. *Math. Med. Biol.* **22** (2005) 163–186.
- [3] A.R.A. Anderson and M.A.J. Chaplain, Continuous and discrete mathematical models of tumor-induced angiogenesis. *Bull. Math. Biol.* **60** (1998) 857–900.
- [4] A.R.A. Anderson, M.A.J. Chaplain, E.L. Newman, R.J.C. Steele and A.M. Thompson, Mathematical modelling of tumour invasion and metastasis. *Comput. Math. Methods Med.* **2** (2000) 129–154.
- [5] D. Arndt, W. Bangerth, T.C. Clevenger, D. Davydov, M. Fehling, D. Garcia-Sanchez, G. Harper, T. Heister, L. Heltai, M. Kronbichler, R.M. Kynch, M. Maier, J.-P. Pelteret, B. Turcksin and D. Wells, The deal.II library, version 9.1. *J. Numer. Math.* **27** (2019) 203–213.
- [6] D. Arndt, W. Bangerth, D. Davydov, T. Heister, L. Heltai, M. Kronbichler, M. Maier, J.-P. Pelteret, B. Turcksin and D. Wells, The deal.II finite element library: design, features, and insights. *Comput. Math. Appl.* **81** (2021) 407–422.
- [7] S. Aznavoorian, M.L. Stracke, H. Krutzsch, E. Schiffmann and L.A. Liotta, Signal transduction for chemotaxis and haptotaxis by matrix molecules in tumor cells. *J. Cell Biol.* **110** (1990) 1427–1438.
- [8] M.A.J. Chaplain and G. Lolas, Mathematical modelling of cancer cell invasion of tissue: the role of the urokinase plasminogen activation system. *Math. Models Methods Appl. Sci.* **15** (2005) 1685–1734.
- [9] M.A.J. Chaplain and G. Lolas, Mathematical modelling of cancer invasion of tissue: dynamic heterogeneity. *Netw. Heterog. Media* **1** (2006) 399–439.
- [10] M. Chapwanya, J.M.-S. Lubuma and R.E. Mickens, Positivity-preserving nonstandard finite difference schemes for cross-diffusion equations in biosciences. *Comput. Math. Appl.* **68** (2014) 1071–1082.
- [11] A. Chertock and A. Kurganov, A second-order positivity preserving central-upwind scheme for chemotaxis and haptotaxis models. *Numer. Math.* **111** (2008) 169–205.
- [12] P.G. Ciarlet, The finite element method for elliptic problems, in *Studies in Mathematics and its Applications*. Vol. 4. North-Holland Publishing Co., Amsterdam, New York, Oxford (1978).
- [13] L. Corrias, B. Perthame and H. Zaag, Global solutions of some chemotaxis and angiogenesis systems in high space dimensions. *Milan J. Math.* **72** (2004) 1–28.
- [14] T.A. Davis and I.S. Duff, An unsymmetric-pattern multifrontal method for sparse LU factorization. *SIAM J. Matrix Anal. Appl.* **18** (1997) 140–158.
- [15] P. Domschke, D. Trucu, A. Gerisch and M.A.J. Chaplain, Mathematical modelling of cancer invasion: implications of cell adhesion variability for tumour infiltrative growth patterns. *J. Theoret. Biol.* **361** (2014) 41–60.
- [16] Y. Epshteyn, Discontinuous Galerkin methods for the chemotaxis and haptotaxis models. *J. Comput. Appl. Math.* **224** (2009) 168–181.
- [17] A. Friedman, *Partial Differential Equations*. R.E. Krieger Pub. Co, Huntington, NY (1976).
- [18] M. Fuest, Global solutions near homogeneous steady states in a multidimensional population model with both predator- and prey-taxis. *SIAM J. Math. Anal.* **52** (2020) 5865–5891.
- [19] A. Gerisch and M.A.J. Chaplain, Mathematical modelling of cancer cell invasion of tissue: local and non-local models and the effect of adhesion. *J. Theoret. Biol.* **250** (2008) 684–704.
- [20] Y. Giga and H. Sohr, Abstract  $L^p$  estimates for the Cauchy problem with applications to the Navier–Stokes equations in exterior domains. *J. Funct. Anal.* **102** (1991) 72–94.
- [21] D. Hanahan and R.A. Weinberg, The hallmarks of cancer. *Cell* **100** (2000) 57–70.
- [22] M. Khalsaraei, S. Heydari and L.D. Algoo, Positivity preserving nonstandard finite difference schemes applied to cancer growth model. *J. Cancer Treat. Res.* **4** (2016) 27–33.
- [23] M. Kolev and B. Zubik-Kowal, Numerical solutions for a model of tissue invasion and migration of tumour cells. *Comput. Math. Methods Med.* **2011** (2011).
- [24] O.A. Ladyženskaja, V.A. Solonnikov and N.N. Ural’ceva, *Linear and quasi-linear equations of parabolic type*, in *Translations of Mathematical Monographs*. Vol. 3. American Mathematical Society, Providence, RI (1988).
- [25] J. Lankeit and M. Winkler, Facing low regularity in chemotaxis systems. *Jahresber. Dtsch. Math. Ver.* **122** (2019) 35–64.
- [26] G.M. Lieberman, Hölder continuity of the gradient of solutions of uniformly parabolic equations with conormal boundary conditions. *Ann. Mat. Pura Appl.* **148** (1987) 77–99.
- [27] G.M. Lieberman, *Second Order Parabolic Differential Equations*. World Scientific Publishing Co., Inc., River Edge, NJ (1996).
- [28] G. Liţcanu and C. Morales-Rodrigo, Asymptotic behavior of global solutions to a model of cell invasion. *Math. Models Methods Appl. Sci.* **20** (2010) 1721–1758.

- [29] J.S. Lowengrub, H.B. Frieboes, F. Jin, Y.-L. Chuang, X. Li, P. Macklin, S.M. Wise and V. Cristini, Nonlinear modelling of cancer: bridging the gap between cells and tumours. *Nonlinearity* **23** (2010) R1–R91.
- [30] B.P. Marchant, J. Norbury and A.J. Perumpanani, Travelling shock waves arising in a model of malignant invasion. *SIAM J. Appl. Math.* **60** (2000) 463–476.
- [31] B.P. Marchant, J. Norbury and J.A. Sherratt, Travelling wave solutions to a haptotaxis-dominated model of malignant invasion. *Nonlinearity* **14** (2001) 1653–1671.
- [32] L. Nirenberg, On elliptic partial differential equations. *Ann. Della Scuola Norm. Super. Pisa Cl. Sci., Ser. 3* **13** (1959) 115–162.
- [33] A.J. Perumpanani and H.M. Byrne, Extracellular matrix concentration exerts selection pressure on invasive cells. *Eur. J. Cancer* **35** (1999) 1274–1280.
- [34] A.J. Perumpanani, J.A. Sherratt, J. Norbury and H.M. Byrne, A two parameter family of travelling waves with a singular barrier arising from the modelling of extracellular matrix mediated cellular invasion. *Phys. D* **126** (1999) 145–159.
- [35] M. Rasche and C. Ziti, Finite time blow-up in some models of chemotaxis. *J. Math. Biol.* **33** (1995) 388–414.
- [36] N. Sfakianakis and M.A.J. Chaplain, Mathematical modelling of cancer invasion: a review, in International Conference by Center for Mathematical Modeling and Data Science, Osaka University, Springer (2020) 153–172.
- [37] R. Strehl, A. Sokolov, D. Kuzmin, D. Horstmann and S. Turek, A positivity-preserving finite element method for chemotaxis problems in 3D. *J. Comput. Appl. Math.* **239** (2013) 290–303.
- [38] C. Surulescu and M. Winkler, Does indirectness of signal production reduce the explosion-supporting potential in chemotaxis–haptotaxis systems? Global classical solvability in a class of models for cancer invasion (and more). *Eur. J. Appl. Math.* **32** (2021) 618–651.
- [39] Y. Tao and M. Winkler, Energy-type estimates and global solvability in a two-dimensional chemotaxis–haptotaxis model with remodeling of non-diffusible attractant. *J. Differ. Equ.* **257** (2014) 784–815.
- [40] Y. Tao and G. Zhu, Global solution to a model of tumor invasion. *Appl. Math. Sci.* **1** (2007) 2385–2398.
- [41] J. Valenciano and M.A.J. Chaplain, Computing highly accurate solutions of a tumour angiogenesis model. *Math. Models Methods Appl. Sci.* **13** (2003) 747–766.
- [42] Ch. Walker and G.F. Webb, Global existence of classical solutions for a haptotaxis model. *SIAM J. Math. Anal.* **38** (2007) 1694–1713.
- [43] T. Wick, Solving monolithic fluid-structure interaction problems in arbitrary Lagrangian Eulerian coordinates with the deal.II library. *Arch. Numer. Soft.* **1** (2013) 1–19.
- [44] X. Zheng, S. Wise and V. Cristini, Nonlinear simulation of tumor necrosis, neo-vascularization and tissue invasion via an adaptive finite-element/level-set method. *Bull. Math. Biol.* **67** (2005) 211–259.

**Please help to maintain this journal in open access!**



This journal is currently published in open access under the Subscribe to Open model (S2O). We are thankful to our subscribers and supporters for making it possible to publish this journal in open access in the current year, free of charge for authors and readers.

Check with your library that it subscribes to the journal, or consider making a personal donation to the S2O programme by contacting [subscribers@edpsciences.org](mailto:subscribers@edpsciences.org).

More information, including a list of supporters and financial transparency reports, is available at <https://edpsciences.org/en/subscribe-to-open-s2o>.

### 3. Papers II and III

This chapter is based on the paper entitled "Flux-corrected transport stabilization of an evolutionary cross-diffusion cancer invasion model", published in Journal of Computational Physics and the paper entitled "Solvability and numerical solution of a cross-diffusion cancer invasion model, accepted for publication in the proceeding of the ENUMATH 2023 conference.

#### 3.1 Flux-corrected transport stabilization of an evolutionary cross-diffusion cancer invasion model

In this paper we considered a haptotactic counterpart of the problem given in the previous chapter as follows:

$$\begin{cases} u_t = -\chi \nabla \cdot (u \nabla c) + \mu u(1 - u) & \text{in } \Omega \times (0, T], \\ c_t = -pc & \text{in } \Omega \times (0, T], \\ p_t = \frac{1}{\epsilon}(uc - p) & \text{in } \Omega \times (0, T], \\ u \frac{\partial c}{\partial n} = 0 & \text{on } \partial\Omega \times (0, T], \\ (u, c, p)(\cdot, 0) = (u_0, c_0, p_0). & \text{in } \Omega, \end{cases} \quad (3.1)$$

As it can be seen, there is no diffusion term in the system and therefore the technique used in the previous section is not applicable anymore. This leaves the question of proving the solvability of this particular problem unanswered from the analytical point of view. Though, we addressed this point from the numerical perspective. Then we showed that, when the convective (or haptotactic) part of the system is dominant, the standard methods for the studied system may become unstable. Next, we employed a high-resolution nonlinear flux-corrected transport method along with an implicit  $\theta$ -method for spatial and temporal discretization, respectively. Using a consequence of Brouwer's fixed point theorem, we then proved that both the nonlinear scheme and its linearized version obtained using the fixed point-iteration are solvable and positivity-preserving. Finally, the numerical analysis were supported by carrying out several numerical simulations in 2D.

Our main results are summarized as follows: After the aforementioned double discretization process, the fully discrete implicit version of (3.1) has the form

$$c_i^{n+1} = c_i^n e^{-\tau_{n+1}(p_i^{n+1} + p_i^n)/2}, \quad (3.2)$$

$$\begin{aligned} p_i^{n+1} = e^{-\tau_{n+1}/\epsilon} p_i^n + \frac{1}{\tau_{n+1}^2} & \left\{ \left( u_i^{n+1} (\epsilon - \tau_{n+1}) - u_i^n \epsilon \right) \left( c_i^{n+1} (\epsilon - \tau_{n+1}) - c_i^n \epsilon \right) \right. \\ & - \left( u_i^{n+1} \epsilon - u_i^n (\epsilon + \tau_{n+1}) \right) \left( c_i^{n+1} \epsilon - c_i^n (\epsilon + \tau_{n+1}) \right) e^{-\tau_{n+1}/\epsilon} \\ & \left. + (u_i^{n+1} - u_i^n)(c_i^{n+1} - c_i^n) \epsilon^2 (1 - e^{-\tau_{n+1}/\epsilon}) \right\}, \end{aligned} \quad (3.3)$$

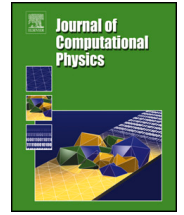
$$(\mathbb{M}_L + \theta \tau_{n+1} \mathbb{L}^{n+1}) \mathbf{u}^{n+1} = (\mathbb{M}_L - (1 - \theta) \tau_{n+1} \mathbb{L}^n) \mathbf{u}^n + \left( \sum_{j=1}^M \alpha_{ij}^{n+1} f_{ij}^{n+1} \right)_{i=1}^M. \quad (3.4)$$

For this discretization the following result holds:

**Theorem.** Consider any  $n \in \{0, \dots, N - 1\}$  and let  $\mathbf{u}^n, \mathbf{c}^n, \mathbf{p}^n \in \mathbb{R}^M$  satisfy  $\mathbf{u}^n \geq 0$ ,  $1 \geq \mathbf{c}^n \geq 0$ ,  $\mathbf{p}^n \geq 0$ . Let the time step  $\tau_{n+1}$  satisfy the conditions

$$(1 - \theta) \tau_{n+1} l_{ii}^n \leq m_i, \quad \theta \tau_{n+1} \left( \mu m_i + \chi n_v \kappa^2 \sum_{K \ni x_i} h_K^{d-2} \right) < m_i, \quad i = 1, \dots, M,$$

where  $(l_{ii}^n)_{i=1}^M$  is the diagonal of  $\mathbb{L}^n$ ,  $n_v$  is the number of vertices of a cell in  $\mathcal{T}_h$ , and  $\kappa$  is a constant satisfying  $\|\nabla \phi_i\|_{L^2(K)} \leq \kappa h_K^{d/2-1}$  for any  $K \in \mathcal{T}_h$  and  $i = 1, \dots, M$ . Then there exist vectors  $\mathbf{u}^{n+1}, \mathbf{c}^{n+1}, \mathbf{p}^{n+1} \in \mathbb{R}^M$  satisfying (3.4), (3.2), (3.3) where the limiters  $\alpha_{ij}^{n+1}$  are computed using the Zalesak algorithm from the fluxes  $f_{ij}^{n+1}$ . Moreover, these vectors satisfy  $\mathbf{u}^{n+1} \geq 0$ ,  $1 \geq \mathbf{c}^{n+1} \geq 0$ , and  $\mathbf{p}^{n+1} \geq 0$ .



# Flux-corrected transport stabilization of an evolutionary cross-diffusion cancer invasion model

Shahin Heydari<sup>a,\*</sup>, Petr Knobloch<sup>a</sup>, Thomas Wick<sup>b</sup>

<sup>a</sup> Charles University, Faculty of Mathematics and Physics, Sokolovská 83, 18675 Praha 8, Czech Republic

<sup>b</sup> Leibniz University Hannover, Institute of Applied Mathematics, Welfengarten 1, 30167 Hannover, Germany

## ARTICLE INFO

### Keywords:

Cancer invasion  
Cross-diffusion equation  
FEM-FCT stabilization  
Positivity preservation  
Existence of solutions

## ABSTRACT

In the present work, we investigate a model of the invasion of healthy tissue by cancer cells which is described by a system of nonlinear PDEs consisting of a cross-diffusion-reaction equation and two additional nonlinear ordinary differential equations. We show that when the convective part of the system, the haptotaxis term, is dominant, then straightforward numerical methods for the studied system may be unstable. We present an implicit finite element method using conforming  $P_1$  or  $Q_1$  finite elements to discretize the model in space and the  $\theta$ -method for discretization in time. The discrete problem is stabilized using a nonlinear flux-corrected transport approach. It is proved that both the nonlinear scheme and the linearized problems used in fixed-point iterations are solvable and positivity preserving. Several numerical experiments are presented in 2D to demonstrate the performance of the proposed method.

## 1. Introduction

Keller and Segel [1,2] proposed the first mathematical model for description of chemotactical processes. Chemotaxis refers to the motion in the direction to (or away from) the position of higher concentration based on the gradient of chemical substances and its chemotacticity character which controls the speed of this motion. Their model has been widely extended and followed to develop more sophisticated and complex chemotaxis models and played a vitally important role in many areas of science, in particular in medical and biological applications, for example, bacteria and cell aggregation [3–5], tumor angiogenesis and invasion [6–9], biological pattern formation [10,5], and immune cell migration [11]. From the analytical point of view, mathematical analysis for chemotaxis systems of equations is a challenge and causes many questions especially in the context of the existence and uniqueness of solutions. In the last three decades, many researchers have been actively involved and answered some of these questions [12–17]. From the numerical point of view, so far a great deal of research on chemotaxis models has been done in various areas, including the finite difference method [7,8,18], discontinuous Galerkin method [19,20], finite element method [21–24], finite volume method [25], operator-splitting methods [26], or fractional step algorithms [27]. However, many analytical and numerical aspects are still untouched and call for further investigation.

The chemotaxis problems are usually strongly coupled nonlinear systems of equations whose solutions represent concentrations or densities and need to be non-negative in order to satisfy the physics behind the system. Hence, it is difficult to construct an efficient and accurate numerical method that does not produce solutions with negative values. Another interesting aspect is the

\* Corresponding author.

E-mail addresses: [heydari@karlin.mff.cuni.cz](mailto:heydari@karlin.mff.cuni.cz) (S. Heydari), [knobloch@karlin.mff.cuni.cz](mailto:knobloch@karlin.mff.cuni.cz) (P. Knobloch), [thomas.wick@ifam.uni-hannover.de](mailto:thomas.wick@ifam.uni-hannover.de) (T. Wick).

<https://doi.org/10.1016/j.jcp.2023.112711>

Received 16 July 2023; Received in revised form 31 October 2023; Accepted 12 December 2023

Available online 14 December 2023

0021-9991/© 2023 Elsevier Inc. All rights reserved.

singular, spiky and oscillatory behavior of the solutions. In particular, when the chemotaxis term dominates the diffusion and reaction terms, standard discretization methods typically provide nonphysical oscillatory numerical solutions. To overcome this problem, stabilization methods can be applied. Up to now, many scientists used flux-corrected transport (FCT) algorithms, i.e., nonlinear high-resolution schemes introduced by Boris and Book [28–30], later developed based on linear finite element discretizations by Kuzmin, Löhner et al. [31–34], and further extended to linear and nonlinear space-time FEM-FCT in [35]. In [36], an implicit flux-corrected transport scheme was developed and applied to three benchmark examples of the general Keller–Segal model in two spatial dimensions. It was shown that the proposed method is positivity preserving and sufficiently accurate, even in the cases where solutions blow up in the center or at the boundary of the domain. The investigations of the blow-up behavior of the solutions were further extended to three spatial dimensions in [37]. In [38,39], an FEM-FCT scheme was coupled with a level-set method to obtain positivity preserving solutions on a stationary surface and evolving-in-time surfaces. It was shown that the proposed method is able to produce accurate numerical solutions, which makes it possible to couple the partial differential equations defined on a specific domain with the PDEs that are defined on the surface of this domain. This scheme was further used with operator-splitting techniques to solve chemotaxis models in 3D. The operator-splitting method splitted a 3D problem into a sequence of 1D subproblems and the FEM-FCT algorithm was used to solve each 1D subproblem separately [40]. In [41], the authors used an efficient adaptive moving mesh finite element approach based on the parabolic Monge–Ampère method for determining the coordinate transformation for the adaptive mesh combined with an FCT scheme which guarantees the non-negativity of the solutions. As a result, the computational cost was significantly reduced. All aforementioned techniques were also applied to the same benchmark examples. A different case was studied in [42,43], where the authors used the pressure-correction scheme and flux-corrected transport algorithm to propose an efficient linear positivity-preserving method and analyzed the error estimate for the solution of chemotaxis–Stokes equations.

In this work, we focus on a cancer-invasion model developed in [44], modeling the motion of cancer cells, degradation of extracellular matrix, and certain enzymes (e.g., protease). These enzymes play an important role in the degradation of the extracellular matrix and they are usually activated whenever cancer cells come in contact with the extracellular matrix. In [45], we extended the proposed model by a diffusion term, gave a rigorous proof for the existence of the global classical solution and presented numerical results for a Galerkin finite element discretization. In the present paper, a diffusion term is not considered, which makes the problem more challenging. In [46], one of the authors of the present paper applied a positivity preserving non-standard finite difference method to solve the nonlinear system in 1D, see also [47] for related approaches. Here, we consider the finite element method and apply the FCT technique to guarantee positivity preservation. First, however, we consider the more diffusive nonlinear low-order method. An additional nonlinearity is then introduced by the flux correction. We prove that both nonlinear problems are solvable and positivity preserving. To the best of our knowledge, the current work is a first attempt to gain an insight into the applicability of the FCT technique to the numerical solution of a haptotaxis system without self-diffusion and to provide a rigorous analysis of the solvability and positivity preservation. Note that the existence and uniqueness for the FEM-FCT method applied to linear evolutionary convection-diffusion equations has been addressed only recently in [48,49]. We also present a fixed-point algorithm for the iterative solution of the FCT discretization and prove that it is well posed and provides a non-negative solution at each step. Consequently, the non-negativity of the approximate solution is guaranteed independently of the choice of a stopping criterion. The properties of the proposed FCT scheme are illustrated by various numerical simulations carried out using our newly designed algorithm in the deal.II library [50,51].

The outline of this paper is as follows. In Section 2, we formulate the mathematical model which is discretized by the Galerkin method in Section 3. Then, the FCT stabilization is introduced in Section 4, where also the solvability and positivity preservation is proved. The fixed-point algorithm is proposed and investigated in Section 5. In Section 6, we report several numerical simulations in two spatial dimensions carried out for various regimes. Finally, our results are summarized in Section 7.

## 2. Mathematical model

In this section, we discuss the following nondimensionalized continuous model of a malignant cancer invasion proposed by Perumpanani et al. in [44,52]. The model contains three unknown variables, namely the cancer cell density  $u = u(x, t)$ , connective tissue  $c = c(x, t)$ , and protease  $p = p(x, t)$ , and it consists of the equations

$$\frac{\partial u}{\partial t} = \mu u(1 - u) - \chi \nabla \cdot (u \nabla c) \quad \text{in } \Omega \times (0, T], \quad (2.1)$$

$$\frac{\partial c}{\partial t} = -pc \quad \text{in } \Omega \times (0, T], \quad (2.2)$$

$$\frac{\partial p}{\partial t} = \epsilon^{-1}(uc - p) \quad \text{in } \Omega \times (0, T], \quad (2.3)$$

where  $\Omega$  is a bounded polyhedral domain in  $\mathbb{R}^d$ ,  $d \in \{1, 2, 3\}$ ,  $[0, T]$  is a time interval, and  $\mu$ ,  $\chi$ ,  $\epsilon$  are positive constants. Here,  $\mu$  and  $\chi$  denote the proliferation and haptotaxis rate of cancer cells, respectively, and the parameter  $\epsilon$  is supposed to be small since the units of connective tissues and invasive cells are much larger than the protease. In the process of invasion, the connective tissue is affected by the invasive flux of  $u \nabla c$  into its compartment. Since the connective tissue does not contain any empty space large enough for passing of passive cancer cells, it degrades by protease which is produced by invasive cancer cells upon contact with connective tissue. It can be shown that if the initial conditions of the above model are non-negative, then the computed solutions stay non-negative at all times, for more details see [52,53,47] and the references therein.



The system (2.1)–(2.3) is subjected to the homogeneous Neumann boundary condition

$$u \frac{\partial c}{\partial n} = 0 \quad \text{on } \partial\Omega \times [0, T], \tag{2.4}$$

where  $n$  is the unit outward normal vector on  $\partial\Omega$ . The above equations are endowed with the initial conditions

$$u(x, 0) = u^0(x), \quad c(x, 0) = c^0(x), \quad p(x, 0) = p^0(x), \quad x \in \Omega, \tag{2.5}$$

where  $u^0, c^0, p^0 : \Omega \rightarrow [0, 1]$  are given functions.

In [45], we considered a modified version of (2.1)–(2.3) containing an extra diffusion term in (2.1). Precisely, instead of the equation (2.1), we considered

$$\frac{\partial u}{\partial t} = \mu u(1 - u) - \chi \nabla \cdot (u \nabla c) + \alpha^{-1} \Delta u \quad \text{in } \Omega \times (0, T] \tag{2.6}$$

with a positive constant  $\alpha$ . This required to replace the boundary condition (2.4) by

$$\alpha^{-1} \frac{\partial u}{\partial n} = \chi u \frac{\partial c}{\partial n} \quad \text{on } \partial\Omega \times [0, T].$$

Thus, the problem considered in this paper corresponds to the limit case  $\alpha \rightarrow \infty$  of the problem from [45]. In that paper, we proved the existence of global classical solutions for two- and three-dimensional bounded domains  $\Omega$  with smooth boundaries and we proved that these solutions are non-negative. However, if (2.6) is replaced by (2.1), i.e., if no diffusion term is present, the technique used in [45] cannot be applied and the solvability of the model is an open problem. Our numerical results in [45] demonstrate that by fixing the proliferation rate  $\mu$  and varying the haptotaxis  $\chi$  one can make either the diffusion or the transport of the cancer cells dominant. The domination of the convection term can produce spurious oscillations and a blow-up in the solution of the system as it is the case to be considered in here.

### 3. A Galerkin discretization

The solution of the problem (2.1)–(2.5) satisfies

$$\left( \frac{\partial u}{\partial t}, v \right) = \mu (u(1 - u), v) + \chi (u \nabla c, \nabla v) \quad \text{in } (0, T] \text{ and for } v \in H^1(\Omega), \tag{3.1}$$

$$c(x, t) = c^0(x) e^{-\int_0^t p(x, s) ds} \quad \forall (x, t) \in \Omega \times [0, T], \tag{3.2}$$

$$p(x, t) = e^{-t/\epsilon} \left[ p^0(x) + \frac{1}{\epsilon} \int_0^t u(x, s) c(x, s) e^{s/\epsilon} ds \right] \quad \forall (x, t) \in \Omega \times [0, T], \tag{3.3}$$

where  $(\cdot, \cdot)$  denotes the inner product in  $L^2(\Omega)$  or  $L^2(\Omega)^d$ . To define an approximate solution of (2.1)–(2.5), we first introduce a triangulation  $\mathcal{T}_h$  of  $\Omega$  consisting of simplicial (for  $d = 1, 2, 3$ ), quadrilateral (for  $d = 2$ ) or hexahedral (for  $d = 3$ ) shape-regular cells possessing the usual compatibility properties (see, e.g., [54]). For any cell  $K \in \mathcal{T}_h$ , we denote by  $h_K$  the diameter of  $K$  and assume that  $h_K \leq h$ . We denote by  $V_h \subset H^1(\Omega)$  the usual conforming  $P_1$  or  $Q_1$  finite element space constructed using the triangulation  $\mathcal{T}_h$ . Let  $\phi_1, \dots, \phi_M$  be the standard basis functions of  $V_h$  associated with the vertices  $x_1, \dots, x_M$  of  $\mathcal{T}_h$ . Thus, the basis functions are non-negative and satisfy  $\phi_i(x_j) = \delta_{ij}$  for  $i, j = 1, \dots, M$ , where  $\delta_{ij}$  is the Kronecker symbol. Any function  $v_h \in V_h$  can be identified with a coefficient vector  $\mathbf{v} = (v_j)_{j=1}^M$  with respect to these basis functions. Precisely, introducing the bijective operator  $\pi_h : \mathbb{R}^M \rightarrow V_h$  by

$$\pi_h \mathbf{v} = \sum_{j=1}^M v_j \phi_j,$$

one has  $v_h = \pi_h \mathbf{v}$ . The assumed shape regularity of  $\mathcal{T}_h$  implies that

$$\|\nabla \phi_i\|_{L^2(K)} \leq \kappa h_K^{d/2-1} \quad \forall K \in \mathcal{T}_h, i = 1, \dots, M, \tag{3.4}$$

where  $\kappa$  is a fixed constant independent of  $i, K$ , and  $h$ . Next, the time interval  $[0, T]$  is decomposed by  $0 = t_0 < t_1 < \dots < t_N = T$  and we set  $\tau_n = t_n - t_{n-1}$ ,  $n = 1, \dots, N$ . At each time level  $t_n$ , the solution of (2.1)–(2.5) will be approximated by functions  $u_h^n, c_h^n, p_h^n \in V_h$ . These functions can be identified with coefficient vectors  $\mathbf{u}^n = (u_j^n)_{j=1}^M$ ,  $\mathbf{c}^n = (c_j^n)_{j=1}^M$ ,  $\mathbf{p}^n = (p_j^n)_{j=1}^M$ , respectively, satisfying  $u_h^n = \pi_h \mathbf{u}^n$ ,  $c_h^n = \pi_h \mathbf{c}^n$ ,  $p_h^n = \pi_h \mathbf{p}^n$ . Note that  $u_h^n(x_i) = u_i^n$ ,  $c_h^n(x_i) = c_i^n$ , and  $p_h^n(x_i) = p_i^n$  for  $i = 1, \dots, M$ . We set

$$u_i^0 = u^0(x_i), \quad c_i^0 = c^0(x_i), \quad p_i^0 = p^0(x_i), \quad i = 1, \dots, M. \tag{3.5}$$

Using linear interpolation with respect to time between the time levels gives functions  $u_{h,\tau}, c_{h,\tau}, p_{h,\tau}$  defined on  $\overline{\Omega} \times [0, T]$ . For example,  $u_{h,\tau}$  satisfies

$$u_{h,\tau}(x, t) = \frac{1}{\tau_{n+1}} [u_h^{n+1}(x)(t - t_n) + u_h^n(x)(t_{n+1} - t)] \quad \forall x \in \overline{\Omega}, t \in [t_n, t_{n+1}], n = 0, \dots, N - 1,$$

or, equivalently,

$$u_{h,\tau}(x_i, t) = \frac{1}{\tau_{n+1}} [u_i^{n+1}(t - t_n) + u_i^n(t_{n+1} - t)] \quad \forall i = 1, \dots, M, \quad t \in [t_n, t_{n+1}], \quad n = 0, \dots, N - 1.$$

Replacing the space  $H^1(\Omega)$  in (3.1) by  $V_h$  and applying the  $\theta$ -method for discretization in time (with  $\theta \in [0, 1]$ ), one obtains

$$\begin{aligned} \left( \frac{u_h^{n+1} - u_h^n}{\tau_{n+1}}, v_h \right) &= \theta \mu (u_h^{n+1}(1 - u_h^{n+1}), v_h) + \theta \chi (u_h^{n+1} \nabla c_h^{n+1}, \nabla v_h) \\ &+ (1 - \theta) \mu (u_h^n(1 - u_h^n), v_h) + (1 - \theta) \chi (u_h^n \nabla c_h^n, \nabla v_h) \quad \forall v_h \in V_h, \quad n = 0, \dots, N - 1. \end{aligned} \tag{3.6}$$

Defining the matrices  $\mathbb{M} = (m_{ij})_{i,j=1}^M$  and  $\mathbb{A}^n = (a_{ij}^n)_{i,j=1}^M$  with

$$m_{ij} = (\phi_j, \phi_i), \quad a_{ij}^n = -\mu (\phi_j(1 - u_h^n), \phi_i) - \chi (\phi_j \nabla c_h^n, \nabla \phi_i),$$

the discrete variational problem (3.6) can be written in the matrix form

$$(\mathbb{M} + \theta \tau_{n+1} \mathbb{A}^{n+1}) \mathbf{u}^{n+1} = (\mathbb{M} - (1 - \theta) \tau_{n+1} \mathbb{A}^n) \mathbf{u}^n, \quad n = 0, \dots, N - 1. \tag{3.7}$$

The relations (3.2) and (3.3) suggest to define the coefficients of  $c_h^n$  and  $p_h^n$  by

$$c_i^n = c^0(x_i) e^{-\int_0^{t_n} p_{h,\tau}(x_i, s) ds}, \quad i = 1, \dots, M, \quad n = 0, \dots, N, \tag{3.8}$$

$$p_i^n = e^{-t_n/\epsilon} \left[ p^0(x_i) + \frac{1}{\epsilon} \int_0^{t_n} u_{h,\tau}(x_i, s) c_{h,\tau}(x_i, s) e^{s/\epsilon} ds \right], \quad i = 1, \dots, M, \quad n = 0, \dots, N. \tag{3.9}$$

Then, for  $i = 1, \dots, M$  and  $n = 0, \dots, N - 1$ , one has

$$c_i^{n+1} = c_i^n e^{-\int_{t_n}^{t_{n+1}} p_{h,\tau}(x_i, s) ds}, \tag{3.10}$$

$$p_i^{n+1} = e^{-\tau_{n+1}/\epsilon} p_i^n + \frac{1}{\epsilon} e^{-t_{n+1}/\epsilon} \int_{t_n}^{t_{n+1}} u_{h,\tau}(x_i, s) c_{h,\tau}(x_i, s) e^{s/\epsilon} ds. \tag{3.11}$$

A direct computation gives

$$c_i^{n+1} = c_i^n e^{-\tau_{n+1} (p_i^{n+1} + p_i^n)/2}, \tag{3.12}$$

$$\begin{aligned} p_i^{n+1} &= e^{-\tau_{n+1}/\epsilon} p_i^n + \frac{1}{\tau_{n+1}^2} \left\{ (u_i^{n+1} (\epsilon - \tau_{n+1}) - u_i^n \epsilon) (c_i^{n+1} (\epsilon - \tau_{n+1}) - c_i^n \epsilon) \right. \\ &\quad - (u_i^{n+1} \epsilon - u_i^n (\epsilon + \tau_{n+1})) (c_i^{n+1} \epsilon - c_i^n (\epsilon + \tau_{n+1})) e^{-\tau_{n+1}/\epsilon} \\ &\quad \left. + (u_i^{n+1} - u_i^n) (c_i^{n+1} - c_i^n) \epsilon^2 (1 - e^{-\tau_{n+1}/\epsilon}) \right\}, \end{aligned} \tag{3.13}$$

for  $i = 1, \dots, M$  and  $n = 0, \dots, N - 1$ . Note that the effects described by the model (2.1)–(2.5), such as haptotaxis, strongly rely on the nonlinear coupling terms. Therefore, all nonlinearities are treated implicitly in the discrete problem (3.7)–(3.9).

To compute a solution of the nonlinear problem (3.7)–(3.9) at time  $t_{n+1}$  (assuming that the solution vectors  $\mathbf{u}^n$ ,  $\mathbf{c}^n$ , and  $\mathbf{p}^n$  at the previous time instant  $t_n$  are known), we apply simple fixed-point iterations leading to sequences  $\mathbf{u}_k^{n+1} = (u_{j,k}^{n+1})_{j=1}^M$ ,  $\mathbf{c}_k^{n+1} = (c_{j,k}^{n+1})_{j=1}^M$ , and  $\mathbf{p}_k^{n+1} = (p_{j,k}^{n+1})_{j=1}^M$ . We set  $\mathbf{u}_0^{n+1} = \mathbf{u}^n$ ,  $\mathbf{c}_0^{n+1} = \mathbf{c}^n$ ,  $\mathbf{p}_0^{n+1} = \mathbf{p}^n$  and then, for  $k > 0$  and  $i = 1, \dots, M$ , we define

$$c_{i,k}^{n+1} = c_i^n e^{-\tau_{n+1} (p_{i,k-1}^{n+1} + p_i^n)/2}, \tag{3.14}$$

$$\begin{aligned} p_{i,k}^{n+1} &= e^{-\tau_{n+1}/\epsilon} p_i^n + \frac{1}{\tau_{n+1}^2} \left\{ (u_{i,k-1}^{n+1} (\epsilon - \tau_{n+1}) - u_i^n \epsilon) (c_{i,k}^{n+1} (\epsilon - \tau_{n+1}) - c_i^n \epsilon) \right. \\ &\quad - (u_{i,k-1}^{n+1} \epsilon - u_i^n (\epsilon + \tau_{n+1})) (c_{i,k}^{n+1} \epsilon - c_i^n (\epsilon + \tau_{n+1})) e^{-\tau_{n+1}/\epsilon} \\ &\quad \left. + (u_{i,k-1}^{n+1} - u_i^n) (c_{i,k}^{n+1} - c_i^n) \epsilon^2 (1 - e^{-\tau_{n+1}/\epsilon}) \right\}. \end{aligned} \tag{3.15}$$

The iterate  $\mathbf{u}_k^{n+1}$  is computed by solving the linear system

$$(\mathbb{M} + \theta \tau_{n+1} \mathbb{A}_{k-1}^{n+1}) \mathbf{u}_k^{n+1} = (\mathbb{M} - (1 - \theta) \tau_{n+1} \mathbb{A}^n) \mathbf{u}^n, \tag{3.16}$$

where the matrix  $\mathbb{A}_{k-1}^{n+1}$  is defined by

$$\mathbb{A}_{k-1}^{n+1} = \left( -\mu (\phi_j (1 - u_{h,k-1}^{n+1}), \phi_i) - \chi (\phi_j \nabla c_{h,k}^{n+1}, \nabla \phi_i) \right)_{i,j=1}^M \quad (3.17)$$

and  $u_{h,k-1}^{n+1} = \pi_h \mathbf{u}_{k-1}^{n+1}$  and  $c_{h,k}^{n+1} = \pi_h \mathbf{c}_k^{n+1}$  are the finite element functions corresponding to the coefficient vectors  $\mathbf{u}_{k-1}^{n+1}$  and  $\mathbf{c}_k^{n+1}$ , respectively.

The linear system (3.16) has the form

$$\mathbb{B} \mathbf{u}^{n+1} = \mathbb{K} \mathbf{u}^n \quad (3.18)$$

and it is desirable that this system is positivity preserving, i.e., that  $\mathbf{u}^{n+1} \geq 0$  if  $\mathbf{u}^n \geq 0$ . A necessary and sufficient condition for this property is  $\mathbb{B}^{-1} \mathbb{K} \geq 0$  but this condition is difficult to verify. Sufficient conditions are formulated in the following lemma. Note that, throughout the paper, an inequality of the type  $\mathbf{u}^n \geq 0$  means that the inequality holds for each component of the vector  $\mathbf{u}^n$ . Similarly, the fact that all entries of a matrix  $\mathbb{K}$  are non-negative is expressed by  $\mathbb{K} \geq 0$ .

**Lemma 3.1.** *Let the matrices  $\mathbb{B} = (b_{ij})_{i,j=1}^M$  and  $\mathbb{K} = (k_{ij})_{i,j=1}^M$  satisfy*

$$b_{ii} \geq 0, \quad k_{ii} \geq 0, \quad b_{ij} \leq 0, \quad k_{ij} \geq 0, \quad \forall i, j = 1, \dots, M, \quad i \neq j,$$

and let  $\mathbb{B}$  be a strictly diagonally dominant or an irreducibly diagonally dominant matrix. Then  $\mathbb{B}$  is an M-matrix and the scheme (3.18) is positivity preserving.

**Proof.** According to [55, Theorem 3.27],  $\mathbb{B}$  is an M-matrix. Thus,  $\mathbb{B}^{-1} \geq 0$  and hence also  $\mathbb{B}^{-1} \mathbb{K} \geq 0$ , which implies the result.  $\square$

In general, the linear system (3.16) originating from a standard Galerkin discretization does not satisfy the above constraints because the mass matrix is non-negative and the stiffness matrix may contain positive off-diagonal entries. Our numerical results in Section 6 show that indeed the concentration  $\mathbf{u}$  may become negative in some parts of the computational domain  $\Omega$ .

#### 4. FCT stabilization

As we will see in Section 6, the magnitude of the solutions gradients can be extremely large in some regions. The solution of the Galerkin discretization from the previous section may become negative especially in these regions. As a remedy, in the following we will modify the Galerkin discretization to guarantee a positivity preservation property. As shown by Kuzmin [32–34], this property can be readily enforced at the discrete level using a conservative manipulation of the mass and stiffness matrices. The former will be approximated by its diagonal counterpart  $\mathbb{M}_L$  constructed using row-sum mass lumping, whereas the latter will be modified by adding an artificial diffusion matrix. To limit the amount of the artificial diffusion, the FEM-FCT approach will be applied following [33].

Since the methods considered in this section guarantee that the approximate solutions are non-negative, it is possible to replace the matrix  $\mathbb{A}^n$  from the previous section by  $\tilde{\mathbb{A}}^n = (\tilde{a}_{ij}^n)_{i,j=1}^M$  with

$$\tilde{a}_{ij}^n = -\mu (\phi_j (1 - |u_h^n|), \phi_i) - \chi (\phi_j \nabla c_h^n, \nabla \phi_i).$$

The matrix  $\tilde{\mathbb{A}}^n$  is more suitable for theoretical considerations than the matrix  $\mathbb{A}^n$ . However, a non-negative approximate solution  $u_h^n$ ,  $c_h^n$ ,  $p_h^n$  satisfying a discrete problem based on the matrix  $\tilde{\mathbb{A}}^n$  will satisfy also the corresponding discrete problem with the original matrix  $\mathbb{A}^n$ .

Using the matrix  $\tilde{\mathbb{A}}^n$ , we introduce a symmetric artificial diffusion matrix  $\mathbb{D}^n = (d_{ij}^n)_{i,j=1}^M$  defined by

$$d_{ij}^n = -\max\{\tilde{a}_{ij}^n, 0, \tilde{a}_{ji}^n\} \quad \text{for } i \neq j, \quad d_{ii}^n = -\sum_{j=1, j \neq i}^M d_{ij}^n,$$

and we set  $\mathbb{L}^n = \tilde{\mathbb{A}}^n + \mathbb{D}^n$ . Note that  $\mathbb{L}^n = (l_{ij}^n)_{i,j=1}^M$  is a Z-matrix (i.e., it has non-positive off-diagonal entries). Furthermore, we introduce the lumped mass matrix  $\mathbb{M}_L = \text{diag}(m_1, \dots, m_M)$  with

$$m_i = \sum_{j=1}^M m_{ij}, \quad i = 1, \dots, M.$$

Now, the simplest way to enforce the positivity preservation is to consider the so-called low-order method corresponding to the so-called high-order method (3.7) which is defined by

$$(\mathbb{M}_L + \theta \tau_{n+1} \mathbb{L}^{n+1}) \mathbf{u}^{n+1} = (\mathbb{M}_L - (1 - \theta) \tau_{n+1} \mathbb{L}^n) \mathbf{u}^n, \quad n = 0, \dots, N - 1. \quad (4.1)$$

Note that the matrix  $\mathbb{L}^{n+1}$  depends on  $\mathbf{u}^{n+1}$  and  $\mathbf{c}^{n+1}$  so that the low-order problem is again nonlinear. In contrast to the Galerkin discretization (3.7), it is now possible to assure the positivity preservation for sufficiently small time steps.

**Lemma 4.1.** Let the time step  $\tau_{n+1}$  satisfy the conditions

$$(1 - \theta) \tau_{n+1} l_{ii}^n \leq m_i, \quad \theta \tau_{n+1} \left( \mu m_i + \chi \left( \nabla c_h^{n+1}, \nabla \phi_i \right) \right) < m_i, \quad i = 1, \dots, M. \quad (4.2)$$

Then the matrix  $\mathbb{M}_L - (1 - \theta) \tau_{n+1} \mathbb{L}^n$  has non-negative entries and  $\mathbb{M}_L + \theta \tau_{n+1} \mathbb{L}^{n+1}$  is an M-matrix.

**Proof.** The first condition in (4.2) implies that  $\mathbb{M}_L - (1 - \theta) \tau_{n+1} \mathbb{L}^n$  has non-negative diagonal entries. The off-diagonal entries of this matrix are non-negative as well, since  $\mathbb{M}_L$  is diagonal and  $\mathbb{L}^n$  is a Z-matrix.

Denoting  $\mathbb{B} = \mathbb{M}_L + \theta \tau_{n+1} \mathbb{L}^{n+1}$ , one has for any  $i \in \{1, \dots, M\}$

$$\sum_{j=1}^M b_{ij} = m_i + \theta \tau_{n+1} \sum_{j=1}^M \tilde{a}_{ij}^{n+1} = m_i - \theta \tau_{n+1} \left( \mu (1 - |u_h^{n+1}|, \phi_i) + \chi \left( \nabla c_h^{n+1}, \nabla \phi_i \right) \right),$$

where we used the fact that  $\sum_{j=1}^M \phi_j = 1$ . Since  $(1, \phi_i) = m_i$ , it follows from the second condition in (4.2) that  $\sum_{j=1}^M b_{ij} > 0$ . Thus,  $b_{ii} > \sum_{j \neq i} |b_{ij}|$ , i.e.,  $\mathbb{B}$  is strictly diagonally dominant and hence non-singular. Moreover,  $\mathbb{B}$  is a matrix of non-negative type and hence it is an M-matrix (see, e.g., [56, Corollary 3.13]).  $\square$

**Corollary 4.2.** Let the time step  $\tau_{n+1}$  satisfy the conditions (4.2). Then the low-order scheme (4.1) is positivity preserving, i.e.,

$$\mathbf{u}^n \geq 0 \quad \Rightarrow \quad \mathbf{u}^{n+1} \geq 0. \quad (4.3)$$

**Proof.** According to Lemma 4.1, the matrix  $\mathbb{M}_L + \theta \tau_{n+1} \mathbb{L}^{n+1}$  is non-singular,  $(\mathbb{M}_L + \theta \tau_{n+1} \mathbb{L}^{n+1})^{-1} \geq 0$ , and  $\mathbb{M}_L - (1 - \theta) \tau_{n+1} \mathbb{L}^n \geq 0$ , which immediately implies (4.3).  $\square$

**Remark 4.3.** The second condition in (4.2) involves  $c_h^{n+1}$  which implicitly depends on  $\tau_{n+1}$  through  $\mathbf{p}^{n+1}$  and hence also through  $\mathbf{u}^{n+1}$ . Therefore, it is desirable to replace this condition by a condition independent of  $c_h^{n+1}$ . This is possible since we will show that the values of  $c_h^{n+1}$  are in the interval  $[0, 1]$ . Then, employing (3.4), one gets

$$\left( \nabla c_h^{n+1}, \nabla \phi_i \right) = \sum_{j=1}^M c_j^{n+1} \left( \nabla \phi_j, \nabla \phi_i \right) \leq \sum_{K \ni x_i} \sum_{j=1}^M \|\nabla \phi_j\|_{L^2(K)} \|\nabla \phi_i\|_{L^2(K)} \leq n_v \kappa^2 \sum_{K \ni x_i} h_K^{d-2},$$

where  $n_v$  is the number of vertices of a cell in  $\mathcal{T}_h$  ( $n_v = d + 1$  for simplices,  $n_v = 4$  for quadrilaterals, and  $n_v = 8$  for hexahedra). Thus, if the time step  $\tau_{n+1}$  satisfies

$$\theta \tau_{n+1} \left( \mu m_i + \chi n_v \kappa^2 \sum_{K \ni x_i} h_K^{d-2} \right) < m_i, \quad i = 1, \dots, M, \quad (4.4)$$

and  $c_h^{n+1} \in [0, 1]$ , then the second condition in (4.2) holds. Note that (4.4) may be significantly more restrictive than (4.2).

To prove that the low-order discretization consisting of the equations (4.1), (3.12), and (3.13) has a solution, we shall use the following consequence of Brouwer's fixed-point theorem.

**Lemma 4.4.** Let  $X$  be a finite-dimensional Hilbert space with inner product  $(\cdot, \cdot)_X$  and norm  $\|\cdot\|_X$ . Let  $P : X \rightarrow X$  be a continuous mapping and  $K > 0$  a real number such that  $(Px, x)_X > 0$  for any  $x \in X$  with  $\|x\|_X = K$ . Then there exists  $x \in X$  such that  $\|x\|_X < K$  and  $Px = 0$ .

**Proof.** See [57, p. 164, Lemma 1.4].  $\square$

**Theorem 4.5.** Consider any  $n \in \{0, \dots, N - 1\}$  and let  $\mathbf{u}^n, \mathbf{c}^n, \mathbf{p}^n \in \mathbb{R}^M$  satisfy  $\mathbf{u}^n \geq 0$ ,  $1 \geq \mathbf{c}^n \geq 0$ ,  $\mathbf{p}^n \geq 0$ . Let the time step  $\tau_{n+1}$  satisfy the conditions

$$(1 - \theta) \tau_{n+1} l_{ii}^n \leq m_i, \quad \theta \tau_{n+1} \left( \mu m_i + \chi n_v \kappa^2 \sum_{K \ni x_i} h_K^{d-2} \right) < m_i, \quad i = 1, \dots, M. \quad (4.5)$$

Then there exist vectors  $\mathbf{u}^{n+1}, \mathbf{c}^{n+1}, \mathbf{p}^{n+1} \in \mathbb{R}^M$  satisfying (4.1), (3.12), (3.13) and  $\mathbf{u}^{n+1} \geq 0$ ,  $1 \geq \mathbf{c}^{n+1} \geq 0$ ,  $\mathbf{p}^{n+1} \geq 0$ .

**Proof.** To get rid of the exponential dependence on  $\mathbf{p}^{n+1}$  when estimating the nonlinear terms in (4.1), we replace (3.12) by

$$c_i^{n+1} = c_i^n e^{-\tau_{n+1} (|p_i^{n+1}| + p_i^n)/2}, \quad i = 1, \dots, M. \quad (4.6)$$

At the end of the proof, we will show that  $\mathbf{p}^{n+1} \geq 0$  so that the original relation (3.12) will be recovered.

For  $u, p \in \mathbb{R}$  and  $i = 1, \dots, M$ , we introduce the notation

$$C_i(p) = c_i^n e^{-\tau_{n+1}(|p|+p_i^n)/2} \tag{4.7}$$

and

$$\begin{aligned} P_i(u, p) = & \frac{1}{\tau_{n+1}^2} u C_i(p) \left\{ (\epsilon - \tau_{n+1})^2 + \epsilon^2 (1 - 2e^{-\tau_{n+1}/\epsilon}) \right\} \\ & + \frac{\epsilon}{\tau_{n+1}^2} u c_i^n \left\{ \tau_{n+1} (1 + e^{-\tau_{n+1}/\epsilon}) - 2\epsilon (1 - e^{-\tau_{n+1}/\epsilon}) \right\} \\ & + \frac{\epsilon}{\tau_{n+1}^2} u_i^n C_i(p) \left\{ \tau_{n+1} (1 + e^{-\tau_{n+1}/\epsilon}) - 2\epsilon (1 - e^{-\tau_{n+1}/\epsilon}) \right\} \\ & + \frac{1}{\tau_{n+1}^2} u_i^n c_i^n \left\{ \epsilon^2 (2 - e^{-\tau_{n+1}/\epsilon}) - (\epsilon + \tau_{n+1})^2 e^{-\tau_{n+1}/\epsilon} \right\} + e^{-\tau_{n+1}/\epsilon} p_i^n. \end{aligned} \tag{4.8}$$

Then, the validity of (4.6) and (3.13) is equivalent to

$$c_i^{n+1} = C_i(p_i^{n+1}), \quad p_i^{n+1} = P_i(u_i^{n+1}, p_i^{n+1}), \quad i = 1, \dots, M.$$

Note that

$$|P_i(u, p)| \leq (|u| + u_i^n) c_i^n \left( \frac{2\epsilon + \tau_{n+1}}{\tau_{n+1}} \right)^2 + p_i^n \quad \forall u, p \in \mathbb{R}, \quad i = 1, \dots, M. \tag{4.9}$$

Furthermore, for  $\mathbf{u}, \mathbf{p} \in \mathbb{R}^M$  and  $i, j = 1, \dots, M$ , we denote

$$A_{ij}(\mathbf{u}, \mathbf{p}) = -\mu (\phi_j(1 - |\pi_h \mathbf{u}|), \phi_i) - \chi (\phi_j \nabla(\pi_h \mathbf{C}(\mathbf{p})), \nabla \phi_i), \tag{4.10}$$

$$D_{ij}(\mathbf{u}, \mathbf{p}) = -\max\{A_{ij}(\mathbf{u}, \mathbf{p}), 0, A_{ji}(\mathbf{u}, \mathbf{p})\} \quad \text{for } i \neq j, \quad D_{ii}(\mathbf{u}, \mathbf{p}) = -\sum_{j=1, j \neq i}^M D_{ij}(\mathbf{u}, \mathbf{p}), \tag{4.11}$$

$$S_i(\mathbf{u}, \mathbf{p}) = m_i u_i + \theta \tau_{n+1} \sum_{j=1}^M (A_{ij}(\mathbf{u}, \mathbf{p}) + D_{ij}(\mathbf{u}, \mathbf{p})) u_j - [(\mathbb{M}_L - (1 - \theta) \tau_{n+1} \mathbb{L}^n) \mathbf{u}^n]_i, \tag{4.12}$$

where  $\mathbf{C}(\mathbf{p}) = (C_i(p_i))_{i=1}^M$ . Then (4.1) with  $c_h^{n+1}$  defined by (4.6) is equivalent to

$$S_i(\mathbf{u}^{n+1}, \mathbf{p}^{n+1}) = 0, \quad i = 1, \dots, M.$$

Therefore, defining the operator  $P : \mathbb{R}^{2M} \rightarrow \mathbb{R}^{2M}$  by

$$P \mathbf{U} = (S_1(\mathbf{u}, \mathbf{p}), \dots, S_M(\mathbf{u}, \mathbf{p}), p_1 - P_1(u_1, p_1), \dots, p_M - P_M(u_M, p_M)) \quad \forall \mathbf{U} = (\mathbf{u}, \mathbf{p}) \in \mathbb{R}^{2M}, \tag{4.13}$$

the vectors  $\mathbf{u}^{n+1}, \mathbf{c}^{n+1}, \mathbf{p}^{n+1}$  are a solution of (4.1), (4.6), (3.13) if and only if  $\mathbf{U} = (\mathbf{u}^{n+1}, \mathbf{p}^{n+1})$  satisfies  $P \mathbf{U} = 0$  and  $\mathbf{c}^{n+1} = \mathbf{C}(\mathbf{p}^{n+1})$ .

To show that the equation  $P \mathbf{U} = 0$  has a solution, we will verify the assumptions of Lemma 4.4. Since it is obvious that the operator  $P$  is continuous, it suffices to investigate the product  $(P \mathbf{U}, \mathbf{U})$ , where  $(\cdot, \cdot)$  is the Euclidean inner product in  $\mathbb{R}^{2M}$ . We will denote the corresponding norm by  $\|\cdot\|$ . The Euclidean norm in  $\mathbb{R}^M$  will be denoted by  $\|\cdot\|_M$ . Since the matrix  $(D_{ij}(\mathbf{u}, \mathbf{p}))_{i,j=1}^M$  is symmetric and has zero row sums and non-positive off-diagonal entries, one obtains

$$\sum_{i,j=1}^M u_i D_{ij}(\mathbf{u}, \mathbf{p}) u_j = -\frac{1}{2} \sum_{i,j=1}^M D_{ij}(\mathbf{u}, \mathbf{p}) (u_i - u_j)^2 \geq 0 \quad \forall \mathbf{u}, \mathbf{p} \in \mathbb{R}^M.$$

Furthermore, since  $0 \leq \mathbf{C}(\mathbf{p}) \leq 1$ , the expressions  $(\phi_j \nabla(\pi_h \mathbf{C}(\mathbf{p})), \nabla \phi_i)$  can be bounded independently of  $\mathbf{p}$ . Therefore, using the equivalence of norms on finite-dimensional spaces, one obtains

$$\theta \tau_{n+1} \sum_{i,j=1}^M u_i (A_{ij}(\mathbf{u}, \mathbf{p}) + D_{ij}(\mathbf{u}, \mathbf{p})) u_j \geq \theta \tau_{n+1} \mu \|\pi_h \mathbf{u}\|_{L^3(\Omega)}^3 - C_1 \|\mathbf{u}\|_M^2 \geq C_2 \|\mathbf{u}\|_M^3 - C_1 \|\mathbf{u}\|_M^2,$$

where  $C_1$  and  $C_2$  are positive constants independent of  $\mathbf{u}$  and  $\mathbf{p}$ . Thus,

$$\sum_{i=1}^M u_i S_i(\mathbf{u}, \mathbf{p}) \geq C_2 \|\mathbf{u}\|_M^3 - C_1 \|\mathbf{u}\|_M^2 - C_3 \|\mathbf{u}\|_M, \tag{4.14}$$

where  $C_3 = \|(\mathbb{M}_L - (1 - \theta) \tau_{n+1} \mathbb{L}^n) \mathbf{u}^n\|_M$ . Finally, using (4.9), it follows that

$$\sum_{i=1}^M p_i (p_i - P_i(u_i, p_i)) \geq \|\mathbf{p}\|_M^2 - C_4 \|\mathbf{p}\|_M \|\mathbf{u}\|_M - C_5 \|\mathbf{p}\|_M,$$

with positive constants  $C_4$  and  $C_5$  independent of  $\mathbf{u}$  and  $\mathbf{p}$ . Applying the Young inequality, the previous two inequalities imply that there exist positive constants  $C_6$  and  $C_7$  such that

$$(P\mathbf{U}, \mathbf{U}) \geq \frac{1}{2} \|\mathbf{U}\|^2 + C_2 \|\mathbf{u}\|_M^3 - C_6 \|\mathbf{u}\|_M^2 - C_7 \geq \frac{1}{2} \|\mathbf{U}\|^2 - \frac{C_6^3}{C_2^2} - C_7 \quad \forall \mathbf{U} = (\mathbf{u}, \mathbf{p}) \in \mathbb{R}^{2M}.$$

Thus, for any  $K > \sqrt{2 C_6^3 / C_2^2 + 2 C_7}$ , one has  $(P\mathbf{U}, \mathbf{U}) > 0$  for any  $\mathbf{U} \in \mathbb{R}^{2M}$  with  $\|\mathbf{U}\| = K$ . Therefore, according to Lemma 4.4, there exists a solution  $\mathbf{U}$  of the equation  $P\mathbf{U} = 0$  and hence also a solution  $\mathbf{u}^{n+1}$ ,  $\mathbf{c}^{n+1}$ ,  $\mathbf{p}^{n+1}$  of (4.1), (4.6), and (3.13).

It immediately follows from (4.6) that  $0 \leq \mathbf{c}^{n+1} \leq 1$ . Thus, according to Corollary 4.2 and Remark 4.3, the solution satisfies  $\mathbf{u}^{n+1} \geq 0$ . Since (3.13) is equivalent to (3.11), one also has  $\mathbf{p}^{n+1} \geq 0$  and hence (3.12) is satisfied as well.  $\square$

Although the solution of (4.1), (3.12), (3.13) does not possess negative values under the time step restrictions (4.2), it is usually very inaccurate since too much artificial diffusion is introduced by the modifications leading to the low-order method (4.1), cf. Section 6.3. Therefore, in the FEM-FCT methodology, a correction term  $\bar{\mathbf{f}}^{n+1}$  is added in such a way that the method becomes less diffusive while negative values are still excluded. This leads to an extension of (4.1) in the form

$$(\mathbb{M}_L + \theta \tau_{n+1} \mathbb{L}^{n+1}) \mathbf{u}^{n+1} = (\mathbb{M}_L - (1 - \theta) \tau_{n+1} \mathbb{L}^n) \mathbf{u}^n + \bar{\mathbf{f}}^{n+1}.$$

The high-order method (3.7) (with  $\mathbb{A}^n$  replaced by  $\bar{\mathbb{A}}^n$ ) is recovered if

$$\bar{\mathbf{f}}^{n+1} = (\mathbb{M}_L - \mathbb{M})(\mathbf{u}^{n+1} - \mathbf{u}^n) + \theta \tau_{n+1} \mathbb{D}^{n+1} \mathbf{u}^{n+1} + (1 - \theta) \tau_{n+1} \mathbb{D}^n \mathbf{u}^n. \tag{4.15}$$

Since  $\mathbb{D}^n$  has zero row sums, one can write

$$(\mathbb{D}^n \mathbf{u}^n)_i = \sum_{j=1}^M d_{ij}^n (u_j^n - u_i^n), \quad i = 1, \dots, M.$$

For the terms with the matrices  $\mathbb{D}^{n+1}$  and  $\mathbb{M}_L - \mathbb{M}$  (which also have zero row sums), one can proceed analogously and hence (4.15) holds if an only if

$$\bar{\mathbf{f}}^{n+1} = \left( \sum_{j=1}^M f_{ij}^{n+1} \right)_{i=1}^M,$$

where the algebraic fluxes  $f_{ij}^{n+1}$  are given by

$$f_{ij}^{n+1} = -m_{ij} (u_j^{n+1} - u_i^{n+1}) + m_{ij} (u_j^n - u_i^n) + \theta \tau_{n+1} d_{ij}^{n+1} (u_j^{n+1} - u_i^{n+1}) + (1 - \theta) \tau_{n+1} d_{ij}^n (u_j^n - u_i^n). \tag{4.16}$$

Because  $\mathbb{M}$ ,  $\mathbb{D}^{n+1}$ , and  $\mathbb{D}^n$  are symmetric matrices, one has  $f_{ij}^{n+1} = -f_{ji}^{n+1}$ . Note also that the fluxes depend on (unknown) values of the approximate solution at time level  $t_{n+1}$ .

Now, the idea of the FCT approach is to limit the fluxes  $f_{ij}^{n+1}$  by solution dependent correction factors  $\alpha_{ij}^{n+1} \in [0, 1]$  called limiters so that the non-negativity of the approximate solution can be guaranteed but less artificial diffusion is introduced than in case of the low-order method. This leads to the discrete problem

$$(\mathbb{M}_L + \theta \tau_{n+1} \mathbb{L}^{n+1}) \mathbf{u}^{n+1} = (\mathbb{M}_L - (1 - \theta) \tau_{n+1} \mathbb{L}^n) \mathbf{u}^n + \left( \sum_{j=1}^M \alpha_{ij}^{n+1} f_{ij}^{n+1} \right)_{i=1}^M. \tag{4.17}$$

The original Galerkin discretization is recovered for  $\alpha_{ij} = 1$  while the largest amount of artificial diffusion is introduced for  $\alpha_{ij} = 0$ . The latter setting is appropriate in the neighborhood of steep fronts and large gradients. The artificial diffusion can be removed in regions where the solution is smooth and where non-positive off-diagonal entries of the stiffness matrix do not pose any threat to non-negativity. The corrected fluxes depend on the approximate solution in a nonlinear way but since the problem in here is already nonlinear, we can treat both nonlinearities simultaneously.

It is convenient to write the nonlinear problem (4.17) in the form

$$\mathbb{M}_L \bar{\mathbf{u}} = (\mathbb{M}_L - (1 - \theta) \tau_{n+1} \mathbb{L}^n) \mathbf{u}^n, \tag{4.18}$$

$$\mathbb{M}_L \bar{\mathbf{u}} = \mathbb{M}_L \bar{\mathbf{u}} + \left( \sum_{j=1}^M \alpha_{ij}^{n+1} f_{ij}^{n+1} \right)_{i=1}^M, \tag{4.19}$$

$$(\mathbb{M}_L + \theta \tau_{n+1} \mathbb{L}^{n+1}) \mathbf{u}^{n+1} = \mathbb{M}_L \bar{\mathbf{u}}. \tag{4.20}$$

According to Lemma 4.1, the steps (4.18) and (4.20) are positivity preserving under the conditions (4.2). To guarantee the positivity preservation of the second step, the limiters  $\alpha_{ij}^{n+1}$  have to be defined appropriately. We will apply the Zalesak algorithm [58] which will be described next.

The solution of the nonlinear problem (4.18)–(4.20) is computed by fixed-point iterations where the algebraic fluxes are calculated using the previous iterate. Since the properties of the Zalesak algorithm do not depend on the form of these fluxes, we will denote them simply by  $f_{ij}$ . Then, the aim is to find limiters  $\alpha_{ij} \in [0, 1]$  such that the solution  $\tilde{\mathbf{u}}$  of

$$\mathbb{M}_L \tilde{\mathbf{u}} = \mathbb{M}_L \bar{\mathbf{u}} + \left( \sum_{j=1}^M \alpha_{ij} f_{ij} \right)_{i=1}^M$$

satisfies

$$\bar{u}_i^{\min} \leq \tilde{u}_i \leq \bar{u}_i^{\max}, \quad i = 1, \dots, M, \quad (4.21)$$

where

$$\bar{u}_i^{\min} = \min_{j \in \mathcal{N}_i \cup \{i\}} \bar{u}_j, \quad \bar{u}_i^{\max} = \max_{j \in \mathcal{N}_i \cup \{i\}} \bar{u}_j, \quad i = 1, \dots, M,$$

and  $\mathcal{N}_i$  is the index set of neighbor vertices to the vertex  $x_i$  (note that two vertices of the triangulation  $\mathcal{T}_h$  are called neighboring if they are contained in the same mesh cell). To preserve conservativity, it is important that the limiters  $\alpha_{ij}$  form a symmetric matrix. The limiting process begins with canceling all fluxes that are diffusive in nature and tend to flatten the solution profiles, cf. [33]. The required modification is

$$f_{ij} := 0 \quad \text{if } f_{ij}(\bar{u}_j - \bar{u}_i) > 0. \quad (4.22)$$

The remaining fluxes are truly antidiffusive and the computation of  $\alpha_{ij}$  involves the following steps:

1. Compute the sum of positive/negative antidiffusive fluxes into node  $i$

$$P_i^+ = \sum_{j \in \mathcal{N}_i} \max\{0, f_{ij}\}, \quad P_i^- = \sum_{j \in \mathcal{N}_i} \min\{0, f_{ij}\}. \quad (4.23)$$

2. Compute the distance to a local extremum of the auxiliary solution  $\bar{\mathbf{u}}$

$$Q_i^+ = m_i(\bar{u}_i^{\max} - \bar{u}_i), \quad Q_i^- = m_i(\bar{u}_i^{\min} - \bar{u}_i). \quad (4.24)$$

3. Compute the nodal correction factors for the net increment to node  $i$

$$R_i^+ = \min \left\{ 1, \frac{Q_i^+}{P_i^+} \right\}, \quad R_i^- = \min \left\{ 1, \frac{Q_i^-}{P_i^-} \right\}. \quad (4.25)$$

If a denominator is zero, set the respective value of  $R_i^+$  or  $R_i^-$  equal to 1.

4. Check the sign of the antidiffusive flux and define the correction factor by

$$\alpha_{ij} = \begin{cases} \min\{R_i^+, R_j^-\} & \text{if } f_{ij} > 0, \\ 1 & \text{if } f_{ij} = 0, \\ \min\{R_i^-, R_j^+\} & \text{if } f_{ij} < 0. \end{cases} \quad (4.26)$$

It can be easily verified (see, e.g., [56]) that this algorithm leads to the property (4.21).

Now we are in a position to prove the solvability and positivity preservation for the above FCT discretization.

**Theorem 4.6.** Consider any  $n \in \{0, \dots, N-1\}$  and let  $\mathbf{u}^n, \mathbf{c}^n, \mathbf{p}^n \in \mathbb{R}^M$  satisfy  $\mathbf{u}^n \geq 0$ ,  $1 \geq \mathbf{c}^n \geq 0$ ,  $\mathbf{p}^n \geq 0$ . Let the time step  $\tau_{n+1}$  satisfy the conditions (4.5). Then there exist vectors  $\mathbf{u}^{n+1}, \mathbf{c}^{n+1}, \mathbf{p}^{n+1} \in \mathbb{R}^M$  satisfying (4.17), (3.12), (3.13) where the fluxes  $f_{ij}^{n+1}$  are given by (4.16) and (4.22) and the limiters  $\alpha_{ij}^{n+1}$  are computed using the Zalesak algorithm (4.23)–(4.26) from the fluxes  $f_{ij}^{n+1}$ . Moreover, these vectors satisfy  $\mathbf{u}^{n+1} \geq 0$ ,  $1 \geq \mathbf{c}^{n+1} \geq 0$ , and  $\mathbf{p}^{n+1} \geq 0$ .

**Proof.** The proof follows the lines of that of Theorem 4.5. Thus, we again start with replacing (3.12) by (4.6). We again define  $C_i$ ,  $P_i$ ,  $A_{ij}$ , and  $D_{ij}$  by (4.7), (4.8), (4.10) and (4.11), respectively, whereas  $S_i$  are now defined by

$$S_i(\mathbf{u}, \mathbf{p}) = m_i u_i + \theta \tau_{n+1} \sum_{j=1}^M (A_{ij}(\mathbf{u}, \mathbf{p}) + D_{ij}(\mathbf{u}, \mathbf{p})) u_j - \sum_{j=1}^M \alpha_{ij}(\mathbf{u}, \mathbf{p}) \tilde{f}_{ij}(\mathbf{u}, \mathbf{p}) - [(\mathbb{M}_L - (1 - \theta) \tau_{n+1} \mathbb{L}^n) \mathbf{u}^n]_i,$$

where  $\alpha_{ij}(\mathbf{u}, \mathbf{p})$  are defined by the Zalesak algorithm (4.23)–(4.26) for the algebraic fluxes  $\tilde{f}_{ij}(\mathbf{u}, \mathbf{p})$  defined by

$$\tilde{f}_{ij}(\mathbf{u}, \mathbf{p}) = \begin{cases} f_{ij}(\mathbf{u}, \mathbf{p}) & \text{if } f_{ij}(\mathbf{u}, \mathbf{p})(\bar{u}_j - \bar{u}_i) \leq 0, \\ 0 & \text{if } f_{ij}(\mathbf{u}, \mathbf{p})(\bar{u}_j - \bar{u}_i) > 0, \end{cases}$$

with  $\bar{\mathbf{u}}$  from (4.18) and

$$f_{ij}(\mathbf{u}, \mathbf{p}) = (-m_{ij} + \theta \tau_{n+1} D_{ij}(\mathbf{u}, \mathbf{p}))(u_j - u_i) + (m_{ij} + (1 - \theta) \tau_{n+1} d_{ij}^n)(u_j^n - u_i^n).$$

Then, defining the operator  $P : \mathbb{R}^{2M} \rightarrow \mathbb{R}^{2M}$  by (4.13), the vectors  $\mathbf{u}^{n+1}$ ,  $\mathbf{c}^{n+1}$ ,  $\mathbf{p}^{n+1}$  are a solution of (4.17), (4.6), (3.13) if and only if  $\mathbf{U} = (\mathbf{u}^{n+1}, \mathbf{p}^{n+1})$  satisfies  $P\mathbf{U} = 0$  and  $\mathbf{c}^{n+1} = \mathbf{C}(\mathbf{p}^{n+1})$ .

The solvability of the equation  $P\mathbf{U} = 0$  will be again proved using Lemma 4.4. To show the continuity of the operator  $P$  at any point  $\tilde{\mathbf{U}} \equiv (\tilde{\mathbf{u}}, \tilde{\mathbf{p}}) \in \mathbb{R}^{2M}$ , it suffices to consider the terms  $\alpha_{ij}(\mathbf{u}, \mathbf{p}) \tilde{f}_{ij}(\mathbf{u}, \mathbf{p})$  since the remaining terms in the definition of  $P$  are clearly continuous. Moreover,  $f_{ij}$  and hence also  $\tilde{f}_{ij}$  are continuous. Thus, if  $\tilde{f}_{ij}(\tilde{\mathbf{U}}) \neq 0$ , then the denominators in the formulas defining  $\alpha_{ij}(\mathbf{U})$  with  $\mathbf{U} = (\mathbf{u}, \mathbf{p})$  do not vanish in a neighborhood of  $\tilde{\mathbf{U}}$  and hence  $\alpha_{ij}$  is continuous at  $\tilde{\mathbf{U}}$ . Consequently, also  $\alpha_{ij} \tilde{f}_{ij}$  is continuous at  $\tilde{\mathbf{U}}$ . If  $\tilde{f}_{ij}(\tilde{\mathbf{U}}) = 0$ , then

$$|(\alpha_{ij} \tilde{f}_{ij})(\mathbf{U}) - (\alpha_{ij} \tilde{f}_{ij})(\tilde{\mathbf{U}})| = |(\alpha_{ij} \tilde{f}_{ij})(\mathbf{U})| \leq |\tilde{f}_{ij}(\mathbf{U})| = |\tilde{f}_{ij}(\mathbf{U}) - \tilde{f}_{ij}(\tilde{\mathbf{U}})|,$$

which shows that  $\alpha_{ij} \tilde{f}_{ij}$  is again continuous at  $\tilde{\mathbf{U}}$ .

To estimate  $(P\mathbf{U}, \mathbf{U})$  from below, let us denote

$$\tilde{\alpha}_{ij}(\mathbf{u}, \mathbf{p}) = \begin{cases} \alpha_{ij}(\mathbf{u}, \mathbf{p}) & \text{if } f_{ij}(\mathbf{u}, \mathbf{p})(\bar{u}_j - \bar{u}_i) \leq 0, \\ 0 & \text{if } f_{ij}(\mathbf{u}, \mathbf{p})(\bar{u}_j - \bar{u}_i) > 0. \end{cases}$$

Then  $\tilde{\alpha}_{ij}$  again form a symmetric matrix and  $\alpha_{ij} \tilde{f}_{ij} = \tilde{\alpha}_{ij} f_{ij}$ . Therefore,  $S_i(\mathbf{u}, \mathbf{p})$  can be written in the form

$$\begin{aligned} S_i(\mathbf{u}, \mathbf{p}) &= \sum_{j=1}^M m_{ij} u_j + \theta \tau_{n+1} \sum_{j=1}^M A_{ij}(\mathbf{u}, \mathbf{p}) u_j \\ &\quad + \sum_{j=1}^M (1 - \tilde{\alpha}_{ij}(\mathbf{u}, \mathbf{p})) (-m_{ij} + \theta \tau_{n+1} D_{ij}(\mathbf{u}, \mathbf{p}))(u_j - u_i) \\ &\quad + \sum_{j=1}^M (1 - \tilde{\alpha}_{ij}(\mathbf{u}, \mathbf{p})) (m_{ij} + (1 - \theta) \tau_{n+1} d_{ij}^n)(u_j^n - u_i^n) \\ &\quad - [(\mathbb{M} - (1 - \theta) \tau_{n+1} \mathbb{A}^n) \mathbf{u}^n]_i. \end{aligned}$$

Denoting  $B_{ij} = (1 - \tilde{\alpha}_{ij}(\mathbf{u}, \mathbf{p})) (-m_{ij} + \theta \tau_{n+1} D_{ij}(\mathbf{u}, \mathbf{p}))$ , one has

$$\sum_{i,j=1}^M u_i (1 - \tilde{\alpha}_{ij}(\mathbf{u}, \mathbf{p})) (-m_{ij} + \theta \tau_{n+1} D_{ij}(\mathbf{u}, \mathbf{p}))(u_j - u_i) = -\frac{1}{2} \sum_{i,j=1}^M B_{ij} (u_i - u_j)^2 \geq 0,$$

since the matrix  $(B_{ij})_{i,j=1}^M$  is symmetric and has non-positive off-diagonal entries. Therefore, one again obtains (4.14) where the constants  $C_1, C_2$  are the same as in the proof of Theorem 4.5 and

$$C_3 = \|\mathbf{g}\|_M + \|(\mathbb{M} - (1 - \theta) \tau_{n+1} \mathbb{A}^n) \mathbf{u}^n\|_M,$$

where

$$g_i = \sum_{j=1}^M |m_{ij} + (1 - \theta) \tau_{n+1} d_{ij}^n| |u_j^n - u_i^n|, \quad i = 1, \dots, M.$$

Thus, in the same way as in the proof of Theorem 4.5, one concludes that there exists a solution  $\mathbf{U}$  of the equation  $P\mathbf{U} = 0$  and hence also a solution  $\mathbf{u}^{n+1}$ ,  $\mathbf{c}^{n+1}$ ,  $\mathbf{p}^{n+1}$  of (4.17), (4.6), and (3.13).

To prove the positivity preservation, we write (4.17) in the form (4.18)–(4.20). Since  $\mathbb{M}_L - (1 - \theta) \tau_{n+1} \mathbb{L}^n \geq 0$  according to Lemma 4.1, one has  $\bar{\mathbf{u}} \geq 0$ . Applying (4.21), one gets  $\tilde{\mathbf{u}} \geq 0$ . Since  $0 \leq \mathbf{c}^{n+1} \leq 1$  due to (4.6), it follows from Lemma 4.1 and Remark 4.3 that the matrix  $\mathbb{M}_L + \theta \tau_{n+1} \mathbb{L}^{n+1}$  is an M-matrix. Consequently,  $\mathbf{u}^{n+1} \geq 0$  in view of (4.20). Since (3.13) is equivalent to (3.11), one also has  $\mathbf{p}^{n+1} \geq 0$  and hence (3.12) is satisfied as well.  $\square$



## 5. Iterative solution of the FCT discretization

To compute a solution of the nonlinear problem (4.17), (3.12), (3.13) at time  $t_{n+1}$ , we will proceed similarly as for the Galerkin discretization in Section 3. Thus, given approximations  $\mathbf{u}_{k-1}^{n+1}$ ,  $\mathbf{c}_{k-1}^{n+1}$ ,  $\mathbf{p}_{k-1}^{n+1}$  (with some  $k > 0$ ) of  $\mathbf{u}^{n+1}$ ,  $\mathbf{c}^{n+1}$ ,  $\mathbf{p}^{n+1}$ , respectively, we compute  $\mathbf{c}_k^{n+1}$ ,  $\mathbf{p}_k^{n+1}$  using (3.14), (3.15). The iterate  $\mathbf{u}_k^{n+1}$  is computed by solving the linear system

$$(\mathbb{M}_L + \theta \tau_{n+1} \mathbb{L}_{k-1}^{n+1}) \mathbf{u}_k^{n+1} = (\mathbb{M}_L - (1 - \theta) \tau_{n+1} \mathbb{L}^n) \mathbf{u}^n + \left( \sum_{j=1}^M \alpha_{ij,k-1}^{n+1} f_{ij,k-1}^{n+1} \right)_{i=1}^M, \quad (5.1)$$

where  $\mathbb{L}_{k-1}^{n+1} = \mathbb{A}_{k-1}^{n+1} + \mathbb{D}_{k-1}^{n+1}$  with the matrix  $\mathbb{A}_{k-1}^{n+1}$  defined in (3.17) and the artificial diffusion matrix  $\mathbb{D}_{k-1}^{n+1}$  defined by

$$d_{ij,k-1}^{n+1} = -\max\{a_{ij,k-1}^{n+1}, 0, a_{ji,k-1}^{n+1}\} \quad \text{for } i \neq j, \quad d_{ii,k-1}^{n+1} = -\sum_{j=1, j \neq i}^M d_{ij,k-1}^{n+1}. \quad (5.2)$$

The algebraic fluxes  $f_{ij,k-1}^{n+1}$  are given by

$$f_{ij,k-1}^{n+1} = (-m_{ij} + \theta \tau_{n+1} d_{ij,k-1}^{n+1})(u_{j,k-1}^{n+1} - u_{i,k-1}^{n+1}) + (m_{ij} + (1 - \theta) \tau_{n+1} d_{ij}^n)(u_j^n - u_i^n) \quad (5.3)$$

and we again consider the prelimiting step

$$f_{ij,k-1}^{n+1} := 0 \quad \text{if } f_{ij,k-1}^{n+1}(\bar{u}_j - \bar{u}_i) > 0, \quad (5.4)$$

with  $\bar{\mathbf{u}}$  from (4.18). The limiters  $\alpha_{ij,k-1}^{n+1}$  are computed from the fluxes  $f_{ij,k-1}^{n+1}$  using the Zalesak algorithm (4.23)–(4.26). The following result shows that, under suitable time step restrictions, the above-defined iterates are uniquely determined and preserve non-negativity. This is important since, in practice, the fixed-point iterations are usually terminated when a stopping criterion is met, i.e., typically before reaching the solution of the nonlinear problem (4.17), (3.12), (3.13).

**Theorem 5.1.** Consider any  $n \in \{0, \dots, N-1\}$  and  $k \in \mathbb{N}$  and let  $\mathbf{u}^n, \mathbf{c}^n, \mathbf{p}^n \in \mathbb{R}^M$  and  $\mathbf{u}_{k-1}^{n+1}, \mathbf{p}_{k-1}^{n+1} \in \mathbb{R}^M$  be arbitrary vectors satisfying  $\mathbf{u}^n \geq 0$ ,  $1 \geq \mathbf{c}^n \geq 0$ ,  $\mathbf{p}^n \geq 0$ , and  $\mathbf{u}_{k-1}^{n+1} \geq 0$ ,  $\mathbf{p}_{k-1}^{n+1} \geq 0$ . Let  $\mathbf{c}_k^{n+1}$ ,  $\mathbf{p}_k^{n+1}$  be given by (3.14), (3.15). Let the time step  $\tau_{n+1}$  satisfy the conditions

$$(1 - \theta) \tau_{n+1} l_{ii}^n \leq m_i, \quad \theta \tau_{n+1} \left( \mu (1 - \pi_h \mathbf{u}_{k-1}^{n+1}, \phi_i) + \chi (\nabla(\pi_h \mathbf{c}_k^{n+1}), \nabla \phi_i) \right) < m_i, \quad i = 1, \dots, M. \quad (5.5)$$

Then the linear system (5.1) has a unique solution  $\mathbf{u}_k^{n+1}$  and one has  $\mathbf{u}_k^{n+1} \geq 0$ ,  $1 \geq \mathbf{c}_k^{n+1} \geq 0$ , and  $\mathbf{p}_k^{n+1} \geq 0$ .

**Proof.** The formula (3.14) immediately implies that  $1 \geq \mathbf{c}_k^{n+1} \geq 0$ . Since (3.15) can be written in the form (3.11) with  $u_{h,\tau}$  and  $c_{h,\tau}$  defined using  $\mathbf{u}_{k-1}^{n+1}$  and  $\mathbf{c}_{k-1}^{n+1}$ , respectively, at time  $t_{n+1}$ , one has  $\mathbf{p}_k^{n+1} \geq 0$ . Since  $\mathbb{M}_L - (1 - \theta) \tau_{n+1} \mathbb{L}^n \geq 0$  according to Lemma 4.1, the solution of (4.18) satisfies  $\bar{\mathbf{u}} \geq 0$ . Then (4.21) implies  $\tilde{\mathbf{u}} \geq 0$  for the solution of

$$\mathbb{M}_L \tilde{\mathbf{u}} = \mathbb{M}_L \bar{\mathbf{u}} + \left( \sum_{j=1}^M \alpha_{ij,k-1}^{n+1} f_{ij,k-1}^{n+1} \right)_{i=1}^M.$$

Finally, we use the fact that  $\mathbf{u}_k^{n+1}$  satisfies

$$(\mathbb{M}_L + \theta \tau_{n+1} \mathbb{L}_{k-1}^{n+1}) \mathbf{u}_k^{n+1} = \mathbb{M}_L \tilde{\mathbf{u}}. \quad (5.6)$$

It follows from the proof of Lemma 4.1 that, under the second condition in (5.5), the matrix  $(\mathbb{M}_L + \theta \tau_{n+1} \mathbb{L}_{k-1}^{n+1})$  is an M-matrix and hence  $\mathbf{u}_k^{n+1}$  is uniquely determined and satisfies  $\mathbf{u}_k^{n+1} \geq 0$ .  $\square$

**Remark 5.2.** From the physical point of view, the quantities  $u$ ,  $c$ , and  $p$  should be not only non-negative but also bounded by 1 from above. We have proved that this is the case for the approximations of  $c$ . Moreover, if this would be true also for the approximations of  $u$ , the integral form (3.9) would provide this property also for the approximations of  $p$ . Unfortunately, a proof of the upper bound for the approximations of  $u$  is not available and numerical results suggest that this bound can be violated. Note that a standard proof of upper bounds for FCT discretizations relies on the decomposition (4.18)–(4.20). Then, in particular, one would need that the solution of (4.18) satisfies  $\bar{\mathbf{u}} \leq \mathbf{1}$  if  $\mathbf{u}^n \leq \mathbf{1}$ . Choosing  $\mathbf{u}^n = \mathbf{1}$  (a vector with all components equal to 1), this requirement implies that  $(1 - \theta) \mathbb{A}^n \mathbf{1} = (1 - \theta) \mathbb{L}^n \mathbf{1} \geq 0$ , i.e., the row sums of the matrix  $(1 - \theta) \mathbb{A}^n$  have to be non-negative. Similarly, to derive an upper bound from (5.6), one would need that  $\theta \mathbb{A}_{k-1}^{n+1} \mathbf{1} \geq 0$ . It is clear that the validity of these row sum conditions cannot be expected.

**Remark 5.3.** The second condition on  $\tau_{n+1}$  in (5.5) depends on  $\mathbf{c}_k^{n+1}$  which itself depends on  $\tau_{n+1}$ . Consequently, in general, one has to proceed iteratively to find  $\tau_{n+1}$  which satisfies (5.5). To avoid this and also the dependence of  $\tau_{n+1}$  on the fixed-point iteration index  $k$ , it is possible to replace (5.5) by (4.5), cf. Remark 4.3.

We summarize the procedure for obtaining a high-resolution positivity preserving scheme for solving (2.1)–(2.5) in Algorithm 5.1.

---

**Algorithm 5.1** Iterative scheme for computing an approximation of the solution to the nonlinear FCT discretization.

---

```

1: Choose a tolerance Tol > 0 and a damping factor  $\beta \in (0, 1]$ .
2: Compute the initial values  $\mathbf{c}^0$ ,  $\mathbf{p}^0$ , and  $\mathbf{u}^0$  by (3.5).
3: Compute the mass matrix  $\mathbb{M}$  and the lumped mass matrix  $\mathbb{M}_L$ .
4: for  $n = 0, 1, \dots, N - 1$  do
5:   Compute the stiffness matrix  $\mathbb{A}^n$  and the artificial diffusion matrix  $\mathbb{D}^n$  and set  $\mathbb{L}^n = \mathbb{A}^n + \mathbb{D}^n$ .
6:   Choose  $\tau_{n+1}$  satisfying (4.5).
7:   Compute the intermediate solution  $\bar{\mathbf{u}}$  from (4.18).
8:   Set  $\mathbf{c}_0^{n+1} = \mathbf{c}^n$ ,  $\mathbf{p}_0^{n+1} = \mathbf{p}^n$ , and  $\mathbf{u}_0^{n+1} = \mathbf{u}^n$ .
9:   for  $k = 1, 2, \dots$  do
10:    Compute  $\mathbf{c}_k^{n+1}$  from (3.14) using  $\mathbf{c}^n$ ,  $\mathbf{p}^n$  and  $\mathbf{p}_{k-1}^{n+1}$ .
11:    Compute  $\mathbf{p}_k^{n+1}$  from (3.15) using  $\mathbf{p}^n$ ,  $\mathbf{c}^n$ ,  $\mathbf{u}^n$ ,  $\mathbf{c}_k^{n+1}$ , and  $\mathbf{u}_{k-1}^{n+1}$ .
12:    Compute the stiffness matrix  $\mathbb{A}_{k-1}^{n+1}$  from (3.17) using  $\mathbf{c}_k^{n+1}$  and  $\mathbf{u}_{k-1}^{n+1}$ .
13:    Compute the artificial diffusion matrix  $\mathbb{D}_{k-1}^{n+1}$  from (5.2) and set  $\mathbb{L}_{k-1}^{n+1} = \mathbb{A}_{k-1}^{n+1} + \mathbb{D}_{k-1}^{n+1}$ .
14:    Compute the algebraic fluxes  $f_{ij,k-1}^{n+1}$  from (5.3) and (5.4).
15:    Compute the limiters  $\alpha_{ij,k-1}^{n+1}$  by the Zalesak algorithm (4.23)–(4.26) using the fluxes  $f_{ij,k-1}^{n+1}$  and the intermediate solution  $\bar{\mathbf{u}}$ .
16:    Compute  $\mathbf{u}^{n+1}$  by solving the linear system (5.1).
17:    if  $\max \{ \|\mathbf{c}_k^{n+1} - \mathbf{c}_{k-1}^{n+1}\|_M, \|\mathbf{p}_k^{n+1} - \mathbf{p}_{k-1}^{n+1}\|_M, \|\mathbf{u}_k^{n+1} - \mathbf{u}_{k-1}^{n+1}\|_M \} < \text{Tol}$  then
18:      Go to line 23.
19:    else
20:      Set  $\mathbf{c}_k^{n+1} := \beta \mathbf{c}_k^{n+1} + (1 - \beta) \mathbf{c}_{k-1}^{n+1}$ ,  $\mathbf{p}_k^{n+1} := \beta \mathbf{p}_k^{n+1} + (1 - \beta) \mathbf{p}_{k-1}^{n+1}$ ,  $\mathbf{u}_k^{n+1} := \beta \mathbf{u}_k^{n+1} + (1 - \beta) \mathbf{u}_{k-1}^{n+1}$ .
21:    end if
22:  end for
23:  Set  $\mathbf{c}^{n+1} = \mathbf{c}_k^{n+1}$ ,  $\mathbf{p}^{n+1} = \mathbf{p}_k^{n+1}$ ,  $\mathbf{u}^{n+1} = \mathbf{u}_k^{n+1}$ .
24: end for

```

---

## 6. Numerical results

In the following, we present several numerical experiments to verify the positivity preserving properties of the proposed scheme for the model (2.1)–(2.5).

The computations are performed on a square domain  $\Omega = (0, 20)^2$  which is decomposed into quadrilateral mesh cells obtained by uniform refinements. Precisely, after  $r$  refinements, the triangulation  $\mathcal{T}_h$  consists of  $2^{2r}$  equal squares. If not otherwise stated, we consider five refinements, i.e.,  $\mathcal{T}_h$  consists of  $32 \times 32$  mesh cells. As explained above, conforming bilinear finite elements are used for approximating all unknown variables. The final time is  $T = 50$  and the parameter  $\epsilon = 0.2$  is used. The values of the remaining parameters of the model will be specified for the particular computations. The initial conditions are defined by

$$u^0(x) = e^{-|x|^2}, \quad c^0(x) = 1 - \frac{1}{2} e^{-|x|^2}, \quad p^0(x) = \frac{1}{2} e^{-|x|^2}.$$

If not otherwise stated, we apply the A-stable Crank-Nicolson method corresponding to  $\theta = 0.5$  for the time discretization. In one case, we will also discuss the application of the unconditionally stable backward Euler method corresponding to  $\theta = 1$ . Algorithm 5.1 is used with the tolerance  $\text{Tol} = 10^{-8}$  and the damping factor  $\beta = 0.5$ . The linear system (5.1) is solved using the sparse direct solver UMFPAK [59]. Our newly developed algorithms are implemented in the open-source finite element library deal.II [50,51]. The code enables to perform computations also in the 3D case, which we demonstrated in [45] for a Galerkin discretization. Since the results for 2D and 3D were qualitatively similar, we omit the 3D case in the present paper. Of course, 3D is interesting from the application viewpoint, which is further ongoing work with FCT, but conceptionally 3D does not add new insight to our proposed FCT scheme.

### 6.1. Comparison between the standard Galerkin FEM and the FEM-FCT scheme in presence of diffusion

To begin with, in the first example we consider the modified model subjected to an extra diffusion term in the equation (2.1) with diffusion coefficient  $\alpha^{-1}$ , as considered in [45], i.e., the equation (2.1) is replaced by (2.6). We consider  $\alpha = 10$ ,  $\chi = 1$ , and  $\mu = 1$ . As can be seen from Figs. 1 and 3, the FEM-FCT scheme introduces slightly more artificial diffusion than the standard Galerkin FEM. One can observe that the cancer cells invade the extracellular matrix and occupy the whole domain completely at the final time. Next, we decrease the amount of the diffusion by setting  $\alpha = 1000$  and keep the proliferation and haptotaxis rate as before. As can be seen from Figs. 5 and 6, the standard Galerkin FEM shows some oscillations in the front layer and the numerical simulation breaks down when the solution reaches the boundary of the computational domain, whereas applying the FEM-FCT removes the oscillations and keeps the solution positive at all times (Fig. 7). The corresponding snapshots of the cancer cell density, extracellular matrix, and protease are plotted along the line  $y = x$  in Figs. 2, 4, 6, and 8.

### 6.2. The FEM-FCT scheme in absence of diffusion for $\chi = 1$ , $\mu = 1$

In this section, we consider the case without the diffusion term, i.e., utilizing (2.1) now, and again set  $\chi = \mu = 1$ . This case was studied in [47,46], where the authors applied nonstandard finite difference (NSFD) schemes using Mickens rules. The proposed

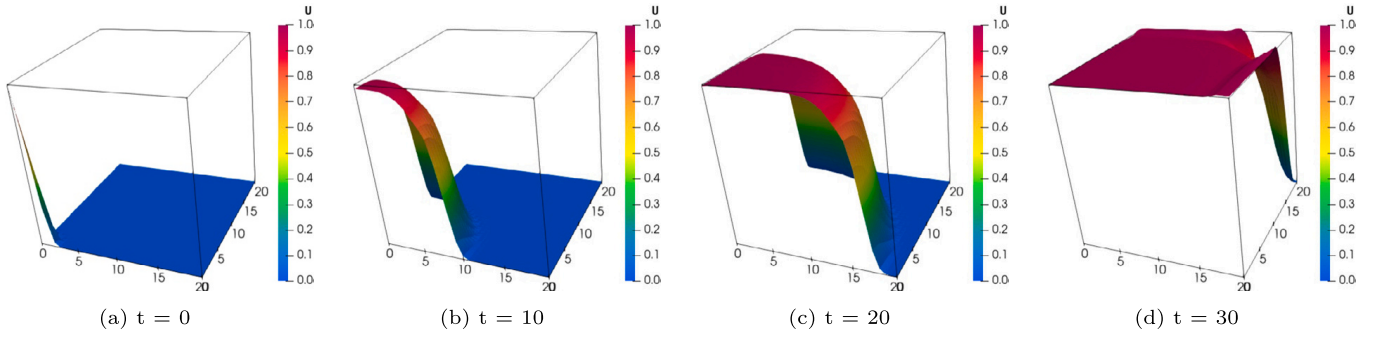


Fig. 1. Cancer cell invasion  $u$  at different time instants  $t = 0, 10, 20, 30$ , obtained with the standard Galerkin FEM for  $\alpha = 10$ ,  $\mu = 1$  and  $\chi = 1$ . (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

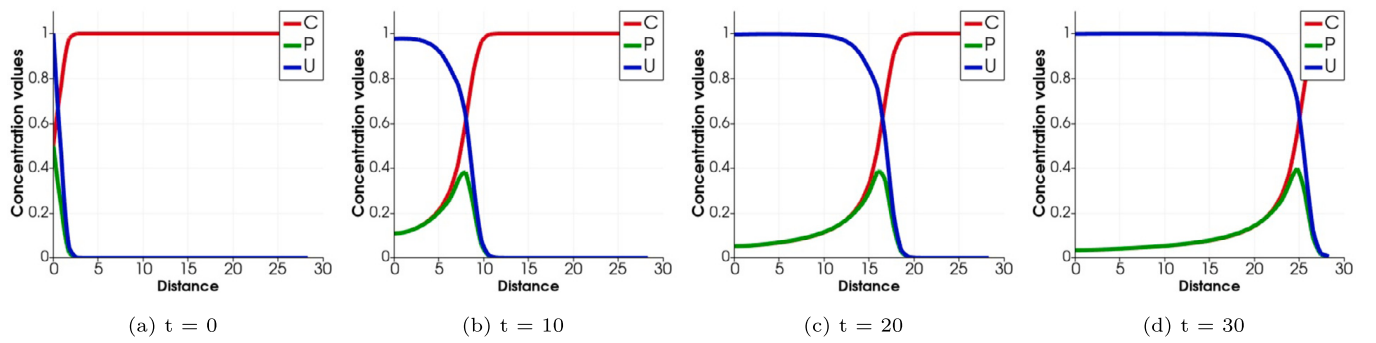


Fig. 2. Cancer cell invasion  $u$ , connective tissue  $c$ , and protease  $p$  at different time instants  $t = 0, 10, 20, 30$ , obtained with the standard Galerkin FEM for  $\alpha = 10$ ,  $\mu = 1$  and  $\chi = 1$ .

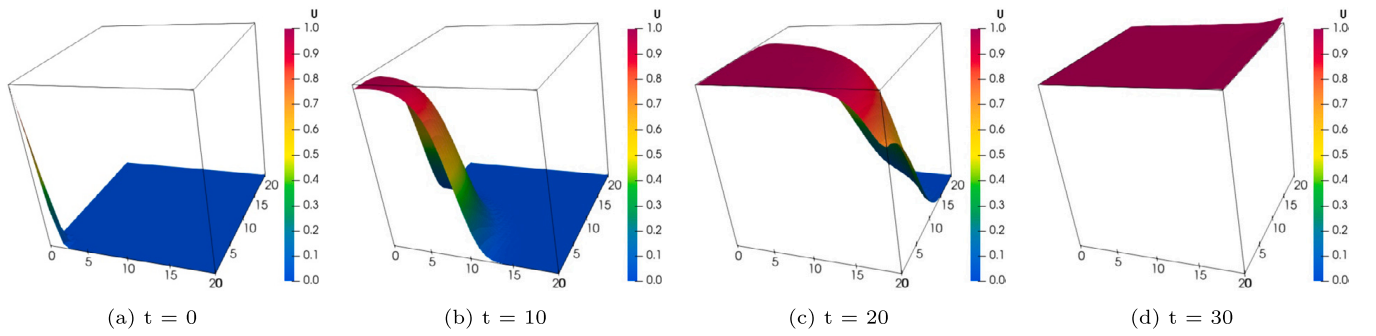


Fig. 3. Cancer cell invasion  $u$  at different time instants  $t = 0, 10, 20, 30$ , obtained with the FEM-FCT scheme for  $\alpha = 10$ ,  $\mu = 1$  and  $\chi = 1$ .

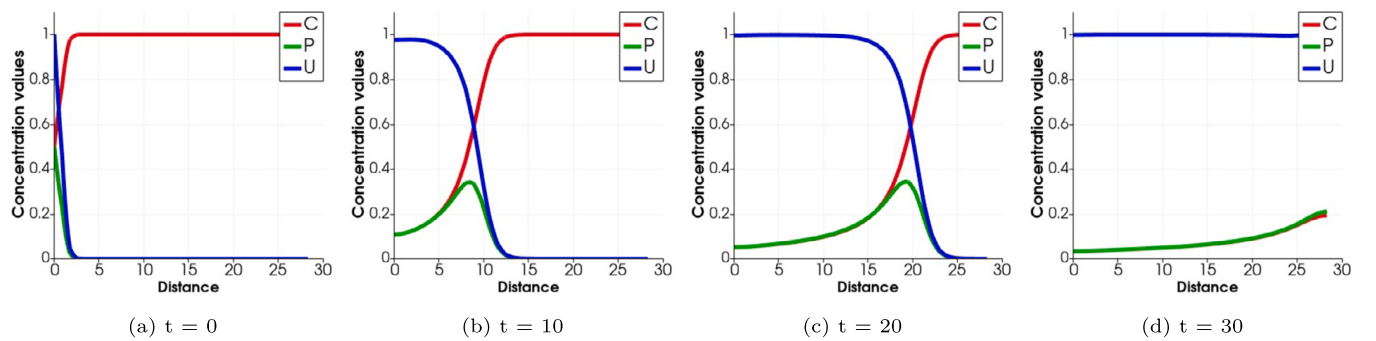


Fig. 4. Cancer cell invasion  $u$ , connective tissue  $c$ , and protease  $p$  at different time instants  $t = 0, 10, 20, 30$ , obtained with the FEM-FCT scheme for  $\alpha = 10$ ,  $\mu = 1$  and  $\chi = 1$ .

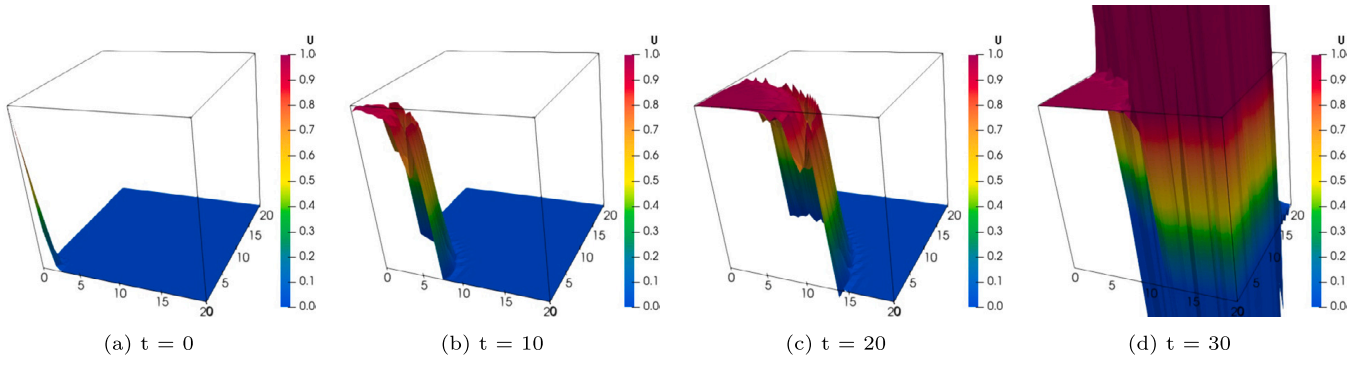


Fig. 5. Cancer cell invasion  $u$  at different time instants  $t = 0, 10, 20, 30$ , obtained with the standard Galerkin FEM for  $\alpha = 1000$ ,  $\mu = 1$  and  $\chi = 1$ .

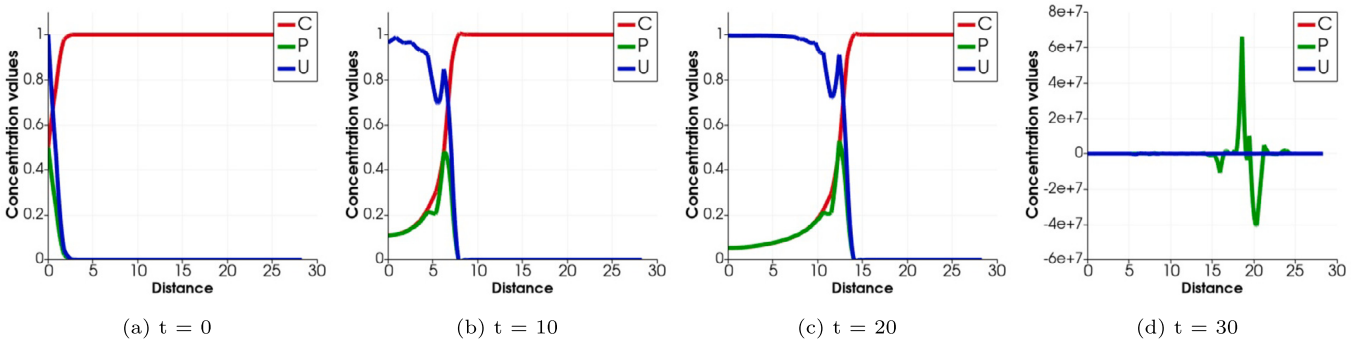


Fig. 6. Cancer cell invasion  $u$ , connective tissue  $c$ , and protease  $p$  at different time instants  $t = 0, 10, 20, 30$ , obtained with the standard Galerkin FEM for  $\alpha = 1000$ ,  $\mu = 1$  and  $\chi = 1$ .

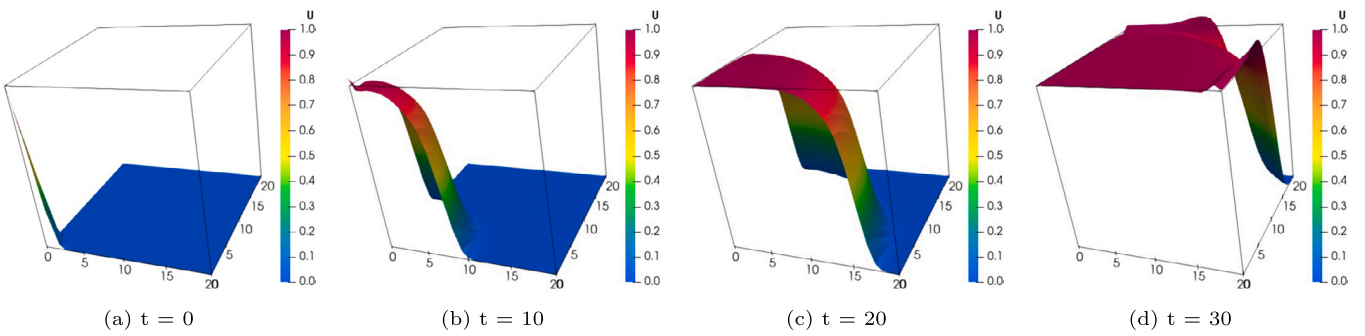


Fig. 7. Cancer cell invasion  $u$  at different time instants  $t = 0, 10, 20, 30$ , obtained with the FEM-FCT scheme for  $\alpha = 1000$ ,  $\mu = 1$  and  $\chi = 1$ .

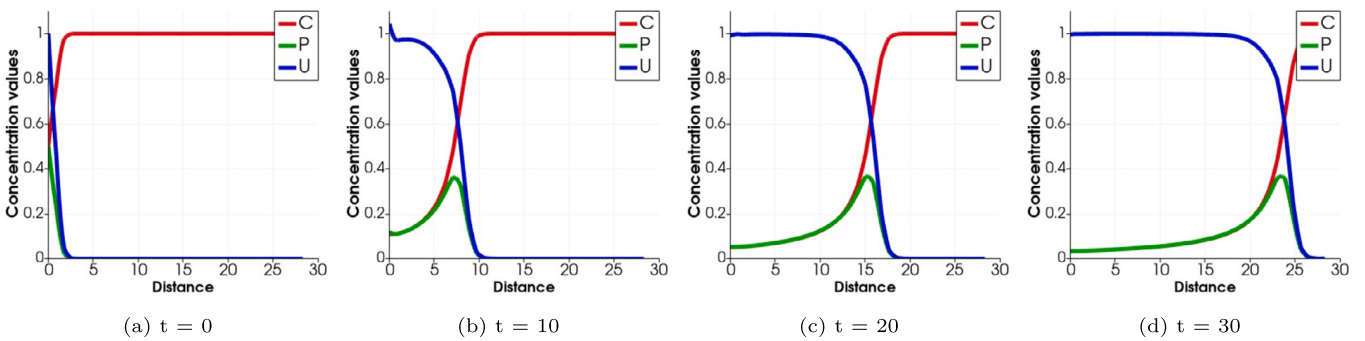


Fig. 8. Cancer cell invasion  $u$ , connective tissue  $c$ , and protease  $p$  at different time instants  $t = 0, 10, 20, 30$ , obtained with the FEM-FCT scheme for  $\alpha = 1000$ ,  $\mu = 1$  and  $\chi = 1$ .

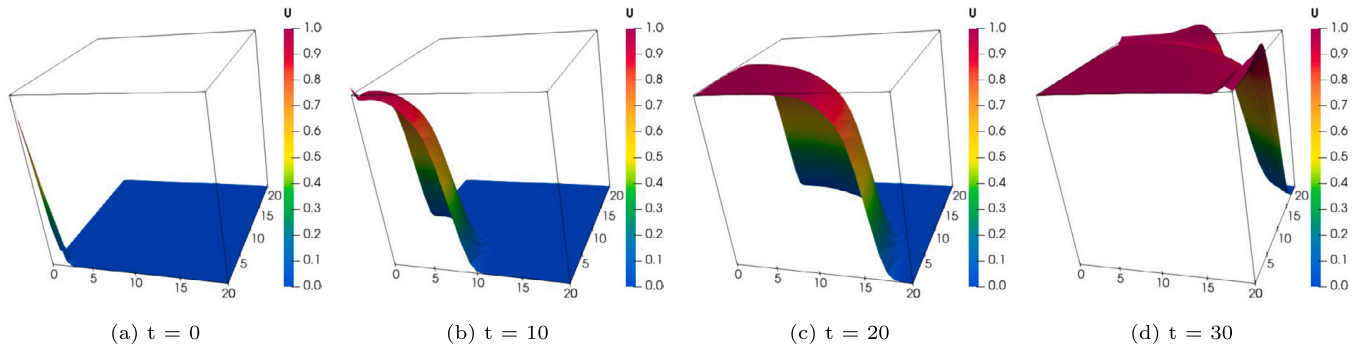


Fig. 9. Cancer cell invasion  $u$  at different time instants  $t = 0, 10, 20, 30$ , obtained with the FEM-FCT scheme for  $\mu = 1$  and  $\chi = 1$ .

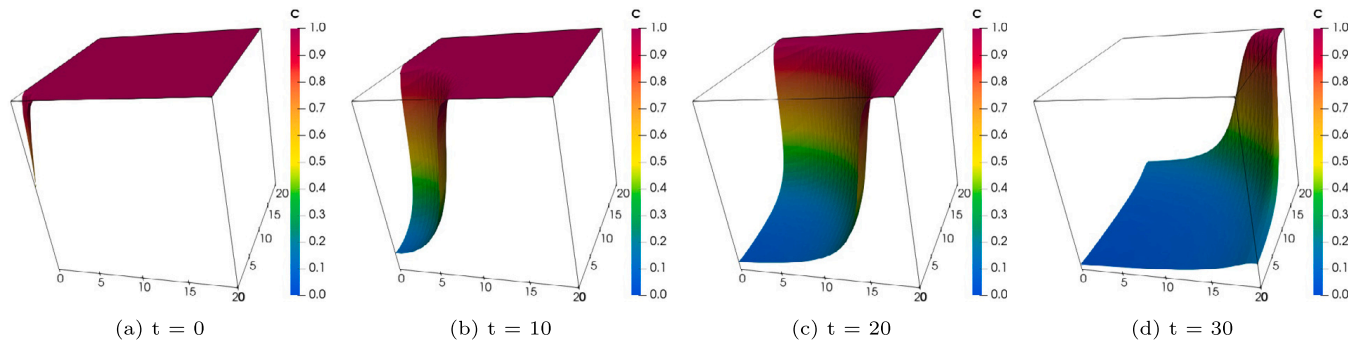


Fig. 10. Decay of the extracellular matrix  $c$  at different time instants  $t = 0, 10, 20, 30$ , computed using the FEM-FCT scheme for  $\mu = 1$  and  $\chi = 1$ .

Table 1

Convergence of the mean values with respect to global mesh refinement at the last time instant  $t = 50$ .

# of refinements	3	4	5	6	7
# DOF	81	289	1089	4225	16641
$\int_{\Omega} c_n(x) dx$	0.02362193	0.02670467	0.03284535	0.03042843	0.03680451
$\int_{\Omega} p_n(x) dx$	0.02373726	0.02685417	0.03308441	0.03062835	0.03712137
$\int_{\Omega} u_n(x) dx$	0.99999999	0.99999998	0.99999976	0.99999943	0.99999889

Table 2

Convergence of the solutions at the point  $(20, 20)$  with respect to the time step  $\tau$  at the last time instant  $t = 50$ .

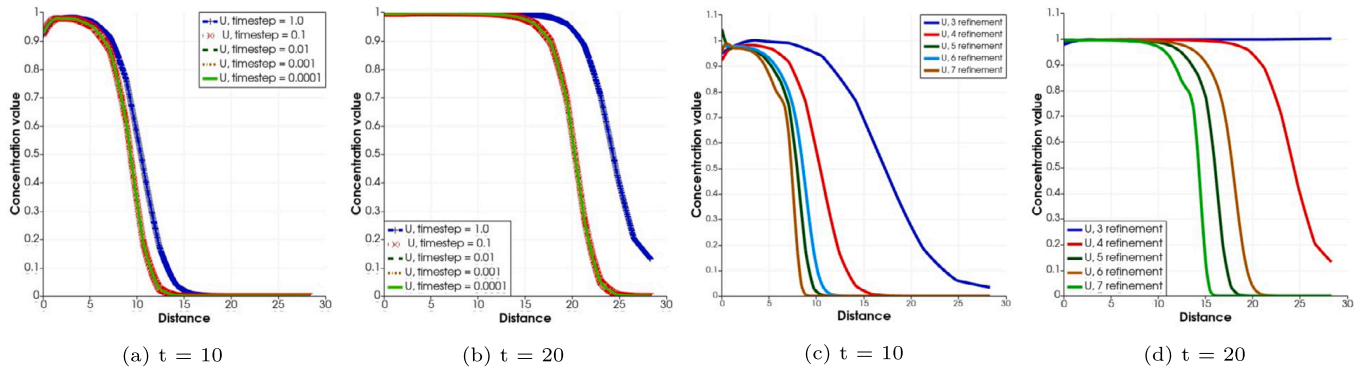
$\tau$	$c_k^{n+1}$	$\ c_k^{n+1} - c_{k-1}^{n+1}\ $	$p_k^{n+1}$	$\ p_k^{n+1} - p_{k-1}^{n+1}\ $	$u_k^{n+1}$	$\ u_k^{n+1} - u_{k-1}^{n+1}\ $
1.0	0.03331	4.64077e-09	0.03356	5.03212e-09	1.00046	6.91815e-11
0.1	0.03884	5.57720e-09	0.03918	5.77532e-09	1.00079	4.47103e-11
0.01	0.03875	9.08222e-09	0.03909	9.21976e-09	1.00070	1.44008e-10
0.001	0.03875	7.27756e-09	0.03909	7.37382e-09	1.00079	1.21648e-10
0.0001	0.03875	5.82277e-09	0.03909	5.89891e-09	1.00079	9.77287e-11

methods were successful in comparison to standard finite difference methods at obtaining positive solutions, however, some wiggles still remained in the vicinity of the front layer. On the other hand, deriving an efficient NSFD scheme heavily depends on the type of the system and the discretization of different terms. Therefore, in this work, we applied the FEM-FCT methodology to remove the oscillations in the front layer while keeping the solutions positive at all times, see Figs. 9 and 10. Next, we check numerically whether the approximate solutions converge. To this end, we computed the integrals of the solutions at the final time  $t = 50$  for different numbers of global refinements, see Table 1. The results correspond to the situation where the tumor is completely malignant and invades the whole extracellular matrix. In Table 2, we study the values of the solutions at the point  $(20, 20)$  and the differences between two consecutive iterative solutions. Moreover, the numbers of fixed-point iterations for various time steps and time instants are shown in Table 3. In particular, we observe that the proposed scheme is convergent with respect to the time step size. The convergence of the cancer cell invasion  $u$  with respect to the time step and the mesh width at two different time instants is also studied in Fig. 11 by means of solution graphs along the line  $y = x$ .

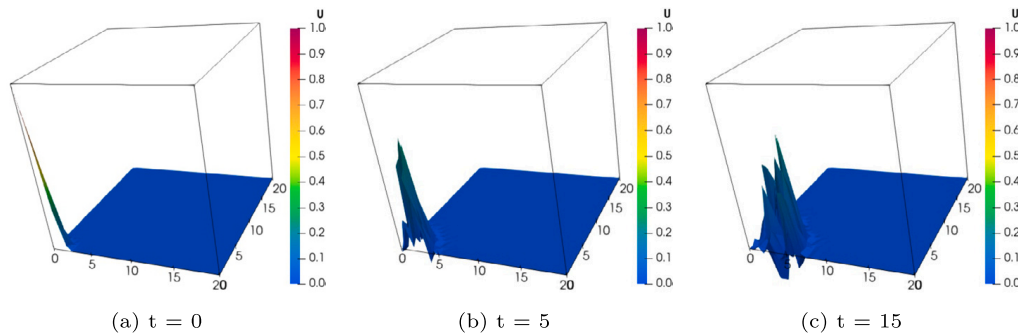
**Table 3**

Numbers of fixed-point iterations for various time steps at time instants  $t = 1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50$ .

$\tau$	# of iterations at time instants										
	1	5	10	15	20	25	30	35	40	45	50
1.0	30	31	31	32	30	23	23	22	21	21	21
0.1	25	25	26	26	26	25	21	20	19	18	18
0.01	19	20	21	21	21	20	18	16	16	15	14
0.001	15	16	18	18	18	17	14	13	12	12	11
0.0001	13	14	14	15	15	14	12	11	10	9	8



**Fig. 11.** Cancer cell invasion  $u$  computed using the FEM-FCT scheme at time instants  $t = 10$  and  $t = 20$  for different time steps (first two pictures) and for different numbers of global refinements (last two pictures).



**Fig. 12.** The effect of the haptotactic rate on the cancer cell invasion  $u$  at different time instants  $t = 0, 5, 15$ , computed using the standard Galerkin FEM for  $\mu = 0.0001$  and  $\chi = 1$ .

### 6.3. Effect of haptotactic domination

In this section, we investigate the effect of directional movement of cancer cells inside the domain. This is a very important property in cancer modeling which can lead to metastasis. In metastasis, the cancer cells are moving to the other parts of the body and start proliferate, forming a new tumor in the new part, and invade the surrounding tissues. In this case, it is very difficult to detect the location of cancerous cells and this is one of the predominant causes of most deaths due to cancer. In the following, we only study a very simple case of haptotactic dominating mechanism of the cancer cell motion. In addition to the absence of the diffusion effect in the system, there is only a small amount of the proliferation rate: we set  $\mu = 0.0001$  and  $\chi = 1$  in the computations. As a result of the haptotactic migration domination, a small cluster of cancer cells builds up at the beginning and this initial amount is expected to move along the direction of the gradient of the extracellular matrix. As Fig. 12 indicates, the numerical simulation by the standard Galerkin FEM breaks down in a very short amount of time after the time instant  $t = 15$ . Next, we apply the FEM-FCT scheme and the low-order method, see Figs. 13 and 14, respectively. We observe that, in both cases, the stabilization prevents the blow-up in the system and leads to non-negative solutions. However, some oscillations still remain in the interior layer. These oscillations could be suppressed by adaptive mesh refinement, which is however out of the scope of this paper. As expected, the low-order method provides a more diffusive solution than the FEM-FCT scheme. It is interesting that, combining the FEM-FCT scheme with the backward Euler method ( $\theta = 1$ ), oscillation-free solutions are obtained, see Fig. 15.

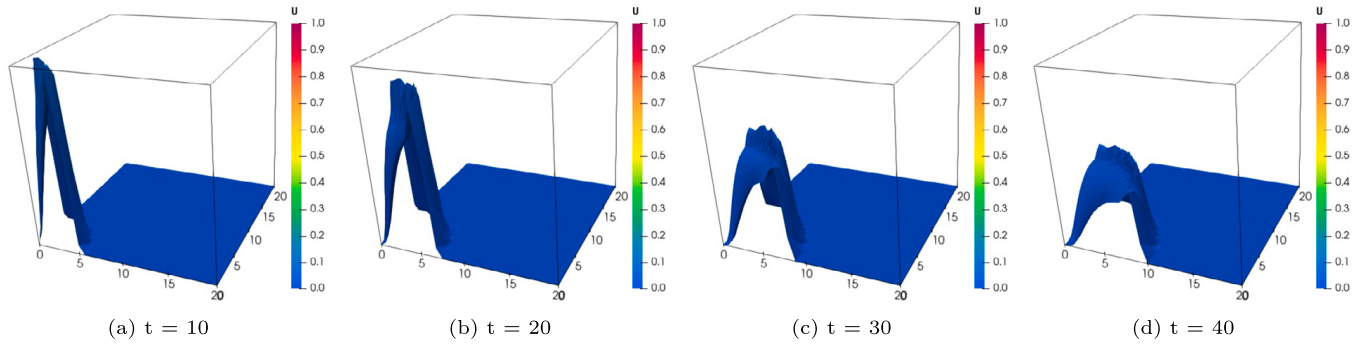


Fig. 13. The effect of the haptotactic rate on the cancer cell invasion  $u$  at different time instants  $t = 10, 20, 30, 40$ , computed using the FEM-FCT scheme for  $\mu = 0.0001$  and  $\chi = 1$ .

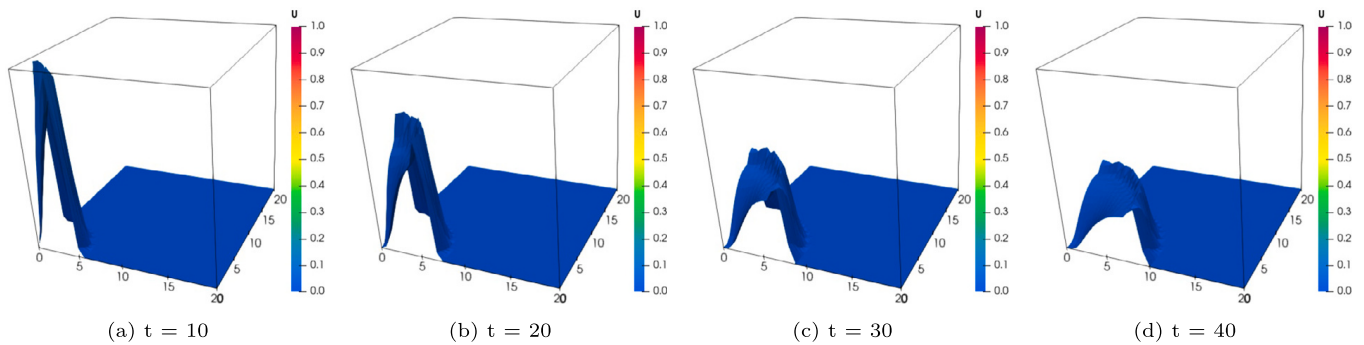


Fig. 14. The effect of the haptotactic rate on the cancer cell invasion  $u$  at different time instants  $t = 10, 20, 30, 40$ , computed using the low-order method for  $\mu = 0.0001$  and  $\chi = 1$ .

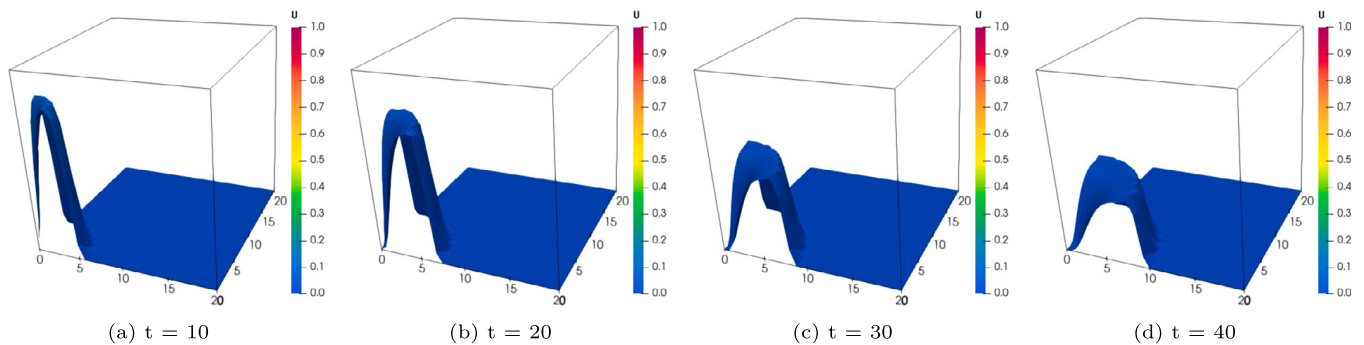


Fig. 15. The effect of the haptotactic rate on the cancer cell invasion  $u$  at different time instants  $t = 10, 20, 30, 40$ , computed using the FEM-FCT scheme with  $\theta = 1$  for  $\mu = 0.0001$  and  $\chi = 1$ .

### 7. Conclusions

In this paper, we proposed a fully discrete nonlinear high-resolution positivity preserving FEM-FCT scheme for haptotaxis equations without self-diffusion term describing a model of cancer invasion. We proved the solvability and positivity preservation of both the nonlinear discrete problem and the linear problems appearing in fixed-point iterations. A series of numerical experiments are shown to verify the robustness of the proposed method. Derivation of error estimates is left to future work.

### CRediT authorship contribution statement

**Shahin Heydari:** Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. **Petr Knobloch:** Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. **Thomas Wick:** Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing.

### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Acknowledgements

This work was initiated during a research stay of the first author at the Institute of Applied Mathematics at the Leibniz University Hanover from November 2021 to April 2022 for which hospitality is still gratefully acknowledged. The work of Shahin Heydari was further supported through the grant No. 396921 of the Charles University Grant Agency and the grant SVV-2023-260711 of Charles University. The work of Petr Knobloch was supported through the grant No. 22-01591S of the Czech Science Foundation.

## References

- [1] E.F. Keller, L.A. Segel, Model for chemotaxis, *J. Theor. Biol.* 30 (2) (1971) 225–234.
- [2] E.F. Keller, L.A. Segel, Initiation of slime mold aggregation viewed as an instability, *J. Theor. Biol.* 26 (3) (1970) 399–415.
- [3] M. Aida, T. Tsujikawa, M. Efendiev, A. Yagi, M. Mimura, Lower estimate of the attractor dimension for a chemotaxis growth system, *J. Lond. Math. Soc.* (2) 74 (2) (2006) 453–474, <https://doi.org/10.1112/S0024610706023015>.
- [4] M. Mimura, T. Tsujikawa, Aggregating pattern dynamics in a chemotaxis model including growth, *Phys. A, Stat. Mech. Appl.* 230 (3–4) (1996) 499–543.
- [5] R. Tyson, S. Lubkin, J.D. Murray, A minimal mechanism for bacterial pattern formation, *Proc. R. Soc. Lond. B, Biol. Sci.* 266 (1416) (1999) 299–304.
- [6] A.R. Anderson, M.A. Chaplain, E.L. Newman, R.J. Steele, A.M. Thompson, Mathematical modelling of tumour invasion and metastasis, *Comput. Math. Methods Med.* 2 (2) (2000) 129–154.
- [7] M.A.J. Chaplain, G. Lolas, Mathematical modelling of cancer cell invasion of tissue: the role of the urokinase plasminogen activation system, *Math. Models Methods Appl. Sci.* 15 (11) (2005) 1685–1734, <https://doi.org/10.1142/S0218202505000947>.
- [8] M.A.J. Chaplain, G. Lolas, Mathematical modelling of cancer invasion of tissue: dynamic heterogeneity, *Netw. Heterog. Media* 1 (3) (2006) 399–439, <https://doi.org/10.3934/nhm.2006.1.399>.
- [9] M.A. Chaplain, A.M. Stuart, A model mechanism for the chemotactic response of endothelial cells to tumour angiogenesis factor, *Math. Med. Biol.* 10 (3) (1993) 149–168.
- [10] M. Aida, A. Yagi, Target pattern solutions for chemotaxis-growth system, *Sci. Math. Jpn.* 59 (3) (2004) 577–590.
- [11] D. Wu, Signaling mechanisms for regulation of chemotaxis, *Cell Res.* 15 (1) (2005) 52–56.
- [12] V. Nanjundiah, Chemotaxis, signal relaying and aggregation morphology, *J. Theor. Biol.* 42 (1) (1973) 63–105.
- [13] L. Corrias, B. Perthame, H. Zaag, Global solutions of some chemotaxis and angiogenesis systems in high space dimensions, *Milan J. Math.* 72 (2004) 1–28, <https://doi.org/10.1007/s00032-003-0026-x>.
- [14] D. Horstmann, M. Winkler, Boundedness vs. blow-up in a chemotaxis system, *J. Differ. Equ.* 215 (1) (2005) 52–107, <https://doi.org/10.1016/j.jde.2004.10.022>.
- [15] Y. Tao, M. Wang, A combined chemotaxis-haptotaxis system: the role of logistic source, *SIAM J. Math. Anal.* 41 (4) (2009) 1533–1558, <https://doi.org/10.1137/090751542>.
- [16] D. Horstmann, M. Lucia, Uniqueness and symmetry of equilibria in a chemotaxis model, *J. Reine Angew. Math.* 654 (2011) 83–124, <https://doi.org/10.1515/CRELLE.2011.030>.
- [17] V. Calvez, L. Corrias, M.A. Ebde, Blow-up, concentration phenomenon and global existence for the Keller–Segel model in high dimension, *Commun. Partial Differ. Equ.* 37 (4) (2012) 561–584, <https://doi.org/10.1080/03605302.2012.655824>.
- [18] M.K. Kolev, M.N. Koleva, L.G. Vulkov, An unconditional positivity-preserving difference scheme for models of cancer migration and invasion, *Mathematics* 10 (1) (2022) 131.
- [19] Y. Epshteyn, A. Kurganov, New interior penalty discontinuous Galerkin methods for the Keller–Segel chemotaxis model, *SIAM J. Numer. Anal.* 47 (1) (2008/09) 386–408, <https://doi.org/10.1137/07070423X>.
- [20] X.H. Li, C.-W. Shu, Y. Yang, Local discontinuous Galerkin method for the Keller–Segel chemotaxis model, *J. Sci. Comput.* 73 (2–3) (2017) 943–967, <https://doi.org/10.1007/s10915-016-0354-y>.
- [21] N. Saito, Conservative upwind finite-element method for a simplified Keller–Segel system modelling chemotaxis, *IMA J. Numer. Anal.* 27 (2) (2007) 332–365, <https://doi.org/10.1093/imanum/drl018>.
- [22] J. Zhang, J. Zhu, R. Zhang, Characteristic splitting mixed finite element analysis of Keller–Segel chemotaxis models, *Appl. Math. Comput.* 278 (2016) 33–44, <https://doi.org/10.1016/j.amc.2016.01.021>.
- [23] S. Zhao, X. Xiao, J. Zhao, X. Feng, A Petrov–Galerkin finite element method for simulating chemotaxis models on stationary surfaces, *Comput. Math. Appl.* 79 (11) (2020) 3189–3205, <https://doi.org/10.1016/j.camwa.2020.01.019>.
- [24] X. Xiao, X. Feng, Y. He, Numerical simulations for the chemotaxis models on surfaces via a novel characteristic finite element method, *Comput. Math. Appl.* 78 (1) (2019) 20–34, <https://doi.org/10.1016/j.camwa.2019.02.004>.
- [25] F. Filbet, A finite volume scheme for the Patlak–Keller–Segel chemotaxis model, *Numer. Math.* 104 (4) (2006) 457–488, <https://doi.org/10.1007/s00211-006-0024-3>.
- [26] D.L. Ropp, J.N. Shadid, Stability of operator splitting methods for systems with indefinite operators: advection-diffusion-reaction systems, *J. Comput. Phys.* 228 (9) (2009) 3508–3516, <https://doi.org/10.1016/j.jcp.2009.02.001>.
- [27] R. Tyson, L.G. Stern, R.J. LeVeque, Fractional step methods applied to a chemotaxis model, *J. Math. Biol.* 41 (5) (2000) 455–475, <https://doi.org/10.1007/s002850000038>.
- [28] J.P. Boris, D.L. Book, Flux-corrected transport. I. SHASTA, a fluid transport algorithm that works, *J. Comput. Phys.* 11 (1) (1973) 38–69, [https://doi.org/10.1016/0021-9991\(73\)90147-2](https://doi.org/10.1016/0021-9991(73)90147-2).
- [29] D. Book, J. Boris, K. Hain, Flux-corrected transport II: Generalizations of the method, *J. Comput. Phys.* 18 (3) (1975) 248–283.
- [30] J. Boris, D. Book, Flux-corrected transport. III. Minimal-error FCT algorithms, *J. Comput. Phys.* 20 (4) (1976) 397–431.
- [31] R. Löhner, K. Morgan, J. Peraire, M. Vahdati, Finite element flux-corrected transport (FEM–FCT) for the Euler and Navier–Stokes equations, *Int. J. Numer. Methods Fluids* 7 (10) (1987) 1093–1109.
- [32] D. Kuzmin, Explicit and implicit FEM–FCT algorithms with flux linearization, *J. Comput. Phys.* 228 (7) (2009) 2517–2534, <https://doi.org/10.1016/j.jcp.2008.12.011>.
- [33] D. Kuzmin, Algebraic flux correction I. Scalar conservation laws, in: D. Kuzmin, R. Löhner, S. Turek (Eds.), *Flux-Corrected Transport. Principles, Algorithms, and Applications*, 2nd edition, Springer, Dordrecht, 2012, pp. 145–192.
- [34] D. Kuzmin, S. Turek, Flux correction tools for finite elements, *J. Comput. Phys.* 175 (2) (2002) 525–558, <https://doi.org/10.1006/jcph.2001.6955>.
- [35] D. Feng, I. Neuweiler, U. Nackenhorst, T. Wick, A time-space flux-corrected transport finite element formulation for solving multi-dimensional advection-diffusion-reaction equations, *J. Comput. Phys.* 396 (2019) 31–53, <https://doi.org/10.1016/j.jcp.2019.06.053>.



- [36] R. Strehl, A. Sokolov, D. Kuzmin, S. Turek, A flux-corrected finite element method for chemotaxis problems, *Comput. Methods Appl. Math.* 10 (2) (2010) 219–232, <https://doi.org/10.2478/cmam-2010-0013>.
- [37] R. Strehl, A. Sokolov, D. Kuzmin, D. Horstmann, S. Turek, A positivity-preserving finite element method for chemotaxis problems in 3D, *J. Comput. Appl. Math.* 239 (2013) 290–303, <https://doi.org/10.1016/j.cam.2012.09.041>.
- [38] A. Sokolov, R. Strehl, S. Turek, Numerical simulation of chemotaxis models on stationary surfaces, *Discrete Contin. Dyn. Syst., Ser. B* 18 (10) (2013) 2689–2704, <https://doi.org/10.3934/dcdsb.2013.18.2689>.
- [39] A. Sokolov, R. Ali, S. Turek, An AFC-stabilized implicit finite element method for partial differential equations on evolving-in-time surfaces, *J. Comput. Appl. Math.* 289 (2015) 101–115, <https://doi.org/10.1016/j.cam.2015.03.002>.
- [40] X. Huang, X. Xiao, J. Zhao, X. Feng, An efficient operator-splitting FEM-FCT algorithm for 3D chemotaxis models, *Eng. Comput.* 36 (4) (2020) 1393–1404.
- [41] M. Sulman, T. Nguyen, A positivity preserving moving mesh finite element method for the Keller–Segel chemotaxis model, *J. Sci. Comput.* 80 (1) (2019) 649–666, <https://doi.org/10.1007/s10915-019-00951-0>.
- [42] X. Huang, X. Feng, X. Xiao, K. Wang, Fully decoupled, linear and positivity-preserving scheme for the chemotaxis–Stokes equations, *Comput. Methods Appl. Mech. Eng.* 383 (2021) 113909, <https://doi.org/10.1016/j.cma.2021.113909>.
- [43] X. Feng, X. Huang, K. Wang, Error estimate of unconditionally stable and decoupled linear positivity-preserving FEM for the chemotaxis–Stokes equations, *SIAM J. Numer. Anal.* 59 (6) (2021) 3052–3076, <https://doi.org/10.1137/21M142085X>.
- [44] A.J. Perumpanani, J.A. Sherratt, J. Norbury, H.M. Byrne, A two parameter family of travelling waves with a singular barrier arising from the modelling of extracellular matrix mediated cellular invasion, *Physica D* 126 (3–4) (1999) 145–159.
- [45] M. Fuest, S. Heydari, P. Knobloch, J. Lankeit, T. Wick, Global existence of classical solutions and numerical simulations of a cancer invasion model, *ESAIM: Math. Model. Numer. Anal.* 57 (4) (2023) 1893–1919, <https://doi.org/10.1051/m2an/2023037>.
- [46] M.M. Khalsaraei, S. Heydari, L.D. Algoo, Positivity preserving nonstandard finite difference schemes applied to cancer growth model, *J. Cancer Treat. Res.* 4 (4) (2016) 27–33.
- [47] M. Chapwanya, J.M.-S. Lubuma, R.E. Mickens, Positivity-preserving nonstandard finite difference schemes for cross-diffusion equations in biosciences, *Comput. Math. Appl.* 68 (9) (2014) 1071–1082, <https://doi.org/10.1016/j.camwa.2014.04.021>.
- [48] V. John, P. Knobloch, P. Korsemeier, On the solvability of the nonlinear problems in an algebraically stabilized finite element method for evolutionary transport-dominated equations, *Math. Comput.* 90 (328) (2021) 595–611, <https://doi.org/10.1090/mcom/3576>.
- [49] V. John, P. Knobloch, Existence of solutions of a finite element flux-corrected-transport scheme, *Appl. Math. Lett.* 115 (2021) 106932, <https://doi.org/10.1016/j.aml.2020.106932>.
- [50] D. Arndt, W. Bangerth, D. Davydov, T. Heister, L. Heltai, M. Kronbichler, M. Maier, J.-P. Pelteret, B. Turcksin, D. Wells, The DEAL.II finite element library: design, features, and insights, *Comput. Math. Appl.* 81 (2021) 407–422, <https://doi.org/10.1016/j.camwa.2020.02.022>.
- [51] D. Arndt, W. Bangerth, M. Feder, M. Fehling, R. Gassmüller, T. Heister, L. Heltai, M. Kronbichler, M. Maier, P. Munch, J.-P. Pelteret, S. Stiecko, B. Turcksin, D. Wells, The deal.II library, Version 9.4, *J. Numer. Math.* 30 (3) (2022) 231–246, <https://doi.org/10.1515/jnma-2022-0054>.
- [52] B.P. Marchant, J. Norbury, A.J. Perumpanani, Travelling shock waves arising in a model of malignant invasion, *SIAM J. Appl. Math.* 60 (2) (2000) 463–476, <https://doi.org/10.1137/S0036139998328034>.
- [53] B.P. Marchant, J. Norbury, J.A. Sherratt, Travelling wave solutions to a haptotaxis-dominated model of malignant invasion, *Nonlinearity* 14 (6) (2001) 1653–1671, <https://doi.org/10.1088/0951-7715/14/6/313>.
- [54] P.G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
- [55] R.S. Varga, *Matrix Iterative Analysis*, Springer-Verlag, Berlin, 2000.
- [56] G.R. Barrenechea, V. John, P. Knobloch, Finite element methods respecting the discrete maximum principle for convection-diffusion equations, *SIAM Rev.* 66 (1) (2024), <https://doi.org/10.1137/22M1488934>.
- [57] R. Temam, *Navier-Stokes Equations. Theory and Numerical Analysis*, North-Holland, Amsterdam, 1977.
- [58] S.T. Zalesak, Fully multidimensional flux-corrected transport algorithms for fluids, *J. Comput. Phys.* 31 (3) (1979) 335–362, [https://doi.org/10.1016/0021-9991\(79\)90051-2](https://doi.org/10.1016/0021-9991(79)90051-2).
- [59] T.A. Davis, Algorithm 832: UMFPAK V4.3—an unsymmetric-pattern multifrontal method, *ACM Trans. Math. Softw.* 30 (2) (2004) 196–199, <https://doi.org/10.1145/992200.992206>.

## **3.2 Solvability and numerical solution of a cross-diffusion cancer invasion model**

In this paper we mainly focused on the idea behind the modeling of the cancer invasion system studied so far in the preceding and summarized the main results. Moreover, we presented new numerical experiments to provide further insight into the behavior of the model utilizing different variations of the parameters appearing in the system and also using the different choices of the limiters for the proposed flux-corrected transport approach.

# Solvability and Numerical Solution of a Cross-diffusion Cancer Invasion Model

Shahin Heydari<sup>[0009-0007-9128-5257]</sup> and  
Petr Knobloch<sup>[0000-0003-2709-5882]</sup>

**Abstract** We consider a model of the invasion of healthy tissue by cancer cells which is described by a system of nonlinear PDEs consisting of a cross-diffusion-reaction equation and two additional nonlinear ordinary differential equations. We discuss the existence of global classical solutions and formulate a positivity preserving finite element discretization stabilized by the flux-corrected transport approach. Moreover, we present a result on the solvability of this nonlinear discrete problem. The properties of both the model and its discretization are illustrated by numerical results computed using the deal.II library.

## 1 Introduction

Mathematical modelling and numerical simulations of cancer invasion are important for a better understanding of the mechanisms governing the growth of malignant tumours and for developing strategies to cure this dangerous disease. In this work, we focus on a nonlinear cancer-invasion model developed in [7] that models the motion of cancer cells, degradation of extracellular matrix, and production of protease. The aim of the paper is to provide a concise description of various aspects of cancer growth modelling: from ideas behind the model, over its analytical investigations, application of various discretization techniques and investigations of the discrete problems, till numerical simulations. We will both review our recent theoretical achievements published in [2, 3] and present new numerical results providing further insight into the behaviour of the model and the proposed discretization.

---

Shahin Heydari  
Charles University, Faculty of Mathematics and Physics, Department of Numerical Mathematics,  
Sokolovská 83, 186 75 Praha 8, Czech Republic, e-mail: heydari@karlin.mff.cuni.cz

Petr Knobloch  
Charles University, Faculty of Mathematics and Physics, Department of Numerical Mathematics,  
Sokolovská 83, 186 75 Praha 8, Czech Republic, e-mail: knobloch@karlin.mff.cuni.cz

The plan of the paper is as follows. In Section 2, we introduce the mathematical model and thoroughly explain its background. Section 3 discusses the existence of global classical solutions to the model and presents the main ideas of the proof in the case when a diffusion term is present. Section 4 is devoted to the discretization of the model. The time derivatives are approximated using finite differences whereas a Galerkin finite element scheme is considered for spatial discretization. To guarantee positivity preservation, an algebraic stabilization is introduced, which finally leads to a high-resolution flux-corrected transport (FCT) scheme. A result on the solvability of this nonlinear discrete problem is formulated and main ideas of the proof are explained. Finally, numerical results are presented in Section 5.

## 2 Mathematical Model

In [7], Perumpanani et al. proposed a model for malignant cancer invasion considering proteolysis and haptotaxis as the main mechanisms. The model was formulated for one spatial coordinate in the direction of the invasion. In more space dimensions, and with a particular choice of the functions describing the considered effects, the model can be written in the form

$$u_t = \mu u(1 - u) - \chi \nabla \cdot (u \nabla c) \quad \text{in } \Omega \times (0, T], \quad (1)$$

$$c_t = -pc \quad \text{in } \Omega \times (0, T], \quad (2)$$

$$p_t = \epsilon^{-1}(uc - p) \quad \text{in } \Omega \times (0, T], \quad (3)$$

where  $\Omega$  is a bounded domain in  $\mathbb{R}^d$ ,  $d \in \{1, 2, 3\}$ ,  $[0, T]$  is a time interval, and  $\mu$ ,  $\chi$ ,  $\epsilon$  are positive constants. The functions  $u$ ,  $c$ , and  $p$  depend on  $x \in \Omega$  and  $t \in [0, T]$  and represent concentrations of the invasive cells, extracellular matrix, and protease, respectively.

Let us explain the background of the model and the ideas considered in [7] to derive it. The extracellular matrix is the non-cellular component present within a connective tissue. It contains collagen in the form of interlacing protein fibers. The extracellular matrix is invaded by tumour cells which produce proteases. Proteases are enzymes that hydrolyze proteins, i.e., they help to digest the surrounding extracellular matrix. The production of proteases is tightly confined to the interface between the invading tumour and the receding connective tissue. Therefore, no protease diffusion is included in the model. This is a difference to many other models which complicates the theoretical investigations. The protease production is modelled by the term  $\epsilon^{-1}uc$ , whereas  $-\epsilon^{-1}p$  describes the natural decay of the protease. The decrease of the extracellular matrix  $c$  is modelled as a simple passive degradation by the activity of the tissue proteases. It is described by the term  $-pc$  since it depends on the amount of collagen still present as well as the protease  $p$ . Finally, in the equation for the concentration  $u$  of the invasive cells, the first term on the right-hand side describes their proliferation whereas the second one describes the spatial movement of the cells due to the gradient of the extracellular matrix. Since the flux of the cancer

cells is proportional to the gradient of the non-diffusible concentration  $c$ , the corresponding effect is called haptotaxis. A similar effect is chemotaxis which occurs in response to a diffusible substrate. Note also that the phenomenon in which a gradient in the concentration of one quantity induces a flux of another quantity is called cross-diffusion. Therefore, the haptotaxis term in (1) describes the cross-diffusion mentioned in the title of this paper.

It is argued in [7] that diffusion of the cancer cell concentration  $u$  can be neglected since the chemokinetic movement was reported to be minimal. Nevertheless, although this diffusion is negligible in numerical simulations, it is important for studying analytical properties of the model. Therefore, we will also consider the case when (1) is replaced by the equation

$$u_t = \mu u(1 - u) - \chi \nabla \cdot (u \nabla c) + \alpha^{-1} \Delta u \quad \text{in } \Omega \times (0, T] \quad (4)$$

with a positive constant  $\alpha$ . Thus, the equation (1) corresponds to the limit case of (4) for  $\alpha \rightarrow \infty$ .

When using (4) instead of (1) in the above model, an appropriate boundary condition leading to a solvable problem (see Section 3) reads

$$\alpha^{-1} u_n = \chi u c_n \quad \text{on } \partial\Omega \times [0, T], \quad (5)$$

where the index  $n$  indicates the derivative in the direction of the outward normal vector on  $\partial\Omega$ . For  $\alpha \rightarrow \infty$  this boundary condition reduces to

$$u c_n = 0 \quad \text{on } \partial\Omega \times [0, T], \quad (6)$$

which will be used for the model (1)–(3). Finally, let us mention that both models are endowed with the initial conditions

$$u(\cdot, 0) = u^0, \quad c(\cdot, 0) = c^0, \quad p(\cdot, 0) = p^0 \quad \text{in } \Omega, \quad (7)$$

where  $u^0, c^0, p^0 : \Omega \rightarrow [0, 1]$  are given functions.

### 3 Global Existence of Classical Solutions

In this section, we discuss the classical solvability of the models introduced above. We start with the system consisting of (4), (2), (3), (5), and (7). In order to construct global classical solutions, it is necessary to control the haptotaxis term  $-\chi \nabla \cdot (u \nabla c)$  in (4) and thus, in particular, to gain information on the spatial derivatives of  $c$ . For similar models including a diffusion term  $\Delta p$  in (3), this has already been achieved in earlier works by applying parabolic regularity theory to (3), first yielding estimates for the spatial derivatives of  $p$  and then also of  $c$ . However, the absence of any spatial regularization in both (2) and (3) makes the corresponding analysis much more difficult.

In our recent paper [2], the lack of regularity estimates was circumvented by transforming the problem under consideration into the equivalent form

$$w_t = (\alpha p c + \mu - \mu e^{\alpha c} w) w + \nabla c \cdot \nabla w + \alpha^{-1} \Delta w \quad \text{in } \Omega \times (0, T], \quad (8)$$

$$c_t = -p c \quad \text{in } \Omega \times (0, T], \quad (9)$$

$$p_t = w e^{\alpha c} c - p \quad \text{in } \Omega \times (0, T], \quad (10)$$

$$w_n = 0 \quad \text{on } \partial\Omega \times [0, T], \quad (11)$$

$$w(\cdot, 0) = w^0, \quad c(\cdot, 0) = c^0, \quad p(\cdot, 0) = p^0 \quad \text{in } \Omega, \quad (12)$$

where

$$w(x, t) = u(x, t) e^{-\alpha c(x, t)}, \quad x \in \overline{\Omega}, \quad t \geq 0. \quad (13)$$

Moreover, to obtain (8)–(12), we set  $\chi = \epsilon = 1$ , which can be done without loss of generality since the general case follows by a simple rescaling of  $x$ ,  $t$ ,  $c$ , and  $p$  (see [2]). To get rid of the nonlinear dependence on  $w$  and to decouple the equations for  $c$  and  $p$  from the first equation, the problem

$$v_t = (\alpha p c + \mu - \mu e^{\alpha c} w) v + \nabla c \cdot \nabla v + \alpha^{-1} \Delta v \quad \text{in } \Omega \times (0, T), \quad (14)$$

$$c_t = -p c \quad \text{in } \Omega \times (0, T), \quad (15)$$

$$p_t = w e^{\alpha c} c - p \quad \text{in } \Omega \times (0, T), \quad (16)$$

$$v_n = 0 \quad \text{on } \partial\Omega \times (0, T), \quad (17)$$

$$v(\cdot, 0) = w^0, \quad c(\cdot, 0) = c^0, \quad p(\cdot, 0) = p^0 \quad \text{in } \Omega \quad (18)$$

is considered for any fixed  $w \in C^0([0, T]; C^1(\overline{\Omega}))$ . Note that, for (14)–(18), the solvability and regularity are easier to investigate than for (8)–(12).

For  $T > 0$  and  $M > 0$ , we introduce the set

$$S_{M, T} = \left\{ w \in C^0([0, T]; C^1(\overline{\Omega})) ; 0 \leq w, \sup_{t \in [0, T]} \|w(\cdot, t)\|_{C^1(\overline{\Omega})} \leq M \right\}$$

and we introduce a mapping  $\Phi$  such that  $\Phi(w)$  is the unique function  $v$  satisfying the above problem for a given  $w \in S_{M, T}$ . If  $M$  is sufficiently large and  $T$  sufficiently small, then  $\Phi$  maps  $S_{M, T}$  to itself and has a compact image. Thus,  $\Phi$  has a fixed point due to Schauder's fixed point theorem, which implies the solvability (and regularity) of the problem (8)–(12). Finally, a priori estimates allow to extend the local solution onto the interval  $(0, \infty)$ , see [2] for details. This leads to the following result.

**Theorem 1** *Suppose that  $\alpha, \chi, \mu, \epsilon$  are positive constants, that  $\Omega$  is a smooth bounded domain in  $\mathbb{R}^d$ ,  $d \in \{1, 2, 3\}$ , and that  $u^0, c^0, p^0 \in \bigcup_{\gamma \in (0, 1)} C^{2, \gamma}(\overline{\Omega})$  are nonnegative functions satisfying  $\alpha^{-1} u_n^0 = \chi u^0 c_n^0$  on  $\partial\Omega$ . Then, for any  $T > 0$ , there exists a unique global classical solution  $(u, c, p)$  of (4), (2), (3), (5), and (7) with regularity  $(u, c, p) \in (C^{2, 1}(\overline{\Omega} \times (0, T]) \cap C^1(\overline{\Omega} \times [0, T]))^3$ , which, moreover, is nonnegative.*

Obviously, when the problem consisting of (1)–(3), (6), and (7) is considered, i.e., if no diffusion term is present, then the technique presented above cannot be applied since the transformation (13) does not make sense in the limit case for  $\alpha \rightarrow \infty$ . In this case, the solvability of the model is an open problem.

#### 4 A Positivity Preserving Discretization

Denoting by  $(\cdot, \cdot)$  the inner product in  $L^2(\Omega)$  or  $L^2(\Omega)^d$ , the solution of our model (1)–(3), (6), (7) satisfies

$$(u_t, v) = \mu (u(1-u), v) + \chi (u \nabla c, \nabla v) \quad \text{in } (0, T] \text{ and for } v \in H^1(\Omega) \quad (19)$$

and

$$c(x, t) = c^0(x) e^{-\int_0^t p(x, s) ds}, \quad (20)$$

$$p(x, t) = e^{-t/\epsilon} \left[ p^0(x) + \frac{1}{\epsilon} \int_0^t u(x, s) c(x, s) e^{s/\epsilon} ds \right] \quad (21)$$

for any  $(x, t) \in \Omega \times [0, T]$ . These relations will be used to define an approximate solution of our model. First, we introduce a triangulation  $\mathcal{T}_h$  of  $\Omega$  consisting of simplicial (for  $d = 1, 2, 3$ ), quadrilateral (for  $d = 2$ ) or hexahedral (for  $d = 3$ ) shape-regular cells possessing the usual compatibility properties. For any cell  $K \in \mathcal{T}_h$ , we denote by  $h_K$  the diameter of  $K$  and assume that  $h_K \leq h$ . We denote by  $V_h \subset H^1(\Omega)$  the usual conforming  $P_1$  or  $Q_1$  finite element space constructed using the triangulation  $\mathcal{T}_h$ . Let  $\phi_1, \dots, \phi_M$  be the standard basis functions of  $V_h$  associated with the vertices  $x_1, \dots, x_M$  of  $\mathcal{T}_h$ . Next, the time interval  $[0, T]$  is decomposed by  $0 = t_0 < t_1 < \dots < t_N = T$  and we set  $\tau_n = t_n - t_{n-1}$ ,  $n = 1, \dots, N$ . At each time level  $t_n$ , the solution of our model will be approximated by functions  $u_h^n, c_h^n, p_h^n \in V_h$ . These functions can be identified with coefficient vectors  $\mathbf{u}^n = (u_i^n)_{i=1}^M$ ,  $\mathbf{c}^n = (c_i^n)_{i=1}^M$ ,  $\mathbf{p}^n = (p_i^n)_{i=1}^M$ , respectively, satisfying  $u_i^n = u_h^n(x_i)$ ,  $c_i^n = c_h^n(x_i)$ , and  $p_i^n = p_h^n(x_i)$  for  $i = 1, \dots, M$ . We set  $u_i^0 = u^0(x_i)$ ,  $c_i^0 = c^0(x_i)$ , and  $p_i^0 = p^0(x_i)$  for  $i = 1, \dots, M$ .

Replacing the space  $H^1(\Omega)$  in (19) by  $V_h$  and applying the  $\theta$ -method for discretization in time (with  $\theta \in [0, 1]$ ), one obtains a discrete variational problem which can be equivalently written in the matrix form

$$(\mathbb{M} + \theta \tau_{n+1} \mathbb{A}^{n+1}) \mathbf{u}^{n+1} = (\mathbb{M} - (1 - \theta) \tau_{n+1} \mathbb{A}^n) \mathbf{u}^n, \quad n = 0, \dots, N-1, \quad (22)$$

where the matrices  $\mathbb{M} = (m_{ij})_{i,j=1}^M$  and  $\mathbb{A}^n = (a_{ij}^n)_{i,j=1}^M$  are defined by

$$m_{ij} = (\phi_j, \phi_i), \quad a_{ij}^n = -\mu (\phi_j(1 - u_h^n), \phi_i) - \chi (\phi_j \nabla c_h^n, \nabla \phi_i).$$

The relation (20) suggests to define the coefficients of  $c_h^n$  by

$$c_i^n = c^0(x_i) e^{-\int_0^{t_n} p_{h,\tau}(x_i, s) ds}, \quad i = 1, \dots, M, \quad n = 0, \dots, N, \quad (23)$$

where  $p_{h,\tau}(\cdot, t_n) = p_h^n$  for all  $n = 0, \dots, N$  and  $p_{h,\tau}(x, \cdot)$  is piecewise linear with respect to the decomposition of  $[0, T]$  for any  $x \in \bar{\Omega}$ . Similarly, (21) leads to

$$p_i^n = e^{-t_n/\epsilon} \left[ p^0(x_i) + \frac{1}{\epsilon} \int_0^{t_n} u_{h,\tau}(x_i, s) c_{h,\tau}(x_i, s) e^{s/\epsilon} ds \right]. \quad (24)$$

A direct computation gives

$$\begin{aligned} c_i^{n+1} &= c_i^n e^{-\tau_{n+1} (p_i^{n+1} + p_i^n)/2}, \quad (25) \\ p_i^{n+1} &= e^{-\tau_{n+1}/\epsilon} p_i^n + \frac{1}{\tau_{n+1}^2} \left\{ \left( u_i^{n+1} (\epsilon - \tau_{n+1}) - u_i^n \epsilon \right) \left( c_i^{n+1} (\epsilon - \tau_{n+1}) - c_i^n \epsilon \right) \right. \\ &\quad - \left( u_i^{n+1} \epsilon - u_i^n (\epsilon + \tau_{n+1}) \right) \left( c_i^{n+1} \epsilon - c_i^n (\epsilon + \tau_{n+1}) \right) e^{-\tau_{n+1}/\epsilon} \\ &\quad \left. + (u_i^{n+1} - u_i^n) (c_i^{n+1} - c_i^n) \epsilon^2 \left( 1 - e^{-\tau_{n+1}/\epsilon} \right) \right\}. \quad (26) \end{aligned}$$

In general, the discretization formulated above does not provide nonnegative solutions. To guarantee the positivity preservation property, we will modify the Galerkin discretization (22) using the diagonal lumped mass matrix  $\mathbb{M}_L = \text{diag}(m_1, \dots, m_M)$  and the symmetric artificial diffusion matrix  $\mathbb{D}^n = (d_{ij}^n)_{i,j=1}^M$  defined by

$$m_i = \sum_{j=1}^M m_{ij}, \quad d_{ij}^n = -\max\{a_{ij}^n, 0, a_{ji}^n\} \quad \text{for } i \neq j, \quad d_{ii}^n = -\sum_{j=1, j \neq i}^M d_{ij}^n,$$

and we set  $\mathbb{L}^n = \mathbb{A}^n + \mathbb{D}^n$ . Then the simplest way to enforce the positivity preservation is to replace (22) by

$$(\mathbb{M}_L + \theta \tau_{n+1} \mathbb{L}^{n+1}) \mathbf{u}^{n+1} = (\mathbb{M}_L - (1 - \theta) \tau_{n+1} \mathbb{L}^n) \mathbf{u}^n, \quad n = 0, \dots, N-1. \quad (27)$$

Note that the matrix  $\mathbb{L}^{n+1}$  depends on  $\mathbf{u}^{n+1}$  and  $\mathbf{c}^{n+1}$  so that (27) is again nonlinear. In contrast to the Galerkin discretization (22), it is now possible to assure the positivity preservation for sufficiently small time steps (i.e.,  $\mathbf{u}^n \geq 0$  implies  $\mathbf{u}^{n+1} \geq 0$ , see [3]). Then (23) and (24) imply that also  $c_h^n$  and  $p_h^n$  will be nonnegative.

The solution of (25)–(27) is usually inaccurate since too much artificial diffusion is introduced by the modifications leading to (27). To limit the amount of the artificial diffusion, we will apply the FCT approach following [6]. First, we note that the Galerkin discretization (27) can be written in the form

$$(\mathbb{M}_L + \theta \tau_{n+1} \mathbb{L}^{n+1}) \mathbf{u}^{n+1} = (\mathbb{M}_L - (1 - \theta) \tau_{n+1} \mathbb{L}^n) \mathbf{u}^n + \left( \sum_{j=1}^M f_{ij}^{n+1} \right)_{i=1}^M$$



with algebraic fluxes  $f_{ij}^{n+1}$  given by

$$f_{ij}^{n+1} = -m_{ij}(u_j^{n+1} - u_i^{n+1}) + m_{ij}(u_j^n - u_i^n) \\ + \theta \tau_{n+1} d_{ij}^{n+1}(u_j^{n+1} - u_i^{n+1}) + (1 - \theta) \tau_{n+1} d_{ij}^n(u_j^n - u_i^n).$$

Now, the idea of the FCT approach is to limit the fluxes  $f_{ij}^{n+1}$  by solution dependent correction factors  $\alpha_{ij}^{n+1} \in [0, 1]$  called limiters so that the nonnegativity of the approximate solution can be guaranteed but less artificial diffusion is introduced than in case of (27). This leads to the discrete problem

$$(\mathbb{M}_L + \theta \tau_{n+1} \mathbb{L}^{n+1}) \mathbf{u}^{n+1} = (\mathbb{M}_L - (1 - \theta) \tau_{n+1} \mathbb{L}^n) \mathbf{u}^n + \left( \sum_{j=1}^M \alpha_{ij}^{n+1} f_{ij}^{n+1} \right)_{i=1}^M. \quad (28)$$

The correction factors have to depend on the fluxes  $f_{ij}^{n+1}$ , which introduces an additional nonlinearity. Since the underlying problem is already nonlinear, both nonlinearities can be treated simultaneously.

A popular choice is the limiter proposed by Zalesak [9]. Given algebraic fluxes  $f_{ij}$  and  $\bar{\mathbf{u}} \in \mathbb{R}^M$ , the Zalesak algorithm defines corrections factors  $\alpha_{ij}$  in such a way that the components of  $\tilde{\mathbf{u}} \in \mathbb{R}^M$  solving the problem

$$\mathbb{M}_L \tilde{\mathbf{u}} = \mathbb{M}_L \bar{\mathbf{u}} + \left( \sum_{j=1}^M \alpha_{ij} f_{ij} \right)_{i=1}^M$$

are bounded from below (above) by local minima (maxima) of  $\bar{\mathbf{u}}$ . Thus, in particular, one gets  $\tilde{\mathbf{u}} \geq 0$  if  $\bar{\mathbf{u}} \geq 0$ . We refer to [3] for details. We have the following result.

**Theorem 2** Consider any  $n \in \{0, \dots, N-1\}$  and let  $\mathbf{u}^n, \mathbf{c}^n, \mathbf{p}^n \in \mathbb{R}^M$  satisfy  $\mathbf{u}^n \geq 0$ ,  $1 \geq \mathbf{c}^n \geq 0$ ,  $\mathbf{p}^n \geq 0$ . Let the time step  $\tau_{n+1}$  satisfy the conditions

$$(1 - \theta) \tau_{n+1} l_{ii}^n \leq m_i, \quad \theta \tau_{n+1} \left( \mu m_i + \chi n_v \kappa^2 \sum_{K \ni x_i} h_K^{d-2} \right) < m_i, \quad i = 1, \dots, M,$$

where  $(l_{ii}^n)_{i=1}^M$  is the diagonal of  $\mathbb{L}^n$ ,  $n_v$  is the number of vertices of a cell in  $\mathcal{T}_h$ , and  $\kappa$  is a constant satisfying  $\|\nabla \phi_i\|_{L^2(K)} \leq \kappa h_K^{d/2-1}$  for any  $K \in \mathcal{T}_h$  and  $i = 1, \dots, M$ . Then there exist vectors  $\mathbf{u}^{n+1}, \mathbf{c}^{n+1}, \mathbf{p}^{n+1} \in \mathbb{R}^M$  satisfying (28), (25), (26) where the limiters  $\alpha_{ij}^{n+1}$  are computed using the Zalesak algorithm from the fluxes  $f_{ij}^{n+1}$ . Moreover, these vectors satisfy  $\mathbf{u}^{n+1} \geq 0$ ,  $1 \geq \mathbf{c}^{n+1} \geq 0$ , and  $\mathbf{p}^{n+1} \geq 0$ .

*Proof.* We mention only the main ideas of the proof and refer to [3] for details. To guarantee that  $\mathbf{c}^{n+1} \leq 1$  and to prove coercivity, we first change the definitions of  $c_i^{n+1}$  and  $a_{ij}^{n+1}$  by introducing absolute values of  $p_i^{n+1}$  and  $u_h^{n+1}$ :

$$c_i^{n+1} = c_i^n e^{-\tau_{n+1} (|p_i^{n+1}| + p_i^n)/2}, \quad a_{ij}^{n+1} = -\mu (\phi_j(1 - |u_h^{n+1}|), \phi_i) - \chi (\phi_j \nabla c_h^{n+1}, \nabla \phi_i).$$

Substituting  $\mathbf{c}^{n+1}$  into (26) and (28), one can introduce an operator  $P : \mathbb{R}^{2M} \rightarrow \mathbb{R}^{2M}$  such that  $\mathbf{U} = (\mathbf{u}^{n+1}, \mathbf{p}^{n+1})$  solves the operator equation  $P\mathbf{U} = 0$ . This operator is continuous and satisfies  $(P\mathbf{U}, \mathbf{U}) \geq \|\mathbf{U}\|^2 - C$  for any  $\mathbf{U} \in \mathbb{R}^{2M}$ , where  $C$  is a positive constant and  $(\cdot, \cdot)$  and  $\|\cdot\|$  are the Euclidean inner product and norm in  $\mathbb{R}^{2M}$ , respectively. Therefore, due to Brouwer's fixed-point theorem, the equation  $P\mathbf{U} = 0$  possesses a solution (cf. [8], p. 164, Lemma 1.4). The assumed time step restrictions and the properties of the Zalesak limiter imply that the solution is nonnegative. Therefore, all results hold also for the original system (28), (25), (26).  $\square$

Note that the use of the discretizations (23)–(24) based on the analytical expressions (20)–(21) was essential for proving the above theoretical results. Discretizing (2)–(3) in a variational form like in [2], we were not able to show the nonnegativity of the approximation of  $p$ .

## 5 Numerical Results

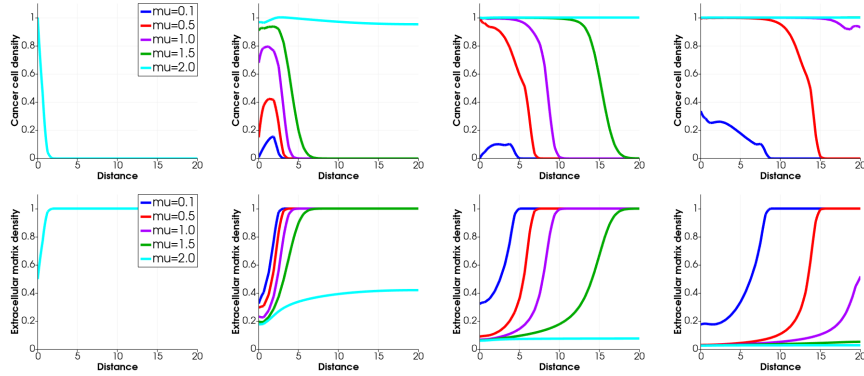
In this section, we present a few numerical results obtained for the discrete problem (28), (25), (26) with  $\theta = 0.5$  (Crank-Nicolson method). This nonlinear problem is solved by fixed-point iterations in the following way. Having an approximation  $\mathbf{u}_k^{n+1}$ ,  $\mathbf{c}_k^{n+1}$ ,  $\mathbf{p}_k^{n+1}$  of  $\mathbf{u}^{n+1}$ ,  $\mathbf{c}^{n+1}$ ,  $\mathbf{p}^{n+1}$  (using  $\mathbf{u}^n$ ,  $\mathbf{c}^n$ ,  $\mathbf{p}^n$  as the initial guess), we compute a new approximation  $\mathbf{c}_{k+1}^{n+1}$  from (25), then  $\mathbf{p}_{k+1}^{n+1}$  from (26), and finally  $\mathbf{u}_{k+1}^{n+1}$  from (28) where  $\mathbb{L}^{n+1}$  and  $f_{ij}^{n+1}$  are computed using  $\mathbf{u}_k^{n+1}$  instead of  $\mathbf{u}^{n+1}$ . Moreover, a damping is used to improve the convergence behaviour. Under the time step restrictions from Theorem 2, it can be proved that all the iterates are uniquely determined and satisfy  $\mathbf{u}_k^{n+1} \geq 0$ ,  $1 \geq \mathbf{c}_k^{n+1} \geq 0$ , and  $\mathbf{p}_k^{n+1} \geq 0$ , see [3]. This is important since, in practice, the fixed-point iterations are usually terminated when a stopping criterion is met, i.e., typically before reaching the solution of the nonlinear problem (28), (25), (26). The described algorithm is implemented in the open-source finite element library deal.II and the linear systems are solved using the sparse direct solver UMFPACK.

The computations were performed on a uniform quadrilateral mesh of the square domain  $\Omega = (0, 20)^2$  and conforming bilinear finite elements were used for approximating all unknown variables. The initial conditions were defined by

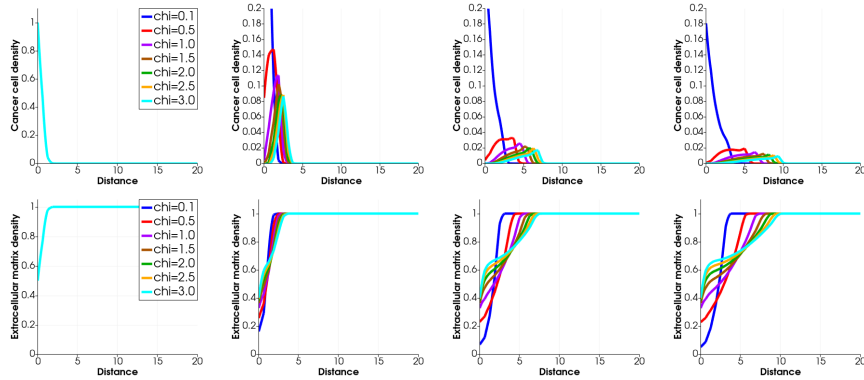
$$u^0(x) = e^{-|x|^2}, \quad c^0(x) = 1 - \frac{1}{2} e^{-|x|^2}, \quad p^0(x) = \frac{1}{2} e^{-|x|^2}$$

and the parameter  $\epsilon = 0.2$  was used. We present the numerical results along the diagonal of the square  $\Omega$  connecting the origin with the point  $(20, 20)$ . Note that more complicated domains  $\Omega$  could be used without any additional difficulties.

Fig. 1 shows the influence of different values of the proliferation rate  $\mu$  on the densities of cancer cells  $u$  and of extracellular matrix (ECM)  $c$  for a fixed haptotactic rate  $\chi = 1.0$ . For a small growth rate  $\mu = 0.1$ , a small cluster of cancer cells is created and migrates over the domain during the time. However, for higher proliferation rates the density of the cancer cells increases considerably and invades the ECM domain



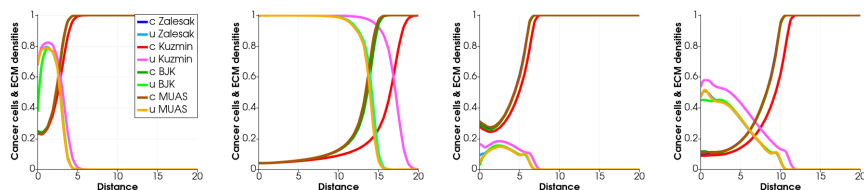
**Fig. 1** Approximations of  $u$  (top) and  $c$  (bottom) for  $\chi = 1.0$  and  $\mu = 0.1, 0.5, 1.0, 1.5, 2.0$  at time instants  $t = 0, t = 5, t = 15$ , and  $t = 35$  (left to right)



**Fig. 2** Approximations of  $u$  (top) and  $c$  (bottom) for  $\mu = 0.001$  and  $\chi = 0.1, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0$  at time instants  $t = 0, t = 5, t = 25$ , and  $t = 45$  (left to right)

more rapidly. Next, we investigate the influence of the haptotactic rate on the cancer cells invasion by varying  $\chi$  and set  $\mu = 0.001$  to be fixed, see Fig. 2. We observe that a small cluster is created at the beginning and migrates through the domain during the time. It can be seen that the migration accelerates by increasing  $\chi$  and degrades the ECM more rapidly.

In Fig. 3, we investigate the influence of the choice of the limiters in (28) on the computed approximations of the solution for parameter choices  $\mu = \chi = 1.0$  and  $\mu = 0.1, \chi = 1.0$ . Apart from the Zalesak limiter [9] considered in the previous section, we use the Kuzmin limiter [5], BJK limiter [1], and MUAS limiter [4]. We observe that, for the Kuzmin limiter, the approximate solutions evolve in time faster than for the other three choices. The results for the Zalesak and MUAS limiters almost coincide whereas, for the BJK limiter, we observe a slightly faster time evolution in the case  $\mu = \chi = 1.0$ .



**Fig. 3** Approximations of  $c$  and  $u$  obtained using the Zalesak, Kuzmin, BJK, and MUAS limiters for  $\mu = \chi = 1.0$  (first two plots) at  $t = 5$  and  $t = 25$  and for  $\mu = 0.1, \chi = 1.0$  (last two plots) at  $t = 25$  and  $t = 45$

It is important to stress that all the presented numerical results clearly demonstrate that the considered methods are positivity preserving as predicted by our theory. Moreover, the concentrations are also bounded by 1 from above.

**Acknowledgements** The work of Petr Knobloch was supported through the grant No. 22-01591S of the Czech Science Foundation. The work of Shahin Heydari was supported through the grant SVV-2023-260711 of the Charles University.

**Competing Interests** The authors have no conflicts of interest to declare that are relevant to the content of this chapter.

## References

1. Barrenechea, G.R., John, V., Knobloch, P.: An algebraic flux correction scheme satisfying the discrete maximum principle and linearity preservation on general meshes. *Math. Models Methods Appl. Sci.* **27**, 525–548 (2017)
2. Fuest, M., Heydari, Sh., Knobloch, P., Lankeit, J., Wick, T.: Global existence of classical solutions and numerical simulations of a cancer invasion model. *ESAIM Math. Model. Numer. Anal.* **57**, 1893–1919 (2023)
3. Heydari, Sh., Knobloch, P., Wick, T.: Flux-corrected transport stabilization of an evolutionary cross-diffusion cancer invasion model. *J. Comput. Phys.* **499**, Art. No. 112711 (2024)
4. John, V., Knobloch, P.: On algebraically stabilized schemes for convection–diffusion–reaction problems. *Numer. Math.* **152**, 553–585 (2022)
5. Kuzmin, D.: Algebraic flux correction for finite element discretizations of coupled systems. In: Papadrakakis, M., Oñate, E., Schrefler, B. (eds.) *Computational Methods for Coupled Problems in Science and Engineering II*, pp. 653–656. CIMNE, Barcelona (2007)
6. Kuzmin, D.: Algebraic flux correction I. Scalar conservation laws. In: Kuzmin, D., Löhner, R., Turek, S. (eds.) *Flux-Corrected Transport. Principles, Algorithms, and Applications*. 2nd edn., pp. 145–192. Springer, Dordrecht (2012)
7. Perumpanani, A.J., Sherratt, J.A., Norbury, J., Byrne, H.M.: A two parameter family of travelling waves with a singular barrier arising from the modelling of extracellular matrix mediated cellular invasion. *Physica D* **126**, 145–159 (1999)
8. Temam, R.: *Navier-Stokes Equations. Theory and Numerical Analysis*. North-Holland, Amsterdam (1977)
9. Zalesak, S.T.: Fully multidimensional flux-corrected transport algorithms for fluids. *J. Comput. Phys.* **31**, 335–362 (1979)

# 4. Paper IV

This chapter is based on the paper entitled "A cross-diffusion system modeling rivaling gangs: global existence of bounded solutions and FCT stabilization for numerical simulation", published in *Mathematical Models and Methods in Applied Sciences*, DOI: 10.1142/S0218202524500349.

## 4.1 A cross-diffusion system modeling rivaling gangs: global existence of bounded solutions and FCT stabilization for numerical simulation

In this paper, we considered a system consisting of two parabolic and two ordinary differential equations describing rivaling gangs interaction [243], as

$$\begin{cases} u_t = D_u \Delta u + \chi_u \nabla \cdot (u \nabla w) & \text{in } \Omega \times (0, \infty), \\ v_t = D_v \Delta v + \chi_v \nabla \cdot (v \nabla z) & \text{in } \Omega \times (0, \infty), \\ w_t = -w + f(v) & \text{in } \Omega \times (0, \infty), \\ z_t = -z + g(u) & \text{in } \Omega \times (0, \infty), \\ D_u \partial_\nu u + \chi_u u \partial_\nu w = D_v \partial_\nu v + \chi_v v \partial_\nu z = 0 & \text{on } \partial\Omega \times (0, \infty), \\ (u, v, w, z)(\cdot, 0) = (u_0, v_0, w_0, z_0) & \text{in } \Omega, \end{cases} \quad (4.1)$$

where  $D_u, D_v, \chi_u, \chi_v$  are positive parameters,  $u_0, v_0, w_0, z_0$  are smooth initial data,  $f, g$  are given sprays rate, and  $u, v$  are the densities of two rivaling gangs which mark their territory by spraying graffiti with densities  $z$  and  $w$ , respectively. Let us summarize which modeling consideration leads to which terms in the system:

- The gangs move around randomly: terms  $D_u \Delta u$  and  $D_v \Delta v$ ,
- The gangs move away from high hostile graffiti concentrations: terms  $+\chi_u \nabla \cdot (u \nabla w)$ ,  $+\chi_v \nabla \cdot (v \nabla z)$ ; the signs of  $+\chi_u$  and  $+\chi_v$  indicate that the gangs are indeed repelled and not attracted by the graffiti,
- The gangs stay in the domain: no-flux boundary conditions  $D_u \partial_\nu u + \chi_u u \partial_\nu w = D_v \partial_\nu v + \chi_v v \partial_\nu z = 0$  on  $\partial\Omega \times (0, \infty)$ ,
- Gang members spray their graffiti at their current locations: terms  $+f(v)$  and  $+g(u)$ ,
- The total amount of members of each gang remains constant throughout the time: *absence* of any zeroth order or external force terms in the first two equations,
- The graffiti decay over time: terms  $-w$  and  $-z$ ,
- The graffiti do not diffuse, i.e., they are immobile: *absence* of terms  $D_w \Delta w$  and  $D_z \Delta z$  for positive  $D_w, D_z$  in the third and fourth equation.

From analytical point of view, we show that under the requirement

$$f, g \in C^1([0, \infty)) \cap L^\infty((0, \infty)) \quad \text{are nonnegative,} \quad (4.2)$$

and for any reasonable smooth initial data, there is a unique, global, non-negative classical solution of (4.1).

**Theorem.** *Let*

$$\Omega \subset \mathbb{R}^n, n \in \{1, 2, 3\}, \text{ be a smooth, bounded domain,} \quad (4.3)$$

$D_u, D_v, \chi_u, \chi_v > 0$ ,  $f, g$  be as in (4.2),  $\alpha \in (0, 1)$  and  $w_0, z_0 \in C^{2+\alpha}(\bar{\Omega}; [0, \infty))$ . Then there exists  $C > 0$  such that for all  $M > 0$  and all nonnegative  $u_0, v_0 \in C^{2+\alpha}(\bar{\Omega})$  with

$$\begin{cases} D_u \partial_\nu u_0 + \chi_u u_0 \partial_\nu w_0 = D_v \partial_\nu v_0 + \chi_v v_0 \partial_\nu z_0 = 0 \text{ on } \partial\Omega \text{ and} \\ \|u_0\|_\infty + \|v_0\|_\infty \leq M, \end{cases} \quad (4.4)$$


there exists a unique, global, nonnegative classical solution  $(u, v, w, z)$  of (4.1), which satisfies the estimates

$$\begin{cases} \|u(\cdot, t)\|_\infty + \|v(\cdot, t)\|_\infty \leq CM \text{ and} \\ \|w(\cdot, t)\|_\infty + \|z(\cdot, t)\|_\infty \leq C, \end{cases} \quad (4.5)$$

for all  $t \geq 0$ .

After having established global existence of a solution to (4.1), a natural next question was whether the first two components of these solutions separate. Answering this question analytically appeared to be a very difficult task, we showed that for small initial data the solutions converge towards homogeneous equilibria, however, for large data it seemed to be very challenging to show any results. Hence, in the next step, we used a numerical scheme to obtain the solution of (4.1), this allowed us to address the asymptotic behavior of large-time solutions and also to illustrate the evolution of gang densities throughout the time for various parameters (mainly diffusion- and convection-dominated regime). To this end, we employed a high-resolution nonlinear finite element flux-corrected transport method altogether with  $\theta$ -method for time discretization and fixed-point iteration to treat the nonlinearities in the proposed scheme. We next proved that under CFL-like conditions, the resulting method is positivity-preserving, we also showed that if certain assumptions on the matrices from the algebraic system hold, the resulting scheme satisfies the DMP. For the last part, we performed several numerical experiments and showed that for small values of  $\chi$  (diffusion-dominated regime), gang populations stay completely mixed and the approximated solutions converge toward constant steady states, however, both partial and complete separation is observed for large values of  $\chi$  (convection-dominated regime).

## A cross-diffusion system modeling rivaling gangs: Global existence of bounded solutions and FCT stabilization for numerical simulation

Mario Fuest \*

*Institut für Angewandte Mathematik, Leibniz Universität Hannover,  
Welfengarten 1, 30167 Hannover, Germany  
fuest@ifam.uni-hannover.de*

Shahin Heydari 

*Faculty of Mathematics and Physics, Charles University,  
Sokolovska 83, 18675 Praha 8, Czech Republic  
heydari@karlin.mff.cuni.cz*

Received 13 December 2023

Revised 5 April 2024

Accepted 27 April 2024

Published 29 June 2024

Communicated by N. Bellomo

In this paper, we study a gang territorial model consisting of two parabolic and two ordinary differential equations, where a taxis-type mechanism models that the two rivaling gangs are repelled by each other's graffiti. Our main analytical finding shows the existence of global, bounded classical solutions. By making use of quantitative global estimates, we prove that these solutions converge to homogeneous steady states if the initial data are sufficiently small.

Moreover, we perform numerical experiments which show that for different choices of parameters, the system may become diffusion- or convection-dominated, where in the former case the solutions converge toward constant steady states while in the latter case nontrivial asymptotic behavior such as segregation is observed. In order to perform these experiments, we apply a nonlinear finite element flux-corrected transport method (FEM-FCT) which is positivity-preserving. Then we treat the nonlinearities in both the system and the proposed nonlinear scheme simultaneously using fixed-point iteration.

*Keywords:* Gang territoriality; cross-diffusion; global existence; asymptotic behavior; separation; FEM-FCT stabilization; positivity preservation.

AMS Subject Classification 2020: 35K55, 35A01, 35B40, 35Q91, 65M22, 65M60, 91D10

\*Corresponding author.

## 1. Introduction

Graffiti, the artful wall writing usually on public property and in open view, is sprayed by urban gang members not only to express themselves, convey their attitudes and communicate with fellow gang members but also to mark their area of control, i.e. the gang's territory.<sup>14, 46</sup> In order to describe the interaction of two rivaling gangs attempting to establish or defend territories by spraying intimidating graffiti, Alsenafi and Barbaro<sup>2</sup> introduce the model

$$\begin{cases} u_t = D_u \Delta u + \chi_u \nabla \cdot (u \nabla w) & \text{in } \Omega \times (0, \infty), \\ v_t = D_v \Delta v + \chi_v \nabla \cdot (v \nabla z) & \text{in } \Omega \times (0, \infty), \\ w_t = -w + f(v) & \text{in } \Omega \times (0, \infty), \\ z_t = -z + g(u) & \text{in } \Omega \times (0, \infty), \\ D_u \partial_\nu u + \chi_u u \partial_\nu w = D_v \partial_\nu v + \chi_v v \partial_\nu z = 0 & \text{on } \partial\Omega \times (0, \infty), \\ (u, v, w, z)(\cdot, 0) = (u_0, v_0, w_0, z_0) & \text{in } \Omega, \end{cases} \quad (1.1)$$

where positive parameters  $D_u, D_v, \chi_u, \chi_v$ , suitably smooth initial data  $u_0, v_0, w_0, z_0$  and spray rates  $f, g$  (the choice  $f = g = \text{id}$  is proposed in Ref. 2) are given. Here,  $u$  and  $v$  denote the densities of two rivaling gangs which mark their territory by spraying graffiti with densities  $z$  and  $w$ , respectively. Let us summarize which modeling considerations lead to which terms in the system.

- The gangs move around randomly: terms  $D_u \Delta u$  and  $D_v \Delta v$ .
- The gangs move away from high hostile graffiti concentrations: terms  $+\chi_u \nabla \cdot (u \nabla w)$ ,  $+\chi_v \nabla \cdot (v \nabla z)$ ; the signs of  $+\chi_u$  and  $+\chi_v$  indicate that the gangs are indeed repelled and not attracted by the graffiti.
- The gangs stay in the domain: no-flux boundary conditions  $D_u \partial_\nu u + \chi_u u \partial_\nu w = D_v \partial_\nu v + \chi_v v \partial_\nu z = 0$  on  $\partial\Omega \times (0, \infty)$ .
- Gang members spray their graffiti at their current locations: terms  $+f(v)$  and  $+g(u)$ .
- The total amount of members of each gang remains constant throughout time: *absence* of any zeroth order or external force terms in the first two equations.
- The graffiti decay over time: terms  $-w$  and  $-z$ .
- The graffiti do not diffuse, i.e. they are immobile: *absence* of terms  $D_w \Delta w$  and  $D_z \Delta z$  for positive  $D_w, D_z$  in the third and fourth equation.

In fact, Alsenafi and Barbaro<sup>2</sup> first design a discrete agent-based model based on similar considerations (see Sec. 2 in Ref. 2) and then obtain (1.1) as the formal limit when both the time step and the grid spacing converge to zero (see Subsec. 3.2 in Ref. 2). For related modeling considerations, we refer to Ref. 8 for an overview of biological phenomena modeled by active particles, to Ref. 10 for a discussion of complex models incorporating a taxis term and to Ref. 9 for the mathematical



modeling of human crowds. Especially, the last two of these surveys examine how such models can be derived by a multiscale approach.

**Mathematically related systems.** In this paper, we study (1.1) both analytically and numerically. Before presenting our results, we compare (1.1) to related problems, namely double cross-diffusion and haptotaxis systems. Under the assumption that the graffiti densities equilibrate instantly (and that  $f = g = \text{id}$ ), Ref. 6 reduces (1.1) to the two-component system

$$\begin{cases} u_t = D_u \Delta u + \chi_u \nabla \cdot (u \nabla v), \\ v_t = D_v \Delta v + \chi_v \nabla \cdot (v \nabla u), \end{cases} \tag{1.2}$$

(with positive parameters), which can also be interpreted as a model for gangs *directly* repelling each other, and proves a weak-stability result as well as convergence of weak solutions (if they exist) to constant steady states under a smallness condition. A key difficulty in establishing even a local solution theory for (1.2) for large data consists of the nonpositive definiteness of the diffusion matrix  $\begin{pmatrix} D_u & \chi_u u \\ \chi_v v & D_v \end{pmatrix}$  whenever  $uv > \frac{D_u D_v}{\chi_u \chi_v}$ . Nonetheless, global existence results have recently been obtained which, however, either need to require a certain regularization<sup>20</sup> or can only guarantee solution properties within the parabolic regime.<sup>65</sup>

The problem that a diffusion matrix is not positive definite can be overcome in multiple ways, for instance by replacing linear diffusion with porous medium-type diffusion, e.g.  $\Delta u$  and  $\Delta v$  by  $\nabla \cdot (u \nabla u)$  and  $\nabla \cdot (v \nabla v)$ , respectively. Indeed, for such a system, global, locally bounded weak solutions exist as long as  $D_u D_v > \chi_u \chi_v$ , see Ref. 44 and also its precedent Ref. 45. Moreover, one can also consider (1.2) with  $\chi_u \chi_v < 0$ , which then models pursuit–evasion dynamics.<sup>62</sup> Again the diffusion matrix is positive definite as long as both components are non-negative but in contrast to the repulsion–repulsion problem with degenerate diffusion above, apparently there only exists a single quasi-energy functional and the *a priori* estimates thereby gained only suffice to construct global weak solutions in the one-dimensional setting<sup>60, 61</sup>; in the higher-dimensional case, global weak solutions are only known to exist if the diffusion is sufficiently enhanced or the taxis is saturated.<sup>28</sup> Furthermore, homogeneous steady states of (1.2) with  $\chi_u \chi_v < 0$  are asymptotically stable in the sense that global classical solutions emanating from nearby initial data (exist and) converge to these equilibria.<sup>27</sup>

Another way of regularizing (1.2) consists in replacing the cross-diffusive contributions with smoother functions; that is, in considering

$$\begin{cases} u_t = D_u \Delta u + \chi_u \nabla \cdot (u \nabla (K * v)), \\ v_t = D_v \Delta v + \chi_v \nabla \cdot (v \nabla (K * u)), \end{cases} \tag{1.3}$$

where  $K$  denotes a spatial averaging kernel, for instance. (If  $K$  is the Dirac delta distribution, one again obtains (1.2).) The system (1.3) can inter alia be used to describe territorial formations of various animals which remember direct

encounters<sup>52</sup>; for existence results we refer to Refs. 31, 38 and references therein. Moreover, for the case when  $K$  is Green's function for  $-\Delta + 1$  with Neumann boundary conditions, i.e. when  $K * \varphi$  is the solution  $\psi$  of the elliptic equation  $-\Delta\psi + \psi = \varphi$  in  $\Omega$  with  $\partial_\nu\psi = 0$  on  $\partial\Omega$ , and when  $\chi_u\chi_v < 0$ , global classical solutions of (1.3) are constructed in Ref. 47. For related systems where the signal equations are parabolic, see for instance Refs. 51 or 64.

In contrast, while the indirect mechanism in (1.1) entails for instance that space-time bounds for  $f(v)$  and  $g(u)$  imply uniform-in-time *a priori* estimates for  $w$  and  $z$ , the third and fourth equations in (1.1) do not regularize in space at all. Thus, mathematically, (1.1) is related to haptotaxis problem such as

$$\begin{cases} u_t = D_u\Delta u - \nabla \cdot (u\chi_u(v)\nabla v), \\ v_t = -uv, \end{cases}$$

studied for instance in Refs. 17 and 18. These systems share the challenge of controlling cross-diffusion terms involving spatial derivatives of the signal(s) without relying on spatial regularity gained due to diffusion terms in the signal equation(s).

**Main analytical results.** For our global existence result regarding (1.1), we need to require that

$$f, g \in C^1([0, \infty)) \cap L^\infty((0, \infty)) \quad \text{are non-negative.} \tag{1.4}$$

(However, Corollary 1.1 and Theorem 1.2 below also apply to unbounded  $f, g$  such as  $f = g = \text{id.}$ ) A prototypical choice is given by  $f(s) = g(s) = \frac{s}{1+s}$  for  $s \geq 0$ , which not only satisfies (1.4) but also guarantees that no graffiti come into existence out of nowhere by fulfilling  $f(0) = g(0) = 0$ . For this example, the amount of sprayed graffiti increases roughly proportionally to the corresponding gang density at that point as long as the latter is rather small but is then limited by some positive constant. Such a saturation effect appears to be reasonable: For large gang densities, the amount of additional wall writings in an area may be limited by available space rather than by the amount of gang members willing to spray graffiti.

For any such choice of graffiti production terms and any reasonably smooth initial data, we can construct globally bounded classical solutions of (1.1).

**Theorem 1.1.** *Let*

$$\Omega \subset \mathbb{R}^n, \quad n \in \{1, 2, 3\}, \quad \text{be a smooth, bounded domain,} \tag{1.5}$$

$D_u, D_v, \chi_u, \chi_v > 0$ ,  $f, g$  be as in (1.4),  $\alpha \in (0, 1)$  and  $w_0, z_0 \in C^{2+\alpha}(\overline{\Omega}; [0, \infty))$ . Then there exists  $C > 0$  such that for all  $M > 0$  and all non-negative  $u_0, v_0 \in C^{2+\alpha}(\overline{\Omega})$  with

$$\begin{cases} D_u\partial_\nu u_0 + \chi_u u_0 \partial_\nu w_0 = D_v\partial_\nu v_0 + \chi_v v_0 \partial_\nu z_0 = 0 & \text{on } \partial\Omega, \quad \text{and} \\ \|u_0\|_{L^\infty(\Omega)} + \|v_0\|_{L^\infty(\Omega)} \leq M, \end{cases} \tag{1.6}$$

there exists a unique, global, non-negative classical solution  $(u, v, w, z)$  of (1.1), which satisfies the estimates

$$\begin{cases} \|u(\cdot, t)\|_{L^\infty(\Omega)} + \|v(\cdot, t)\|_{L^\infty(\Omega)} \leq CM, & \text{and} \\ \|w(\cdot, t)\|_{L^\infty(\Omega)} + \|z(\cdot, t)\|_{L^\infty(\Omega)} \leq C, \end{cases} \quad (1.7)$$

for all  $t \geq 0$ .

A direct consequence of this theorem, especially of the bound (1.7), is the existence of global small data solutions of the original system proposed in Ref. 2.

**Corollary 1.1.** *Assume (1.5), let  $D_u, D_v, \chi_u, \chi_v > 0$ , let  $f(s) = g(s) = s$  for  $s \geq 0$ ,  $\alpha \in (0, 1)$  and let  $w_0, z_0 \in C^{2+\alpha}(\bar{\Omega}; [0, \infty))$ . Then there exists  $M > 0$  such that for all non-negative  $u_0, v_0 \in C^{2+\alpha}(\bar{\Omega})$  fulfilling (1.6), there exists a unique, global, bounded, non-negative classical solution  $(u, v, w, z)$  of (1.1).*

Having established global existence of solutions to (1.1), a natural next question is whether the first two components of these solutions separate. We first give a negative answer for small initial data and show that the solutions converge toward homogeneous equilibria. A corresponding result for the two-component system (1.2) (with  $\chi_1, \chi_2 > 0$ ) has already been observed for global weak solutions (whose existence, however, is not known yet) in Theorem 5.1 in Ref. 6 and also numerically in Sec. 6 in Ref. 6.

**Theorem 1.2.** *Assume (1.5), let  $D_u, D_v, \chi_u, \chi_v > 0$ , suppose that  $f, g \in C^1([0, \infty))$  are non-negative and let  $w_0, z_0 \in C^{2+\alpha}(\bar{\Omega})$  for some  $\alpha \in (0, 1)$  also be non-negative. Then there exists  $M > 0$  such that for all non-negative  $u_0, v_0 \in C^{2+\alpha}(\bar{\Omega})$  fulfilling (1.6), there is a global classical solution  $(u, v, w, z)$  of (1.1) fulfilling*

$$u(\cdot, t) \rightarrow \bar{u}_0 \quad \text{and} \quad v(\cdot, t) \rightarrow \bar{v}_0 \quad \text{in } L^p(\Omega), \quad (1.8)$$

$$w(\cdot, t) \rightarrow f(\bar{v}_0) \quad \text{and} \quad z(\cdot, t) \rightarrow g(\bar{u}_0) \quad \text{in } W^{1,2}(\Omega) \text{ and in } L^p(\Omega), \quad (1.9)$$

as  $t \rightarrow \infty$  for all  $p \in [1, \infty)$ . (Here, we have set  $\bar{\varphi} := \frac{1}{|\Omega|} \int_{\Omega} \varphi$  for  $\varphi \in L^1(\Omega)$ .)

While Theorem 1.2 settles the asymptotic behavior of small data solutions, it does not address the situation for large data; in particular, nontrivial large-time stabilization leading to some kind of segregation may still be possible. Usually, such questions are linked to the stability of heterogeneous steady states: For instance, related systems modeling other types of segregation such as the Shigesada–Kawasaki–Teramoto model<sup>54</sup> or the Cahn–Hilliard equation<sup>16</sup> feature a rich structure of heterogeneous steady states which may attract various solutions, see for instance Refs. 39 and 48 for the former and Refs. 53 and 63 for the latter model as well as references therein. For (1.1), however, the situation is entirely

different: Convergence toward nonconstant smooth steady states is impossible simply because there are no such equilibria; that is, all smooth solutions of

$$\begin{cases} 0 = D_u \Delta u + \chi_u \nabla \cdot (u \nabla w) & \text{in } \Omega, \\ 0 = D_v \Delta v + \chi_v \nabla \cdot (v \nabla z) & \text{in } \Omega, \\ 0 = -w + f(v) & \text{in } \Omega, \\ 0 = -z + g(u) & \text{in } \Omega, \\ D_u \partial_\nu u + \chi_u u \partial_\nu w = D_v \partial_\nu v + \chi_v v \partial_\nu z = 0 & \text{on } \partial\Omega, \end{cases} \quad (1.10)$$

are constant. In Proposition 4.2 in Ref. 6, this has already been shown in the two-dimensional setting for  $f(s) = g(s) = s$ ,  $s \geq 0$ . (In fact, there even the existence of heterogeneous weak solutions to (1.10) fulfilling an entropy estimate is ruled out.) We show that for many natural choices of  $f$  and  $g$ , no nontrivial smooth steady states exist.

**Proposition 1.1.** *Let  $\Omega \subset \mathbb{R}^n$ ,  $n \in \mathbb{N}$ , be a smooth, bounded domain, let  $f, g$  be non-negative, real analytic functions on  $(-\varepsilon, \infty)$  for some  $\varepsilon > 0$  and suppose that  $(u, v, w, z) \in (C^2(\bar{\Omega}))^4$  is a non-negative classical solution of (1.10). Then  $u, v, w$  and  $z$  are constant.*

This leaves open the question whether separation may occur at all. However, Proposition 1.1 does not rule out the most drastic way of separation, namely convergence toward multiples of characteristic functions of disjoint sets. Likewise, other forms of large-time behavior involving infinite time gradient blow up are not excluded either.

**Numerical simulations.** As answering the question whether the gangs may separate analytically appears to be very difficult, we instead perform numerical experiments which not only address the asymptotic behavior of large-time solutions but also generally illustrate the evolution of gang densities throughout time for various parameter regimes.

Cross-diffusion systems such as (1.1) can be considered as representatives of diffusion–convection–reaction equations (DCR) in computational fluid dynamics and often standard discretization methods for approximating the numerical solutions of DCR equations produce spuriously oscillating solutions whenever the convection is much larger than diffusion or reaction. Consequently, a vast variety of stabilization techniques has been introduced over the years to overcome these problems. The most popular among them is the streamline upwind/Petrov–Galerkin (SUPG) method introduced in Ref. 13, for which another stabilization acting in crosswind direction has been added by nonlinear so-called spurious oscillations at layers diminishing (SOLD) methods.<sup>35</sup> Local-projection stabilization (LPS) schemes,<sup>30</sup> unusual stabilized finite element methods,<sup>23</sup> Mizukami–Hughes method,<sup>49</sup> Galerkin-Least-Square methods<sup>34</sup> are among the other methods which have also been developed for stabilizing DCR problems in the convection-dominated

regime. While all the aforementioned approaches attempt to stabilize the finite element method by adding additional terms to the Galerkin finite element discretization, flux-corrected transport (FCT) schemes have been developed<sup>11, 40–42</sup> as a different technique which is a nonlinear scheme working on the algebraic level by modifying the algebraic equation obtained from the Galerkin finite element method. Stabilization methods were further investigated and developed for time-dependent DCR models, see for instance Refs. 15, 24, 36 and 37, to just mention a few. However, most of these finite element stabilization techniques deal with the cases where the convection terms are linear and their implementation to nonlinear convection terms still calls for further investigation. To this end, applications of finite element flux-corrected transport method (FEM-FCT) have been studied in Refs. 55–57 for Keller–Segel models; Ref. 33 considered a different model containing chemotaxis–Stokes equations and analyzed the error in Ref. 21, and the analysis of the solvability and positivity preservation of the FEM-FCT for a model of cancer invasion has been studied recently.<sup>32</sup>

Accordingly, it is not surprising that also for the double cross-diffusion system (1.1) studied in this paper, standard discretization schemes such as Galerkin finite element method give rise to nonphysical oscillations leading to negative values in the approximate solutions when the sensitivity magnitudes  $\chi_u$  and  $\chi_v$  of the strongly nonlinear convection terms are large. Thus as a remedy, we employ a high-resolution nonlinear FEM-FCT to reduce the oscillations and preserve the positivity of the solution. Moreover, we deal with strong nonlinear coupling in the system and nonlinearity of the proposed scheme simultaneously using a fixed-point iteration.

To answer the question whether the gangs’ populations separate from each other or not, we perform a series of numerical experiments using our newly designed algorithm which is implemented in finite element library deal.II.<sup>4, 5</sup> We show that for small values of  $\chi$ , gang populations stay completely mixed and that the approximate solutions converge toward constant steady states, see Sec. 6.1.1. However, both partial and complete separation is observed for large values of  $\chi$ , see Secs. 6.1.2 and 6.1.3. Moreover, we compare our outcome with the reported results for the two-component version of (1.1) (see Ref. 6) and a related agent-based model.<sup>2, 3</sup>

**Plan of the paper.** This paper is organized as follows. Following a brief local existence result in Sec. 2, Sec. 3 establishes various *a priori* bounds which eventually allow us to prove global existence of solutions, i.e. Theorem 1.1 and Corollary 1.1. Next, Theorem 1.2 and Proposition 1.1, that is, statements on the asymptotical stability and existence of steady states, are derived in Sec. 4. In Sec. 5, we first discretize the system using the  $\theta$ -method in time and a Galerkin finite element scheme in space and then enforce the positivity using the FEM-FCT scheme whenever the Galerkin method fails. We demonstrate the numerical results for different choices of parameters and study the convergence of our proposed method with respect to time step and mesh in Sec. 6.

## 2. Local Existence of Classical Solutions to a Transformed System

As the third and fourth equations in (1.1) do not regularize in space, the graffito-taxis terms in the first two equations are particularly challenging to deal with. To overcome this issue, we introduce the transformations

$$a := ue^{\xi_u w} \quad \text{and} \quad b := ve^{\xi_v z}, \quad \text{where} \quad \xi_u := \frac{\chi_u}{D_u} \quad \text{and} \quad \xi_v := \frac{\chi_v}{D_v},$$

variations of which have been already used for the analysis of several haptotaxis systems (cf. Refs. 22 and 25 for early examples). Indeed, as

$$\begin{aligned} a_t &= u_t e^{\xi_u w} + \xi_u w_t u e^{\xi_u w} \\ &= e^{\xi_u w} \nabla \cdot (D_u \nabla (a e^{-\xi_u w}) + \chi_u a e^{-\xi_u w} \nabla w) + \xi_u w_t a \\ &= D_u e^{\xi_u w} \nabla \cdot (e^{-\xi_u w} \nabla a) - \xi_u a w + \xi_u a f(v), \end{aligned}$$

and likewise

$$b_t = D_v e^{\xi_v z} \nabla \cdot (e^{-\xi_v z} \nabla b) - \xi_v b z + \xi_v b g(u),$$

in  $\Omega \times (0, \infty)$  whenever  $(u, v, w, z)$  is a global (classical) solution of (1.1), the system (1.1) is equivalent to

$$\begin{cases} a_t = D_u e^{\xi_u w} \nabla \cdot (e^{-\xi_u w} \nabla a) - \xi_u a w + \xi_u a f(b e^{-\xi_v z}) & \text{in } \Omega \times (0, \infty), \\ b_t = D_v e^{\xi_v z} \nabla \cdot (e^{-\xi_v z} \nabla b) - \xi_v b z + \xi_v b g(a e^{-\xi_u w}) & \text{in } \Omega \times (0, \infty), \\ w_t = -w + f(b e^{-\xi_v z}) & \text{in } \Omega \times (0, \infty), \\ z_t = -z + g(a e^{-\xi_u w}) & \text{in } \Omega \times (0, \infty), \\ \partial_\nu a = \partial_\nu b = 0 & \text{on } \partial\Omega \times (0, \infty), \\ (a, b, w, z)(\cdot, 0) = (a_0, b_0, w_0, z_0) & \text{in } \Omega, \end{cases} \quad (2.1)$$

where  $a_0 = u_0 e^{\xi_u w_0}$  and  $b_0 := v_0 e^{\xi_v z_0}$ . For this transformed system we have the following local existence result.

**Lemma 2.1.** *Let  $\Omega \subset \mathbb{R}^n$ ,  $n \in \{1, 2, 3\}$ , be a smooth, bounded domain, let  $f, g$  be as in (1.4),  $\alpha \in (0, 1)$  and*

$$a_0, b_0, w_0, z_0 \in C^{2+\alpha}(\bar{\Omega}) \quad \text{with} \quad \partial_\nu a_0 = \partial_\nu b_0 = 0 \quad \text{on } \partial\Omega. \quad (2.2)$$

*Then there exists  $T_{\max} \in (0, \infty]$  and a unique quadruple of non-negative functions*

$$(a, b, w, z) \in (C^{2+\alpha, 1+\frac{\alpha}{2}}(\bar{\Omega} \times (0, T_{\max})) \cap C^1(\bar{\Omega} \times [0, T_{\max}]))^4,$$

*with*

$$(\nabla a, \nabla b, \nabla w, \nabla z) \in (C^1(\bar{\Omega} \times [0, T_{\max}]))^4,$$

solving (2.1) classically with the property that if  $T_{\max} < \infty$ , then

$$\limsup_{t \nearrow T_{\max}} (\|a(\cdot, t)\|_{C^{1+\tilde{\alpha}}(\bar{\Omega})} + \|b(\cdot, t)\|_{C^{1+\tilde{\alpha}}(\bar{\Omega})}) = \infty, \tag{2.3}$$

for all  $\tilde{\alpha} \in (0, 1)$ .

**Proof.** This can be shown by means of a fixed-point argument and parabolic regularity theory, see for instance Lemma 2.5 in Ref. 29 or Lemmata 2.1 and 2.2 in Ref. 59 for details. □

### 3. Global Existence and Boundedness

In this section, we always assume that

$$\Omega \text{ is a smooth bounded domain in } \mathbb{R}^n, n \in \{1, 2, 3\} \text{ and } f, g \text{ are as in (1.4).} \tag{3.1}$$

In order to prove that the solution constructed in Lemma 2.1 is global in time, we need to show that (2.3) does not hold for some  $\tilde{\alpha} \in (0, 1)$ ; that is, that all solution components remain bounded in  $C^{2+\tilde{\alpha}}(\bar{\Omega})$ . This is achieved by a series of *a priori* estimates, which in part rely on previously established bounds. In particular, in Lemma 3.5 we apply Lemma 3.4 (and hence indirectly also Lemmata 3.1–3.3) to the solution of (2.1) with initial data  $(a, b, z, w)(\cdot, t_0)$  for some  $t_0 \in (0, T_{\max})$ . Therefore, we need to carefully track the dependency of the constants in the estimates below on the initial data and thus introduce the condition

$$a_0, b_0, w_0, z_0 \text{ fulfill (2.2) for some } \alpha \in (0, 1) \text{ and } \begin{cases} \|a_0\|_{L^\infty(\Omega)} \leq M, \\ \|b_0\|_{L^\infty(\Omega)} \leq M, \\ \|w_0\|_{L^\infty(\Omega)} \leq L, \\ \|z_0\|_{L^\infty(\Omega)} \leq L, \end{cases} \tag{3.2}$$

for  $M, L > 0$ . That is, for fixed  $M, L > 0$ , the constants given by Lemmas 3.1–3.4 do not depend on the precise form of the initial data, provided those fulfill (3.2). Also, the constants in the former three lemmata may not depend on  $M$ .

#### 3.1. $L^\infty$ estimates

We start with two rather basic estimates, namely  $L^\infty$  bounds for the last two and  $L^1$  bounds for the first two equations in (1.1). Already at this point we make use of the fact that (1.4) contains boundedness of  $f$  and  $g$ .

**Lemma 3.1.** *Assume (3.1) and let  $L > 0$ . Then there exists  $C_1 > 0$  such that for all  $M > 0$  and all initial data satisfying (3.2), the corresponding solution  $(a, b, w, z)$  of (2.1) given by Lemma 2.1 fulfills*

$$\|w(\cdot, t)\|_{L^\infty(\Omega)} \leq C_1 \quad \text{and} \quad \|z(\cdot, t)\|_{L^\infty(\Omega)} \leq C_1 \quad \text{for all } t \in (0, T_{\max}). \tag{3.3}$$

**Proof.** The functions  $\bar{w} := \max\{\|w_0\|_{L^\infty(\Omega)}, \|f\|_{L^\infty((0,\infty))}\}$  and  $\bar{z} := \max\{\|z_0\|_{L^\infty(\Omega)}, \|g\|_{L^\infty((0,\infty))}\}$  are bounded supersolutions of the third and fourth subproblems in (2.1), respectively. Moreover, both  $w$  and  $z$  are non-negative by Lemma 2.1.  $\square$

**Lemma 3.2.** *Assume (3.1) and let  $L > 0$ . Then there exists  $C_2 > 0$  such that for all  $M > 0$  and all initial data satisfying (3.2), the corresponding solution  $(a, b, w, z)$  of (2.1) given by Lemma 2.1 fulfills*

$$\|a(\cdot, t)\|_{L^1(\Omega)} \leq C_2 M \quad \text{and} \quad \|b(\cdot, t)\|_{L^1(\Omega)} \leq C_2 M \quad \text{for all } t \in (0, T_{\max}).$$

**Proof.** Integrating the first equation in (1.1) shows that  $\int_\Omega u(\cdot, t) = \int_\Omega u_0$  for  $t \in (0, T_{\max})$ . With  $C_1$  as given by Lemma 3.1, the definition of  $a$  thus implies

$$\int_\Omega a(\cdot, t) = \int_\Omega (ue^{\xi_u w})(\cdot, t) \leq e^{\xi_u C_1} \int_\Omega u_0 \leq M|\Omega|e^{\xi_u C_1} \quad \text{for all } t \in (0, T_{\max}).$$

The bound for  $b$  can be derived analogously.  $\square$

As to  $L^\infty$  bounds of  $a$  and  $b$ , we note that

$$\bar{a}(x, t) := \|a_0\|_{L^\infty(\Omega)} e^{\xi_u t \|f\|_{L^\infty((0,\infty))}} \quad \text{and} \quad \bar{b}(x, t) := \|b_0\|_{L^\infty(\Omega)} e^{\xi_v t \|g\|_{L^\infty((0,\infty))}},$$

$(x, t) \in \bar{\Omega} \times [0, T_{\max})$ , are supersolutions of the first two subproblems in (2.1) and that accordingly  $a$  and  $b$  are bounded locally in time. However, by employing testing procedures and a Moser-type iteration (following Refs. 1 and 58), we are also able to obtain  $L^\infty$  bounds which are not only time-independent but which additionally depend favorably on  $M$  as well.

**Lemma 3.3.** *Assume (3.1) and let  $L > 0$ . Then there exists  $C_3 > 0$  such that for all  $M > 0$  and all initial data satisfying (3.2), the corresponding solution  $(a, b, w, z)$  of (2.1) given by Lemma 2.1 fulfills*

$$\|a(\cdot, t)\|_{L^\infty(\Omega)} \leq C_3 M \quad \text{and} \quad \|b(\cdot, t)\|_{L^\infty(\Omega)} \leq C_3 M, \tag{3.4}$$

for all  $t \in (0, T_{\max})$ .

**Proof.** We fix initial data satisfying (3.2) and the corresponding solution  $(a, b, w, z)$  of (2.1) given by Lemma 2.1.

Moreover, by a quantitative version of Ehrling’s lemma proved in Lemma 2.5 in Ref. 26, there exist  $c_1 > 0$  and  $\mu > 0$  such that

$$c_2 p \int_\Omega \varphi^2 \leq \frac{2D_u e^{-\xi_u C_1}}{p} \int_\Omega |\nabla \varphi|^2 + c_1 p^\mu \left( \int_\Omega |\varphi| \right)^2,$$



for all  $\varphi \in W^{1,2}(\Omega)$  and all  $p \geq 2$ , where  $c_2 := \xi_u \|f\|_{L^\infty((0,\infty))} + 1$  and where  $C_1$  is given by Lemma 3.1. For  $p \geq 2$  and  $T \in (0, T_{\max})$ , we can thus calculate

$$\begin{aligned} & \frac{d}{dt} \int_{\Omega} e^{-\xi_u w} a^p \\ &= p \int_{\Omega} e^{-\xi_u w} a^{p-1} a_t - \xi_u \int_{\Omega} e^{-\xi_u w} a^p w_t \\ &= -D_u p \int_{\Omega} e^{-\xi_u w} \nabla a \cdot \nabla a^{p-1} + \xi_u (p-1) \int_{\Omega} e^{-\xi_u w} a^p (-w + f(b e^{-\xi_v z})) \\ &\leq -\frac{4D_u (p-1) e^{-\xi_u C_1}}{p^2} \int_{\Omega} |\nabla a^{\frac{p}{2}}|^2 + c_2 (p-1) \int_{\Omega} e^{-\xi_u w} a^p \\ &\leq -c_2 \int_{\Omega} e^{-\xi_u w} a^p - \frac{2D_u e^{-\xi_u C_1}}{p} \int_{\Omega} |\nabla a^{\frac{p}{2}}|^2 + c_2 p \int_{\Omega} (a^{\frac{p}{2}})^2 \\ &\leq -c_2 \int_{\Omega} e^{-\xi_u w} a^p + c_1 p^\mu \left( \int_{\Omega} a^{\frac{p}{2}} \right)^2 \\ &\leq -c_2 \int_{\Omega} e^{-\xi_u w} a^p + c_1 p^\mu \sup_{s \in (0, T)} \left( \int_{\Omega} a^{\frac{p}{2}}(\cdot, s) \right)^2 \quad \text{in } (0, T), \end{aligned}$$

which in combination with an ordinary differential equation (ODE) comparison argument implies

$$\int_{\Omega} (e^{-\xi_u w} a^p)(\cdot, t) \leq \max \left\{ \int_{\Omega} e^{-\xi_u w_0} a_0^p, \frac{c_1 p^\mu}{c_2} \sup_{s \in (0, T)} \left( \int_{\Omega} a^{\frac{p}{2}}(\cdot, s) \right)^2 \right\},$$

for all  $T \in (0, T_{\max})$ , all  $t \in (0, T)$  and all  $p \geq 2$ . Setting  $c_3 := e^{\xi_u C_1} \max\{1, \frac{c_1}{c_2}\}$ , we thus obtain

$$\int_{\Omega} a^p(\cdot, t) \leq c_3 \max \left\{ \int_{\Omega} a_0^p, p^\mu \sup_{s \in (0, T)} \left( \int_{\Omega} a^{\frac{p}{2}}(\cdot, s) \right)^2 \right\}, \tag{3.5}$$

for all  $T \in (0, T_{\max})$ , all  $t \in (0, T)$  and all  $p \geq 2$ . We next set

$$p_j := 2^j \quad \text{and} \quad A_j(T) := \sup_{s \in (0, T)} \|a(\cdot, s)\|_{L^{p_j}(\Omega)} \quad \text{for } j \in \mathbb{N}_0 \text{ and } T \in (0, T_{\max}),$$

so that with  $c_4 := c_3 \max\{1, |\Omega|\}$  and as the condition  $\|a_0\|_{L^\infty(\Omega)} \leq M$  in (3.2) implies  $\|a_0\|_{L^{p_j}(\Omega)} \leq |\Omega|^{\frac{1}{p_j}} M \leq \max\{1, |\Omega|\} M$ , we further infer from (3.5) that

$$A_j(T) \leq \max\{c_4 M, (c_3 p_j^\mu)^{\frac{1}{p_j}} A_{j-1}(T)\} \quad \text{for all } j \in \mathbb{N} \text{ and } T \in (0, T_{\max}).$$

We now fix  $T \in (0, T_{\max})$ . If there are infinitely many  $j \in \mathbb{N}_0$  with  $A_j(T) \leq c_4 M$ , then

$$\|a\|_{L^\infty(\Omega \times (0, T))} = \liminf_{j \rightarrow \infty} A_j(T) \leq c_4 M. \tag{3.6}$$

Else there exists  $j_0 \in \mathbb{N}_0$  such that

$$A_j(T) \leq (c_3 p_j^\mu)^{\frac{1}{p_j}} A_{j-1}(T) \quad \text{for all } j > j_0. \tag{3.7}$$

(We note that while  $j_0$  may depend on the initial data, the constant  $C_3$  defined below will not.) Choosing  $j_0$  minimal, we can moreover assume that

$$A_{j_0}(T) \leq \max\{C_2, c_4\}M, \tag{3.8}$$

where  $C_2$  is as in Lemma 3.2. (That lemma guarantees  $A_0(T) \leq C_2M$ .) By induction and as  $p_j = 2^j$  for  $j \in \mathbb{N}$ , (3.7) implies

$$A_j(T) \leq \left( \prod_{k=j_0+1}^j (c_3 p_k^\mu)^{\frac{1}{p_k}} \right) A_{j_0} = c_3^{\sum_{k=j_0+1}^j 2^{-k}} \cdot 2^{\mu \sum_{k=j_0+1}^j k 2^{-k}} \cdot A_{j_0},$$

for all  $j > j_0$ . In combination with (3.8), we arrive at

$$\begin{aligned} \|a\|_{L^\infty(\Omega \times (0, T))} &= \lim_{j \rightarrow \infty} A_j(T) \leq c_3^{\sum_{k=j_0+1}^\infty 2^{-k}} \cdot 2^{\mu \sum_{k=j_0+1}^\infty k 2^{-k}} \cdot A_{j_0} \\ &\leq c_3^{\sum_{k=0}^\infty 2^{-k}} \cdot 2^{\mu \sum_{k=0}^\infty k 2^{-k}} \cdot \max\{C_2, c_4\}M =: c_5M. \end{aligned}$$

Together with (3.6) this implies  $\|a\|_{L^\infty(\Omega \times (0, T))} \leq C_3M$  for all  $T \in (0, T_{\max})$ , where  $C_3 := \max\{c_4, c_5\}$ . Letting  $T \nearrow T_{\max}$ , we obtain the first statement in (3.4), while the second one follows upon an analogous computation for the second solution component of (2.1), possibly after enlarging  $C_3$ .  $\square$

### 3.2. Gradient estimates

Our next step toward showing that (2.3) cannot hold consists in deriving uniform-in-time  $L^4$  estimates for  $\nabla w$  and  $\nabla z$ . To that end, we follow a technique introduced in Lemmata 3.13–3.15 in Ref. 59 for a haptotaxis system. As a preparation, we first note that the time regularization in the third and fourth equation in (2.1) implies that space–time gradient estimates for  $a$  and  $b$  imply uniform-in-time gradient estimates for  $w$  and  $z$ .

**Lemma 3.4.** *Assume (3.1) and let  $L, M > 0$ . Moreover, let  $T_0 > 0$  and  $p \in (1, \infty)$ . Then there exists  $C_4 > 0$  such that for all initial data satisfying (3.2), the corresponding solution  $(a, b, w, z)$  of (2.1) given by Lemma 2.1 fulfills*

$$\begin{aligned} &\int_{\Omega} |\nabla w(\cdot, t)|^p + \int_{\Omega} |\nabla z(\cdot, t)|^p \\ &\leq C_4 \left( \int_{\Omega} w_0^p + \int_{\Omega} z_0^p + \int_0^t \int_{\Omega} |\nabla a|^p + \int_0^t \int_{\Omega} |\nabla b|^p \right), \end{aligned} \tag{3.9}$$

for all  $t \in (0, T)$ , where  $T := \min\{T_0, T_{\max}\}$ .

**Proof.** We again fix initial data satisfying (3.2) and the solution  $(a, b, w, z)$  of (2.1) given by Lemma 2.1. By Lemma 3.3, we can find  $C_3 > 0$  such that  $a, b \leq C_3 M$  in  $\Omega \times (0, T)$ . Since according to the variation-of-constants formula we may write  $w(x, t) = e^{-t}w_0(x) + \int_0^t e^{-(t-s)} f((be^{-\xi_v z})(x, s)) ds$  for  $(x, t) \in \Omega \times (0, T_{\max})$ , we have

$$\begin{aligned} \int_{\Omega} |\nabla w(\cdot, t)|^p &\leq 2^{p-1} \int_{\Omega} |e^{-t} \nabla w_0|^p + 2^{p-1} \int_{\Omega} \int_0^t e^{-(t-s)} |\nabla(f((be^{-\xi_v z})(\cdot, s)))|^p ds \\ &\leq 2^{p-1} \int_{\Omega} |\nabla w_0|^p + 2^{p-1} \|f'\|_{C^0([0, C_3 M])} \int_0^t \int_{\Omega} |\nabla(be^{-\xi_v z})|^p, \end{aligned}$$

for  $t \in (0, T)$ . As

$$\begin{aligned} |\nabla(be^{-\xi_v z})|^p &\leq 2^{p-1} e^{-\xi_v p z} |\nabla b|^p + 2^{p-1} \xi_v^p b^p e^{-\xi_v p z} |\nabla z|^p \\ &\leq 2^{p-1} |\nabla b|^p + 2^{p-1} \xi_v^p C_3^p M^p |\nabla z|^p \quad \text{in } \Omega \times (0, T), \end{aligned}$$

and together with an analogous argumentation for  $\int_{\Omega} |\nabla z(\cdot, t)|^p$ , we can conclude

$$\begin{aligned} \int_{\Omega} |\nabla w(\cdot, t)|^p + \int_{\Omega} |\nabla z(\cdot, t)|^p &\leq c_1 \left( \int_{\Omega} |\nabla w_0|^p + \int_{\Omega} |\nabla z_0|^p + \int_0^t \int_{\Omega} |\nabla a|^p + \int_0^t \int_{\Omega} |\nabla b|^p \right) \\ &\quad + c_2 \int_0^t \left( \int_{\Omega} |\nabla w|^p + \int_{\Omega} |\nabla z|^p \right), \end{aligned}$$

for  $t \in (0, T)$ , where

$$c_1 := \max\{2^{p-1}, 2^{2p-2} \|f'\|_{C^0([0, C_3 M])}, 2^{2p-2} \|g'\|_{C^0([0, C_3 M])}\},$$

and

$$c_2 := 2^{2p-2} C_3^p M^p \max\{\xi_v^p \|f'\|_{C^0([0, C_3 M])}, \xi_u^p \|g'\|_{C^0([0, C_3 M])}\}.$$

Thus, Grönwall's inequality asserts the statement for  $C_4 := c_1 e^{c_2 T_0}$ . □

In order to show that the right-hand side in (3.9) is bounded for  $p = 4$ , we consider the evolution of the function  $\frac{1}{2}(\int_{\Omega} |\nabla a|^2 + \int_{\Omega} |\nabla b|^2)$ . As it turns out, however, we can only control its time derivative on small timescales and thus aim to derive estimates in  $(t_0, T_{\max})$  for  $t_0$  close to  $T_{\max}$  (which may be assumed to be finite for the sake of contradiction) only. To that end, it is crucial that Lemma 3.3 provides  $L^\infty$  estimates for  $a$  and  $b$  in terms of the  $L^\infty(\Omega)$  norm of  $a_0$  and  $b_0$ , as we then may apply Lemma 3.4 to the solution with initial data  $(a, b, w, z)(\cdot, t_0)$  without worrying about the dependency of the constant  $C_4$  thereby obtained on  $t_0$ .

**Lemma 3.5.** *Assume (3.1) as well as (3.2) for some  $L, M > 0$  and denote the solution of (2.1) given by Lemma 2.1 by  $(a, b, w, z)$ . Moreover, let  $T \in (0, T_{\max}] \cap$*

$(0, \infty)$ . Then there exists  $C_5 > 0$  such that

$$\int_{\Omega} |\nabla w(\cdot, t)|^4 + \int_{\Omega} |\nabla z(\cdot, t)|^4 \leq C_5 \quad \text{for } t \in (0, T). \tag{3.10}$$

**Proof.** It follows from the Gagliardo–Nirenberg inequality (that is, from Ref. 50; or, more directly, from Lemma A.3 in Ref. 27) that there is  $c_1 > 0$  such that

$$\frac{\max\{D_u, D_v\}}{\min\{D_u, D_v\}} \int_{\Omega} |\nabla \varphi|^4 \leq c_1 \left( \int_{\Omega} |\Delta \varphi|^2 \right) \|\varphi\|_{L^\infty(\Omega)}^2 + c_1 \|\varphi\|_{L^\infty(\Omega)}^4, \tag{3.11}$$

for all  $\varphi \in C^2(\bar{\Omega})$  with  $\partial_\nu \varphi = 0$  on  $\partial\Omega$ . Moreover, we let  $C_1$  and  $C_3$  be as given by Lemmas 3.1 and 3.3, respectively. Then (3.11) and Lemma 3.3 imply

$$\int_{\Omega} |\nabla a|^4 + \int_{\Omega} |\nabla b|^4 \leq C_3^2 M^2 c_1 \left( \int_{\Omega} |\Delta a|^2 + \int_{\Omega} |\Delta b|^2 \right) + 2C_3^4 M^4 c_1. \tag{3.12}$$

Next, we apply Lemma 3.4 (with  $M = C_3 M$ ,  $L = C_1$ ,  $T_0 = 1$  and  $p = 4$ ) to obtain  $\widehat{C}_4 > 0$  with the property that whenever a quadruple of functions  $(\widehat{a}_0, \widehat{b}_0, \widehat{w}_0, \widehat{z}_0)$  satisfies (3.2) with  $M$  replaced by  $C_3 M$  (and  $(a_0, b_0, w_0, z_0)$  replaced by  $(\widehat{a}_0, \widehat{b}_0, \widehat{w}_0, \widehat{z}_0)$ ), then the corresponding solution  $(\widehat{a}, \widehat{b}, \widehat{w}, \widehat{z})$  of (2.1) with maximal existence time  $\widehat{T}_{\max}$  given by Lemma 2.1 fulfills

$$\begin{aligned} & \int_{\Omega} |\nabla \widehat{w}(\cdot, t)|^4 + \int_{\Omega} |\nabla \widehat{z}(\cdot, t)|^4 \\ & \leq \widehat{C}_4 \left( \int_{\Omega} (\widehat{w}_0)^4 + \int_{\Omega} (\widehat{z}_0)^4 + \int_0^t \int_{\Omega} |\nabla \widehat{a}|^4 + \int_0^t \int_{\Omega} |\nabla \widehat{b}|^4 \right), \end{aligned}$$

for all  $t \in (0, \max\{1, \widehat{T}_{\max}\})$ . Setting  $D := \min\{D_u, D_v\}$ ,  $\chi := \max\{\chi_u, \chi_v\}$ ,  $\xi := \frac{\chi}{D}$

$$t_0 := \max \left\{ 0, T - \frac{1}{16\xi^4 C_3^4 \widehat{C}_4 M^4 c_1^2}, T - 1 \right\}, \tag{3.13}$$

as well as  $c_2 := \widehat{C}_4 (\int_{\Omega} w^4(\cdot, t_0) + \int_{\Omega} z^4(\cdot, t_0))$  and recalling that classical solutions of (2.1) are unique by Lemma 2.1, we can conclude

$$\int_{\Omega} |\nabla w(\cdot, t)|^4 + \int_{\Omega} |\nabla z(\cdot, t)|^4 \leq c_2 + \widehat{C}_4 \left( \int_{t_0}^t \int_{\Omega} |\nabla a|^4 + \int_{t_0}^t \int_{\Omega} |\nabla b|^4 \right), \tag{3.14}$$

for all  $t \in (t_0, T)$ . Since  $D_u e^{\xi_u w} \nabla \cdot (e^{-\xi_u w} \nabla a) = D_u \Delta a - \chi_u \nabla a \cdot \nabla w$ , testing the equation for  $a$  with  $-\Delta a$  and applying Young’s inequality thrice gives

$$\begin{aligned} & \frac{1}{2} \frac{d}{dt} \int_{\Omega} |\nabla a|^2 \\ & = -D_u \int_{\Omega} |\Delta a|^2 + \chi_u \int_{\Omega} (\nabla a \cdot \nabla w) \Delta a - \xi_u \int_{\Omega} (-aw + af(be^{-\xi_v z})) \Delta a \end{aligned}$$

$$\begin{aligned} &\leq -\frac{D_u}{2} \int_{\Omega} |\Delta a|^2 + \chi \xi \int_{\Omega} |\nabla a|^2 |\nabla w|^2 + \frac{\xi^2}{D} \int_{\Omega} a^2 (w + f(be^{-\xi v z}))^2 \\ &\leq -\frac{D_u}{2} \int_{\Omega} |\Delta a|^2 + \frac{D}{8C_3^2 M^2 c_1} \int_{\Omega} |\nabla a|^4 + 2\xi^4 DC_3^2 M^2 c_1 \int_{\Omega} |\nabla w|^4 + c_3, \end{aligned}$$

in  $(0, T)$ , where  $c_3 := \frac{C_3^2 M^2 \xi^2}{D} (C_1 + \|f\|_{L^\infty((0, \infty))})^2 |\Omega|$ . By integrating in time from  $t_0$  to  $t$  and noting that  $T - t_0 \leq 1$ , we obtain

$$\begin{aligned} &\int_{\Omega} |\nabla a(\cdot, t)|^2 - \int_{\Omega} |\nabla a(\cdot, t_0)|^2 + D_u \int_{t_0}^t \int_{\Omega} |\Delta a|^2 \\ &\leq \frac{D}{4C_3^2 M^2 c_1} \int_{t_0}^t \int_{\Omega} |\nabla a|^4 + 4\xi^4 DC_3^2 M^2 c_1 (t - t_0) \\ &\quad \times \sup_{s \in (t_0, t)} \int_{\Omega} |\nabla w(\cdot, s)|^4 + 2c_3, \end{aligned} \tag{3.15}$$

for all  $t \in (t_0, T)$ . Likewise, there is  $c_4 > 0$  such that

$$\begin{aligned} &\int_{\Omega} |\nabla b(\cdot, t)|^2 - \int_{\Omega} |\nabla b(\cdot, t_0)|^2 + D_v \int_{t_0}^t \int_{\Omega} |\Delta b|^2 \\ &\leq \frac{D}{4C_3^2 M^2 c_1} \int_{t_0}^t \int_{\Omega} |\nabla b|^4 + 4\xi^4 DC_3^2 M^2 c_1 (t - t_0) \\ &\quad \times \sup_{s \in (t_0, t)} \int_{\Omega} |\nabla z(\cdot, s)|^4 + 2c_4, \end{aligned} \tag{3.16}$$

for all  $t \in (t_0, T)$ . By (3.15), (3.16), (3.14), (3.13) and (3.12), we therefore have

$$\begin{aligned} &\int_{\Omega} |\nabla a(\cdot, t)|^2 + \int_{\Omega} |\nabla b(\cdot, t)|^2 - \int_{\Omega} |\nabla a(\cdot, t_0)|^2 - \int_{\Omega} |\nabla b(\cdot, t_0)|^2 \\ &\quad + D_u \int_0^t \int_{\Omega} |\Delta a|^2 + D_v \int_0^t \int_{\Omega} |\Delta b|^2 \\ &\leq \frac{D}{4C_3^2 M^2 c_1} \left( \int_{t_0}^t \int_{\Omega} |\nabla a|^4 + \int_{t_0}^t \int_{\Omega} |\nabla b|^4 \right) + 4\xi^4 DC_3^2 M^2 c_1 (t - t_0) \\ &\quad \times \sup_{s \in (t_0, t)} \left( \int_{\Omega} |\nabla w(\cdot, s)|^4 + \int_{\Omega} |\nabla z(\cdot, s)|^4 \right) + 2(c_3 + c_4) \\ &\leq \underbrace{\left( \frac{D}{4C_3^2 M^2 c_1} + 4\xi^4 DC_3^2 \widehat{C}_4 M^2 c_1 (T - t_0) \right)}_{\leq D/(2C_3^2 M^2 c_1)} \\ &\quad \times \left( \int_{t_0}^t \int_{\Omega} |\nabla a|^4 + \int_{t_0}^t \int_{\Omega} |\nabla b|^4 \right) + c_5 \end{aligned}$$

$$\begin{aligned} &\leq D_u \int_{t_0}^t \int_{\Omega} |\Delta a|^2 + D_v \int_{t_0}^t \int_{\Omega} |\Delta b|^2 \\ &\quad - \frac{D}{2C_3^2 M^2 c_1} \left( \int_{t_0}^t \int_{\Omega} |\nabla a|^4 + \int_{t_0}^t \int_{\Omega} |\nabla b|^4 \right) + c_6, \end{aligned}$$

for  $t \in (t_0, T)$ , where  $c_5 := 4\xi^4 DC_3^2 M^2 c_1 c_2 + 2c_3 + 2c_4$  and  $c_6 := c_5 + 2DC_3^2 M^2$ . By Beppo Levi's theorem, this implies

$$\int_{t_0}^T \int_{\Omega} |\nabla a|^4 + \int_{t_0}^T \int_{\Omega} |\nabla b|^4 \leq \frac{2C_3^2 M^2 c_1}{D} \left( \int_{\Omega} |\nabla a(\cdot, t_0)|^2 + \int_{\Omega} |\nabla b(\cdot, t_0)|^2 + c_6 \right).$$

Another application of Lemma 3.4 then shows that the desired estimate (3.10) holds for all  $t \in (t_0, T)$  and some  $C_5 > 0$ , while the inclusions  $w, z \in C^1(\bar{\Omega} \times [0, t_0])$  trivially entail (3.10) also for  $t \in [0, t_0]$  (possibly after enlarging  $C_5$ ).  $\square$

### 3.3. Solutions are global in time: Proof of Theorem 1.1 and Corollary 1.1

With Lemma 3.5 at hand, globality in time can be shown as in Lemma 2.14 in Ref. 29 (or Lemma 2.2 in Ref. 59): Parabolic regularity theory rapidly upgrades the bounds implied by (3.3), (3.4) and (3.10) to Hölder estimates.

**Lemma 3.6.** *Assume (3.1) as well as (3.2) for some  $L, M > 0$ . Then the solution  $(a, b, w, z)$  given by Lemma 2.1 is global in time; that is,  $T_{\max} = \infty$ .*

**Proof.** Suppose  $T_{\max} < \infty$ . We rewrite the first two equations in (2.1) as

$$a_t = D_u \Delta a - \chi_u \nabla a \cdot \nabla w + \psi_a \quad \text{and} \quad b_t = D_v \Delta b - \chi_v \nabla b \cdot \nabla z + \psi_b,$$

in  $\Omega \times (0, \infty)$ , where

$$\psi_a = -\xi_u a w + \xi_u a f(b e^{-\xi_v z}) \quad \text{and} \quad \psi_b = -\xi_v b z + \xi_v b g(a e^{-\xi_u w}).$$

As  $a, b, w, z$  belong to  $L^\infty(\Omega \times (0, T_{\max}))$  by Lemmas 3.3 and 3.1, so do  $\psi_1, \psi_2$ . Also recalling (2.2) and Lemma 3.5, we may thus apply (a consequence of) maximal Sobolev regularity (cf. Lemma 2.13 in Ref. 29) to obtain  $c_1 > 0$  such that

$$\|\nabla a\|_{L^{12}((0, T_{\max}); L^\infty(\Omega))} + \|\nabla b\|_{L^{12}((0, T_{\max}); L^\infty(\Omega))} \leq c_1.$$

According to Lemma 3.4, there then exists  $c_2 > 0$  with

$$\|\nabla w\|_{L^\infty((0, T_{\max}); L^{12}(\Omega))} + \|\nabla z\|_{L^\infty((0, T_{\max}); L^{12}(\Omega))} \leq c_2.$$

This allows us to again invoke Lemma 2.13 in Ref. 29 to obtain  $c_3 > 0$  such that

$$\|a_t\|_{L^{12}(Q_T)} + \|\Delta a\|_{L^{12}(Q_T)} + \|b_t\|_{L^{12}(Q_T)} + \|\Delta b\|_{L^{12}(Q_T)} \leq c_3,$$

where  $Q_T := \Omega \times (0, T_{\max})$ . Thus, by Lemma II.3.3 in Ref. 43 there is  $c_4 > 0$  such that

$$\|a\|_{C^{\frac{19}{12}, \frac{19}{24}}(\bar{\Omega} \times [0, T_{\max}])} + \|b\|_{C^{\frac{19}{12}, \frac{19}{24}}(\bar{\Omega} \times [0, T_{\max}])} \leq c_4,$$

which contradicts the extensibility criterion in Lemma 2.1.  $\square$

Theorem 1.1 follows now easily from the lemmata above.

**Proof of Theorem 1.1.** That the solution  $(a, b, w, z)$  of (2.1) constructed in Lemma 2.1 is global in time has been asserted in Lemma 3.6. Upon setting  $u := ae^{-\xi_u w}$  and  $v := be^{-\xi_v z}$ , we also obtain a global classical solution of (1.1) which due to Lemma 3.3, the evident estimates  $u \leq a, v \leq b$  in  $\bar{\Omega} \times [0, \infty)$  and Lemma 3.1 moreover fulfills (1.7) for some  $C > 0$ .  $\square$

Finally, we show that Theorem 1.1 allows for a quick proof of Corollary 1.1; that is, of the existence of small data solutions for (1.1) with  $f(s) = g(s) = s$  for  $s \geq 0$ .

**Proof of Corollary 1.1.** We let  $\Omega, \alpha$  and  $w_0, z_0$  be as in the statement of Corollary 1.1. Moreover, we fix a non-negative cutoff function  $\zeta \in C^\infty([0, \infty))$  with  $\zeta(s) = s$  for  $s \in [0, 1]$  and  $\zeta \equiv 2$  in  $[2, \infty)$ . Then  $\tilde{f} = \tilde{g} = \zeta$  fulfill (1.4), so that for each  $M > 0$ , Theorem 1.1 provides us with a unique, global, bounded, non-negative classical solution  $(u, v, w, z)$  of (1.1) (with  $f$  and  $g$  replaced by  $\tilde{f}$  and  $\tilde{g}$ ) and  $C > 0$  such that (1.7) holds. In particular, if  $M \leq \frac{1}{C}$ , then  $u, v \leq 1$  so that in that case  $(u, v, w, z)$  also solves (1.1) with  $f(s) = g(s) = s$  for  $s \geq 0$ .  $\square$

#### 4. Smooth Steady States

While for all  $M_1, M_2 > 0$  the tuple  $((M_1, M_2, f(M_2), g(M_1)))$  forms a smooth steady state of (1.1), the conservation of mass for the first two solution components implies that these steady states may appear as limits only for the choices  $M_1 = \bar{u}_0$  and  $M_2 = \bar{v}_0$ . In Sec. 4.1, we prove Theorem 1.2; that is, that this equilibrium indeed attracts solutions whenever the initial data are small.

The asymptotic behavior of solutions not covered by Theorem 1.2 will be a main aspect of our numerical experiments performed in Sec. 6. Analytically, we can at least rule out stabilization toward nonconstant smooth steady states: We show that there are no such equilibria in Sec. 4.2.

##### 4.1. Convergence to homogeneous steady states for small data: Proof of Theorem 1.2

Theorem 1.1 already contains a key step of the convergence proof, namely the fact that smallness of  $u_0$  and  $v_0$  implies smallness of  $u$  and  $v$  for all times. With these estimates at hand, we can show that

$$y := c \int_{\Omega} (u - \bar{u}_0)^2 + c \int_{\Omega} (v - \bar{v}_0)^2 + \int_{\Omega} (w - f(\bar{v}_0))^2 + \int_{\Omega} |\nabla w|^2 + \int_{\Omega} (z - g(\bar{u}_0))^2 + \int_{\Omega} |\nabla z|^2,$$

(for suitably chosen  $c > 0$ ) is a subsolution of a homogeneous ODE of the form  $y' = -c'y$  for some  $c' > 0$  and hence converges exponentially fast to 0.

**Proof of Theorem 1.2.** We fix  $\Omega, D_u, D_v, \chi_u, \chi_v, f, g, \alpha, w_0$  and  $z_0$  as in the statement of Theorem 1.2. By the Poincaré inequality, there is  $c_1 > 0$  such that

$$\int_{\Omega} (\varphi - \bar{\varphi})^2 \leq c_1 \int_{\Omega} |\nabla \varphi|^2 \quad \text{for all } \varphi \in W^{1,2}(\Omega), \tag{4.1}$$

and due to the assumptions on  $f$  and  $g$ , there is  $c_2 > 0$  such that

$$\|f'\|_{C^0([0,1])} \leq c_2 \quad \text{and} \quad \|g'\|_{C^0([0,1])} \leq c_2. \tag{4.2}$$

As in the proof of Corollary 1.1 we fix  $\zeta \in C^\infty([0, \infty))$  with  $\zeta(s) = 1$  for  $s \in [0, 1]$  and  $\zeta \equiv 2$  in  $[2, \infty)$ . Since  $\tilde{f} := \zeta f$  and  $\tilde{g} := \zeta g$  fulfill (1.4), Theorem 1.1 asserts that there exists  $c_3 > 0$  such that for all  $M > 0$  the following holds: If  $u_0, v_0 \in C^{2+\alpha}(\bar{\Omega})$  are non-negative and satisfy (1.6) for some  $M > 0$ , there exists a global, non-negative classical solution  $(u, v, w, z)$  of (1.1) (with  $f, g$  replaced by  $\tilde{f}, \tilde{g}$ ) with

$$\|u\|_{L^\infty(\Omega \times (0, \infty))} \leq M c_3 \quad \text{and} \quad \|v\|_{L^\infty(\Omega \times (0, \infty))} \leq M c_3. \tag{4.3}$$

We choose  $M > 0$  so small that

$$M^2 c_3^2 \leq \min \left\{ \frac{1}{4c_4 \max\{\frac{\chi_u^2}{D_u}, \frac{\chi_v^2}{D_v}\}}, 1 \right\}, \quad \text{where } c_4 := \frac{3c_2^2 \max\{c_1, 1\}}{\min\{D_u, D_v\}}, \tag{4.4}$$

and fix non-negative  $u_0, v_0 \in C^{2+\alpha}(\bar{\Omega})$  fulfilling (1.6) as well as the solution  $(u, v, w, z)$  of (1.1) (with  $f, g$  replaced by  $\tilde{f}, \tilde{g}$ ) given by Theorem 1.1. We note that (4.3) and the second estimate contained in (4.4) imply  $u, v \leq 1$  in  $\bar{\Omega} \times [0, \infty)$ . Thus,  $(\tilde{f}(u), \tilde{g}(v)) = (f(u), g(v))$  in  $\bar{\Omega} \times [0, \infty)$  and (4.2) results in

$$\max\{|f'(u)|, |g'(v)|\} \leq c_2 \quad \text{in } \bar{\Omega} \times [0, \infty). \tag{4.5}$$

By testing the first equation in (1.1) with  $u - \bar{u}_0$  and making use of Young's inequality, (4.3) and (4.1) (we note that integrating the first equation implies  $\int_{\Omega} u(\cdot, t) = \int_{\Omega} u_0$  for all  $t \geq 0$ , so that (4.1) is indeed applicable), we obtain

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \int_{\Omega} (u - \bar{u}_0)^2 &= -D_u \int_{\Omega} |\nabla u|^2 - \chi_u \int_{\Omega} u \nabla u \cdot \nabla w \\ &\leq -\frac{3D_u}{4} \int_{\Omega} |\nabla u|^2 + \frac{M^2 c_3^2 \chi_u^2}{D_u} \int_{\Omega} |\nabla w|^2 \\ &\leq -\frac{D_u}{2} \int_{\Omega} |\nabla u|^2 - \frac{D_u}{4c_1} \int_{\Omega} (u - \bar{u}_0)^2 + \frac{M^2 c_3^2 \chi_u^2}{D_u} \int_{\Omega} |\nabla w|^2, \end{aligned}$$



in  $(0, \infty)$ . Moreover, testing the third equation in (1.1), namely  $w_t = -w + f(v) = -(w - f(\bar{v}_0)) + f(v) - f(\bar{v}_0)$ , with  $w - f(\bar{v}_0)$  and Young's inequality, the mean value theorem and (4.5) yield

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \int_{\Omega} (w - f(\bar{v}_0))^2 &= - \int_{\Omega} (w - f(\bar{v}_0))^2 + \int_{\Omega} (f(v) - f(\bar{v}_0))(w - f(\bar{v}_0)) \\ &\leq -\frac{1}{2} \int_{\Omega} (w - f(\bar{v}_0))^2 + \frac{c_2^2}{2} \int_{\Omega} (v - \bar{v}_0)^2 \quad \text{in } (0, \infty), \end{aligned}$$

while testing the same equation with  $-\Delta w$  gives

$$\frac{1}{2} \frac{d}{dt} \int_{\Omega} |\nabla w|^2 = - \int_{\Omega} |\nabla w|^2 + \int_{\Omega} f'(v) \nabla v \cdot \nabla w \leq -\frac{1}{2} \int_{\Omega} |\nabla w|^2 + \frac{c_2^2}{2} \int_{\Omega} |\nabla v|^2,$$

in  $(0, \infty)$ . Analogously

$$\frac{1}{2} \frac{d}{dt} \int_{\Omega} (v - \bar{v}_0)^2 \leq -\frac{D_v}{2} \int_{\Omega} |\nabla v|^2 - \frac{D_v}{4c_1} \int_{\Omega} (v - \bar{v}_0)^2 + \frac{M^2 c_3^2 \chi_v^2}{D_v} \int_{\Omega} |\nabla z|^2,$$

and

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \left( \int_{\Omega} (z - g(\bar{u}_0))^2 + \int_{\Omega} |\nabla z|^2 \right) &\leq -\frac{1}{2} \int_{\Omega} (z - g(\bar{u}_0))^2 + \frac{c_2^2}{2} \int_{\Omega} (u - \bar{u}_0)^2 \\ &\quad - \frac{1}{2} \int_{\Omega} |\nabla z|^2 + \frac{c_2^2}{2} \int_{\Omega} |\nabla u|^2, \end{aligned}$$

hold in  $(0, \infty)$ . Recalling (4.4), we conclude that the function  $y: [0, \infty) \rightarrow [0, \infty)$  defined by

$$\begin{aligned} y(t) &:= \frac{c_4}{2} \int_{\Omega} (u - \bar{u}_0)^2 + \frac{c_4}{2} \int_{\Omega} (v - \bar{v}_0)^2 \\ &\quad + \frac{1}{2} \int_{\Omega} (w - f(\bar{v}_0))^2 + \frac{1}{2} \int_{\Omega} (z - g(\bar{u}_0))^2 + \frac{1}{2} \int_{\Omega} |\nabla w|^2 + \frac{1}{2} \int_{\Omega} |\nabla z|^2, \end{aligned}$$

for  $t \geq 0$  fulfills

$$\begin{aligned} y' &\leq - \left( \frac{\min\{D_u, D_v\}c_4}{4c_1} - \frac{c_2^2}{2} \right) \left( \int_{\Omega} (u - \bar{u}_0)^2 + \int_{\Omega} (v - \bar{v}_0)^2 \right) \\ &\quad - \left( \frac{\min\{D_u, D_v\}c_4}{2} - \frac{c_2^2}{2} \right) \left( \int_{\Omega} |\nabla u|^2 + \int_{\Omega} |\nabla v|^2 \right) \\ &\quad - \frac{1}{2} \left( \int_{\Omega} (w - f(\bar{v}_0))^2 + \int_{\Omega} (z - g(\bar{u}_0))^2 \right) \\ &\quad - \left( \frac{1}{2} - M^2 c_3^2 c_4 \max \left\{ \frac{\chi_u^2}{D_u}, \frac{\chi_v^2}{D_v} \right\} \right) \left( \int_{\Omega} |\nabla w|^2 + \int_{\Omega} |\nabla z|^2 \right) \\ &\leq -c_5 y \quad \text{in } (0, \infty), \end{aligned}$$

where  $c_5 := \min\{\frac{c_2^2}{2c_4}, \frac{1}{2}\}$ . Thus,  $y(t) \leq e^{-c_5 t}y(0) \rightarrow 0$  for  $t \rightarrow \infty$ . This entails (1.8) and (1.9) for  $p = 2$ , upon which the claims for  $p \in (2, \infty)$  follow from (4.3) and the interpolation inequality  $\|\varphi\|_{L^p(\Omega)} \leq \|\varphi\|_{L^\infty(\Omega)}^{\frac{p-2}{p}} \|\varphi\|_{L^2(\Omega)}^{\frac{2}{p}}$ , valid for all  $\varphi \in L^\infty(\Omega)$ .  $\square$

#### 4.2. Lack of smooth heterogeneous steady states

We now show that all smooth steady states of (1.1), that is, solutions to (1.10), are spatially homogeneous, provided  $f$  and  $g$  belong to  $\bigcup_{\varepsilon>0} C^\omega((-\varepsilon, \infty); [0, \infty))$ . In a rather straightforward manner, this result can be extended to wider classes of functions  $f$  and  $g$ , but as the prototypical choices mentioned in the introduction are covered by Proposition 1.1, we confine ourselves to analytical functions  $f$  and  $g$ , for which the proof is particularly short.

**Proof of Proposition 1.1.** Inserting the third and fourth equation in (1.10) into the first two and the last equations therein yields

$$\begin{cases} 0 = \Delta u + \xi_u \nabla \cdot (u \nabla f(v)) & \text{in } \Omega, \\ 0 = \Delta v + \xi_v \nabla \cdot (v \nabla g(u)) & \text{in } \Omega, \\ \partial_\nu u + \xi_u u \partial_\nu f(v) = 0 & \text{on } \partial\Omega, \\ \partial_\nu v + \xi_v v \partial_\nu g(u) = 0 & \text{on } \partial\Omega, \end{cases}$$

where we have again set  $\xi_u := \frac{X_u}{D_u}$  and  $\xi_v := \frac{X_v}{D_v}$ . Due to the supposed regularity of  $u$  and  $v$ , Lemma 4.1 in Ref. 12 asserts that there are  $\alpha, \beta \in \mathbb{R}$  such that

$$u = \alpha e^{-\xi_u f(v)} \quad \text{and} \quad v = \beta e^{-\xi_v g(u)} \quad \text{in } \bar{\Omega}, \tag{4.6}$$

so that in particular

$$u = \alpha e^{-\xi_u f(\beta e^{-\xi_v g(u)})} \quad \text{in } \bar{\Omega}. \tag{4.7}$$

Both  $(-\varepsilon, \infty) \ni s \mapsto \alpha e^{-\xi_u f(\beta e^{-\xi_v g(s)})}$  and  $(-\varepsilon, \infty) \ni s \mapsto s$  are analytical functions on  $(-\varepsilon, \infty)$  for some  $\varepsilon > 0$ . As the former only vanishes at 0 (and then everywhere) if  $\alpha = 0$ , these functions differ. Therefore, the set  $S = \{s \in (-\varepsilon, \infty) \mid s = \alpha e^{-\xi_u f(\beta e^{-\xi_v g(s)})}\}$  is discrete by the identity theorem. According to (4.7), the image of the continuous function  $u: \bar{\Omega} \rightarrow \mathbb{R}$  is contained in  $S$ , which is only possible if  $u$  is constant. Recalling (4.6), we conclude that  $v$  is constant, whenceupon the third and fourth equations in (1.10) direct imply that also  $w$  and  $z$  are constant.  $\square$

## 5. Numerical Method

### 5.1. Galerkin approximation

In this section, we will give the details on the implicit finite element discretization for solving the problem (1.1) numerically. A finite element discretization of the system (1.1) is based on its weak formulation, which reads: Find  $u, v, w, z \in$

$L^\infty(0, T; L^2(\Omega)) \cap L^2(0, T; H^1(\Omega))$  with  $u_t, v_t, w_t, z_t \in L^2(0, T; (H^1(\Omega))^*)$  for given  $u^0, v^0, w^0, z^0 \in L^2(\Omega)$  such that a.e. in  $(0, \infty)$  we have

$$\begin{aligned} \langle u_t, \psi \rangle &= -D_u \int_{\Omega} \nabla u \cdot \nabla \psi dx - \chi_u \int_{\Omega} u \nabla w \cdot \nabla \psi dx, \\ \langle v_t, \psi \rangle &= -D_v \int_{\Omega} \nabla v \cdot \nabla \psi dx - \chi_v \int_{\Omega} v \nabla z \cdot \nabla \psi dx, \\ \langle w_t, \psi \rangle &= - \int_{\Omega} w \psi dx + \int_{\Omega} f(v) \psi dx, \\ \langle z_t, \psi \rangle &= - \int_{\Omega} z \psi dx + \int_{\Omega} g(u) \psi dx, \end{aligned} \tag{5.1}$$

for all  $\psi \in C^\infty(\overline{\Omega})$ . Here,  $\langle \cdot, \cdot \rangle$  represents the duality pairing between  $(H^1(\Omega))^*$  and  $H^1(\Omega)$ .

To define a finite element discretization of problem (1.1), we first consider  $\mathcal{T}_h$ , a uniformly regular triangulation of  $\Omega$  with a mesh size  $h$ . Then we construct the finite element space  $X_h$  consisting of continuous piecewise multi-linear functions

$$X_h = \{ \phi \in H^1(\Omega); \phi|_K \in Q_1(K), \forall K \in \mathcal{T}_h \}, \quad Z_h = X_h \cap H_0^1(\Omega),$$

with basis functions  $\psi_j, j = 1, \dots, M$ , such that  $X_h = \text{span}\{\psi_j\}$ , where  $M$  is the number of degrees of freedom and  $Q_1$  is a space consisting of piecewise multi-linear functions. Any function  $u_h \in X_h$  can be written in a unique way with respect to these basis functions as

$$u_h = \sum_{j=1}^M u_j \psi_j,$$

and hence it can be identified with the coefficient vector  $\mathbf{u} = (u_1, \dots, u_M)$ ;  $v_h, w_h$ , and  $z_h$  can be defined similarly. Next, let  $0 = t_0 < t_1 < \dots < t_N = T$  be an equidistant decomposition of the time interval  $[0, T]$  with  $\Delta t = t^{n+1} - t^n, n = 0, \dots, N - 1$ . We use  $u_h^n, v_h^n, w_h^n, z_h^n \in X_h$  to denote the approximation of the solutions at each time level  $t^n$ . Furthermore, an important feature of the considered system consists in the nonlinear terms and coupling between the equations, so that a fully implicit discretization leads to a coupled nonlinear algebraic system. We compute the solution of this nonlinear problem using fixed-point iterations. As a result, after applying the usual approach for deriving a Galerkin finite element scheme for space discretization, considering the  $\theta$ -method for time discretization, and using a fixed-point iteration to treat the nonlinear terms in the system the linearized algebraic form corresponding to the system (5.1) reads as

$$(\mathbb{M} + \theta \Delta t \mathbb{A}_{k-1}^{n+1, u}) \mathbf{u}_k^{n+1} = (\mathbb{M} - (1 - \theta) \Delta t \mathbb{A}^{n, u}) \mathbf{u}^n, \tag{5.2}$$

$$(\mathbb{M} + \theta \Delta t \mathbb{A}_{k-1}^{n+1, v}) \mathbf{v}_k^{n+1} = (\mathbb{M} - (1 - \theta) \Delta t \mathbb{A}^{n, v}) \mathbf{v}^n, \tag{5.3}$$

$$(1 + \theta \Delta t)\mathbb{M} \mathbf{w}_k^{n+1} = (1 - (1 - \theta)\Delta t)\mathbb{M} \mathbf{w}^n + \Delta t (\theta G_k^{n+1,v} + (1 - \theta)G^{n,v}), \tag{5.4}$$

$$(1 + \theta \Delta t)\mathbb{M} \mathbf{z}_k^{n+1} = (1 - (1 - \theta)\Delta t)\mathbb{M} \mathbf{z}^n + \Delta t (\theta G_k^{n+1,u} + (1 - \theta)G^{n,u}), \tag{5.5}$$

for  $\theta \in [0, 1]$ , where  $\mathbf{u}_k^{n+1} = (u_{j,k}^{n+1})_{j=1}^M$ ,  $\mathbf{v}_k^{n+1} = (v_{j,k}^{n+1})_{j=1}^M$ ,  $\mathbf{w}_k^{n+1} = (w_{j,k}^{n+1})_{j=1}^M$  and  $\mathbf{z}_k^{n+1} = (z_{j,k}^{n+1})_{j=1}^M$  denote the vectors of unknowns at time level  $t^{n+1}$  and iteration step  $k, k = 1, \dots$ , and  $\mathbf{u}^n = (u_j^n)_{j=1}^M$ ,  $\mathbf{v}^n = (v_j^n)_{j=1}^M$ ,  $\mathbf{w}^n = (w_j^n)_{j=1}^M$  and  $\mathbf{z}^n = (z_j^n)_{j=1}^M$  are the known solutions from the previous time level  $t^n$ . Setting  $\mathbf{u}_0^{n+1} = \mathbf{u}^n$ ,  $\mathbf{v}_0^{n+1} = \mathbf{v}^n$ ,  $\mathbf{w}_0^{n+1} = \mathbf{w}^n$  and  $\mathbf{z}_0^{n+1} = \mathbf{z}^n$  with  $\mathbf{u}^0 = \mathbf{u}(x, 0)$ ,  $\mathbf{v}^0 = \mathbf{v}(x, 0)$ ,  $\mathbf{w}^0 = \mathbf{w}(x, 0)$  and  $\mathbf{z}^0 = \mathbf{z}(x, 0)$ , the entries of the mass matrix, stiffness matrices and vectors above are given by

$$\begin{aligned} \mathbb{M}_{ij} &= \int_{\Omega} \psi_j \psi_i, \\ \mathbb{A}_{ij,k-1}^{n+1,u} &= D_u \int_{\Omega} \nabla \psi_j \cdot \nabla \psi_i + \chi_u \int_{\Omega} \psi_j \nabla w_{h,k-1}^{n+1} \cdot \nabla \psi_i, \\ \mathbb{A}_{ij}^{n,u} &= D_u \int_{\Omega} \nabla \psi_j \cdot \nabla \psi_i + \chi_u \int_{\Omega} \psi_j \nabla w_h^n \cdot \nabla \psi_i, \\ \mathbb{A}_{ij,k-1}^{n+1,v} &= D_v \int_{\Omega} \nabla \psi_j \cdot \nabla \psi_i + \chi_v \int_{\Omega} \psi_j \nabla z_{h,k-1}^{n+1} \cdot \nabla \psi_i, \\ \mathbb{A}_{ij}^{n,v} &= D_v \int_{\Omega} \nabla \psi_j \cdot \nabla \psi_i + \chi_v \int_{\Omega} \psi_j \nabla z_h^n \cdot \nabla \psi_i, \\ G_{i,k}^{n+1,v} &= \int_{\Omega} f(v_{h,k}^{n+1}) \psi_i, \\ G_i^{n,v} &= \int_{\Omega} f(v_h^n) \psi_i, \\ G_{i,k}^{n+1,u} &= \int_{\Omega} g(u_{h,k}^{n+1}) \psi_i, \\ G_i^{n,u} &= \int_{\Omega} g(u_h^n) \psi_i \end{aligned}$$

for  $i, j = 1, \dots, M$  where  $\psi_i \in Z_h$  and  $u_{h,k}^{n+1}, v_{h,k}^{n+1}, w_{h,k}^{n+1}, z_{h,k}^{n+1}$  denote the fully discrete solution functions at fixed-point step  $k = 1, 2, \dots$

### 5.2. Positivity-preserving FCT scheme

Since all components of the system (1.1) represent densities of populations or graffiti, they should be non-negative, and Lemma 2.1 asserts that the analytical solutions considered in the first part of this paper are indeed non-negative. Therefore, the numerical solutions of the model problem must also be non-negative in order

to satisfy the physics behind the system. Thus, the numerical methods should be constructed in such a way that the qualitative properties of the exact solutions are preserved. The standard Galerkin discretization described in the previous section usually produces oscillatory nonpositive solutions, which leads to numerical instabilities, especially when the convective part of the system is dominant. Therefore, a stabilization has to be applied. One appropriate possibility is to modify the algebraic system resulting from the Galerkin discretization, which will be discussed in the following.

The numerical behavior of the gang concentrations  $u$  and  $v$  heavily depends on the properties of the matrices on the left- and right-hand sides of (5.2) and (5.3). In the following, we will introduce a positivity-preserving FCT scheme following the work of Kuzmin<sup>40–42</sup> which can preserve the positivity of the concentrations  $u$  at all times, the positivity of  $v$  can be obtained similarly by repeating the same process.

**Definition 5.1.** A matrix  $\mathbb{A}$  is called a Z-matrix if it has only nonpositive off-diagonal entries, monotone if  $\mathbb{A}^{-1} \geq 0$  and an M-matrix if it is a monotone Z-matrix.

**Lemma 5.1.** Consider a fully discrete system of the form

$$\mathbb{B} \mathbf{u}^{n+1} = \mathbb{K} \mathbf{u}^n, \tag{5.6}$$

and suppose that the coefficients of  $\mathbb{B} = (b_{ij})_{i,j=1}^M$  and  $\mathbb{K} = (k_{ij})_{i,j=1}^M$  satisfy

$$b_{ii} \geq 0, \quad k_{ii} \geq 0, \quad b_{ij} \leq 0, \quad k_{ij} \geq 0, \quad \forall i, j = 1, \dots, M, \quad i \neq j.$$

If  $\mathbb{B}$  is strictly or irreducibly diagonally dominant, then it is an M-matrix and

- (1) the scheme (5.6) is globally positivity-preserving, i.e.  $\mathbf{u}^{n+1} \geq 0$  if  $\mathbf{u}^n \geq 0$ ,
- (2) the global discrete maximum principle (DMP) holds if  $\sum_j b_{ij} = \sum_j k_{ij} \forall i = 1, \dots, M$ , i.e.

$$(\min \mathbf{u}^n)^- \leq \mathbf{u}^{n+1} \leq (\max \mathbf{u}^n)^+,$$

where  $(\max \mathbf{u}^n)^+ = \max\{0, \mathbf{u}^n\}$  and  $(\min \mathbf{u}^n)^- = \min\{0, \mathbf{u}^n\}$ ,  $n = 0, \dots, N-1$ .

**Proof.** See Theorem 4 in Ref. 41. □

For the system (5.2), which can be written as (5.6), the above properties do not hold since the mass matrix  $\mathbb{M}$  may contain some non-negative off-diagonal entries and also some positive off-diagonal entries might appear in the stiffness matrices. Therefore, as a remedy, following the work of Kuzmin,<sup>40–42</sup> we not only replace the mass matrix  $\mathbb{M}$  by its diagonal counterpart, the lump matrix  $\mathbb{M}_L$

$$\mathbb{M}_L = \text{diag}(m_1, \dots, m_M), \quad m_i = \sum_{j=1}^M m_{ij}, \quad i = 1, \dots, M,$$

but also add symmetric artificial diffusion matrices  $\mathbb{D}_{k-1}^{n+1,u} = (d_{ij,k-1}^{n+1,u})_{i,j=1}^M$  and  $\mathbb{D}^{n,u} = (d_{ij}^{n,u})_{i,j=1}^M$  to the stiffness matrices  $\mathbb{A}_{k-1}^{n+1,u}$  and  $\mathbb{A}^{n,u}$  to eliminate their non-negative off-diagonal entries as

$$d_{ij,k-1}^{n+1,u} = -\max\{a_{ij,k-1}^{n+1,u}, 0, a_{ji,k-1}^{n+1,u}\} \quad \text{for } i \neq j, \quad d_{ii,k-1}^{n+1,u} = -\sum_{j=1, j \neq i}^M d_{ij,k-1}^{n+1,u},$$

$n = 0, \dots, N - 1, k = 1, 2, \dots$  and

$$d_{ij}^{n,u} = -\max\{a_{ij}^{n,u}, 0, a_{ji}^{n,u}\} \quad \text{for } i \neq j, \quad d_{ii}^{n,u} = -\sum_{j=1, j \neq i}^M d_{ij}^{n,u},$$

$n = 0, \dots, N - 1.$

Denoting  $\tilde{\mathbb{A}}_{k-1}^{n+1,u} = \mathbb{A}_{k-1}^{n+1,u} + \mathbb{D}_{k-1}^{n+1,u}$  and  $\tilde{\mathbb{A}}^{n,u} = \mathbb{A}^{n,u} + \mathbb{D}^{n,u}$ , the low-order form of the (5.2) can be written as

$$(\mathbb{M}_L + \theta \Delta t \tilde{\mathbb{A}}_{k-1}^{n+1,u}) \mathbf{u}_k^{n+1} = (\mathbb{M}_L - (1 - \theta) \Delta t \tilde{\mathbb{A}}^{n,u}) \mathbf{u}^n, \quad n = 0, \dots, N - 1. \quad (5.7)$$

**Lemma 5.2.** *Let the time step  $\Delta t$  satisfy*

$$m_i - (1 - \theta) \Delta t \tilde{a}_{ii}^{n,u} \geq 0, \quad m_i + \theta \Delta t \sum_{j=1}^M a_{ij,k-1}^{n+1,u} > 0, \quad (5.8)$$

$n = 0, \dots, N - 1, k = 1, 2, \dots$ , then the low-order scheme (5.7) is positivity-preserving.

**Proof.** Denoting  $\mathbb{K} = (\mathbb{M}_L - (1 - \theta) \Delta t \tilde{\mathbb{A}}^{n,u})$ , since  $\mathbb{M}_L$  is diagonal and  $\tilde{\mathbb{A}}^{n,u}$  is a Z-matrix the off-diagonal entries of  $\mathbb{K}$  are non-negative and it also has non-negative diagonal entries if the first condition in (5.8) is satisfied, thus  $\mathbb{K} \geq 0$ . Now, set  $\mathbb{B} = (\mathbb{M}_L + \theta \Delta t \tilde{\mathbb{A}}_{k-1}^{n+1,u})$ , since  $\mathbb{D}_{k-1}^{n+1,u}$  has zero row sum, if the second condition in (5.8) holds we can conclude that  $\mathbb{B}$  is strictly diagonally dominant and hence nonsingular. Furthermore  $\mathbb{B}$  is a matrix of non-negative type hence it is an M-matrix, therefore according to Lemma 5.1, (5.7) is positivity-preserving.  $\square$

**Remark 5.1.** The scheme (5.7) can be simplified as

$$\mathbb{B} \mathbf{u}^{n+1} = \mathbb{K} \mathbf{u}^n,$$

where  $\mathbb{B}, \mathbb{K} \in \mathbb{R}^{M \times M}$  and  $\mathbf{u}^{n+1}, \mathbf{u}^n \in \mathbb{R}^M$ . Assuming that

$$\mathbb{B}^{-1} \geq 0, \quad \mathbb{K} \geq 0, \quad \mathbb{B} \mathbb{I}_M \geq \mathbb{K} \mathbb{I}_M,$$

where  $\mathbb{I}_M$  denotes a vector with all entries equal to 1, one obtains with  $M = (\max \mathbf{u}^n)^+$  and  $m = (\min \mathbf{u}^n)^-$ ,  $n = 0, \dots, N - 1$  that

$$\mathbf{u}^{n+1} = \mathbb{B}^{-1} \mathbb{K} \mathbf{u}^n \leq M \mathbb{B}^{-1} \mathbb{K} \mathbb{I}_M \leq M \mathbb{B}^{-1} \mathbb{B} \mathbb{I}_M = M \mathbb{I}_M,$$

$$\mathbf{u}^{n+1} = \mathbb{B}^{-1} \mathbb{K} \mathbf{u}^n \geq m \mathbb{B}^{-1} \mathbb{K} \mathbb{I}_M \geq m \mathbb{B}^{-1} \mathbb{B} \mathbb{I}_M = m \mathbb{I}_M,$$

thus that

$$(\min \mathbf{u}^n)^- \leq \mathbf{u}^{n+1} \leq (\max \mathbf{u}^n)^+, \quad i = 1, \dots, M,$$

i.e. that the DMP is satisfied. We would like to notice that due to the construction and following Lemma 5.2 the conditions  $\mathbb{B}^{-1} \geq 0$  ( $\mathbb{B}$  is an M-matrix and thus invertible) and  $\mathbb{K} \geq 0$  already hold. However, without further assumptions on the matrices  $\tilde{\mathbb{A}}_{k-1}^{n+1,u}$  and  $\tilde{\mathbb{A}}^{n,u}$ , proving the last assumption above (which is a replacement of the row-sum property in Lemma 5.1) is very challenging, see Ref. 7 for more information.

Although the solution of (5.7) suppresses the spurious oscillations and does not produce negative solutions, it often becomes inaccurate since the amount of the added artificial diffusion is usually too large. Therefore, the idea of FEM-FCT is to modify the right-hand side of (5.7) in such a way that the solutions become less diffusive in the smooth regions while their positivity is still conserved. By construction, the difference between (5.2) and (5.7) reads

$$(\mathbb{M}_L - \mathbb{M})(\mathbf{u}_k^{n+1} - \mathbf{u}^n) + \theta \Delta t \mathbb{D}_{k-1}^{n+1,u} \mathbf{u}_k^{n+1} + (1 - \theta) \Delta t \mathbb{D}^{n,u} \mathbf{u}^n, \quad (5.9)$$

$n = 0, \dots, N - 1, k = 1, 2, \dots$ , which is nonlinear since it depends on the approximate solution  $\mathbf{u}_k^{n+1}$ . To be aligned with treating nonlinearities in Eq. (5.2) using fixed-point iteration, we replace  $u_k^{n+1}$  by  $u_{k-1}^{n+1}$ , which leads to

$$\bar{\mathbf{f}}_{k-1}^{n+1,u} := (\mathbb{M}_L - \mathbb{M})(\mathbf{u}_{k-1}^{n+1} - \mathbf{u}^n) + \theta \Delta t \mathbb{D}_{k-1}^{n+1,u} \mathbf{u}_{k-1}^{n+1} + (1 - \theta) \Delta t \mathbb{D}^{n,u} \mathbf{u}^n,$$

$n = 0, \dots, N - 1$ , which admits a decomposition into a sum of discrete internodal fluxes

$$\bar{\mathbf{f}}_{k-1}^{n+1,u} = \sum_{j=1}^M f_{ij,k-1}^{n+1,u}, \quad f_{ij,k-1}^{n+1,u} = -f_{ji,k-1}^{n+1,u}, \quad i, j = 1, \dots, M, \quad (5.10)$$

since the matrices  $\mathbb{D}_{k-1}^{n+1,u}, \mathbb{D}^{n,u}$  and  $(\mathbb{M}_L - \mathbb{M})$  are symmetric and have zero row sums. Moreover, the so-called algebraic fluxes  $f_{ij,k-1}^{n+1,u}$  are given by

$$f_{ij,k-1}^{n+1,u} = (-m_{ij} + \theta \Delta t d_{ij,k-1}^{n+1,u})(u_{j,k-1}^{n+1} - u_{i,k-1}^{n+1}) + (m_{ij} + (1 - \theta) \Delta t d_{ij}^{n,u})(u_j^n - u_i^n).$$

Now, the amount of algebraic fluxes inserted into each node must be limited by a limiter  $\alpha_{ij,k-1}^{n+1,u} \in [0, 1]$  in such a way that in addition to keeping the numerical solutions positive, it also controls the amount of added artificial diffusion especially in the regions where the solutions are smooth and well resolved where  $\alpha_{ij,k-1}^{n+1,u} = 1$  is appropriate, therefore one replaces (5.10) by

$$\mathbf{f}_{k-1}^{n+1,u} = \sum_{j=1}^M \alpha_{ij,k-1}^{n+1,u} f_{ij,k-1}^{n+1,u}, \quad \alpha_{ij,k-1}^{n+1,u} = \alpha_{ji,k-1}^{n+1,u}, \quad i, j = 1, \dots, M,$$

this leads to

$$(\mathbb{M}_L + \theta \Delta t \tilde{\mathbb{A}}_{k-1}^{n+1,u}) \mathbf{u}_k^{n+1} = (\mathbb{M}_L - (1 - \theta) \Delta t \tilde{\mathbb{A}}^{n,u}) \mathbf{u}^n + \mathbf{f}_{k-1}^{n+1,u}, \quad (5.11)$$

$n = 0, \dots, N - 1, k = 1, 2, \dots$  and it can be rewritten in the form

$$\mathbb{M}_L \bar{\mathbf{u}} = (\mathbb{M}_L - (1 - \theta) \Delta t \tilde{\mathbb{A}}^{n,u}) \mathbf{u}^n,$$

$$\mathbb{M}_L \tilde{\mathbf{u}} = \mathbb{M}_L \bar{\mathbf{u}} + \mathbf{f}_{k-1}^{n+1,u},$$

$$(\mathbb{M}_L + \theta \Delta t \tilde{\mathbb{A}}_{k-1}^{n+1,u}) \mathbf{u}_k^{n+1} = \mathbb{M}_L \tilde{\mathbf{u}},$$

which is a high-resolution finite element scheme. Moreover, it is positivity-preserving under the conditions (5.8) and for appropriate choice of flux limiters.<sup>32</sup>

We use the limiting strategy based on Zalesak's algorithm<sup>66</sup> to determine an appropriate value of  $\alpha_{ij,k-1}^{n+1,u}$ . The limiting process begins with canceling all fluxes that are diffusive in nature and tend to flatten the solutions profile.<sup>41</sup> The required modification is

$$f_{ij,k-1}^{n+1,u} := 0 \quad \text{if } f_{ij,k-1}^{n+1,u} (\bar{u}_j - \bar{u}_i) > 0.$$

To define the limiter, we perform the following steps:

- (1) Compute the sum of positive/negative antidiffusive fluxes into node  $i$

$$P_i^+ = \sum_{j \in \mathcal{N}_i} \max\{0, f_{ij,k-1}^{n+1,u}\}, \quad P_i^- = \sum_{j \in \mathcal{N}_i} \min\{0, f_{ij,k-1}^{n+1,u}\}, \quad (5.12)$$

where  $\mathcal{N}_i$  is the set of the nearest neighbors of the node  $i$ .

- (2) Compute the distance to a local extremum of the auxiliary solution  $\bar{\mathbf{u}}$  at the neighboring nodes that share an edge with the node  $i$

$$Q_i^+ = m_i(\bar{u}_i^{\max} - \bar{u}_i), \quad Q_i^- = m_i(\bar{u}_i^{\min} - \bar{u}_i), \quad (5.13)$$

with  $\bar{u}_i^{\min} = \min_{j \in \mathcal{N}_i \cup \{i\}} \bar{u}_j, \bar{u}_i^{\max} = \max_{j \in \mathcal{N}_i \cup \{i\}} \bar{u}_j, i = 1, \dots, M$ .

- (3) Compute the nodal correction factors for the net increment at the node  $i$

$$R_i^+ = \min \left\{ 1, \frac{Q_i^+}{P_i^+} \right\}, \quad R_i^- = \min \left\{ 1, \frac{Q_i^-}{P_i^-} \right\}, \quad (5.14)$$

if  $P_i^+$  or  $P_i^-$  vanishes then set  $R_i^+ = 1$  or  $R_i^- = 1$ , respectively.

- (4) Check the sign of the antidiffusive flux and apply the correction factor by

$$\alpha_{ij} = \begin{cases} \min\{R_i^+, R_j^-\} & \text{if } f_{ij,k-1}^{n+1,u} > 0, \\ 1 & \text{if } f_{ij,k-1}^{n+1,u} = 0, \\ \min\{R_i^-, R_j^+\} & \text{if } f_{ij,k-1}^{n+1,u} < 0. \end{cases} \quad (5.15)$$

Another way how to derive (5.11) is to apply the FCT stabilization to the non-linear problem obtained by applying the Galerkin discretization and the  $\theta$ -method before introducing the fixed-point iterations. The algebraic fluxes defined in this way depend on the unknown solution  $u$  and hence the FCT stabilization introduces an additional nonlinearity. However, since the problem at hand is already



---

**Algorithm 1.** FEM-FCT.

---

```

for time step  $n \leftarrow 0, \dots, N$  do
  for iteration step  $k = 1, 2, \dots$  do
    solve for  $\mathbf{u}_k^{n+1}$ 
      Solve  $\mathbb{M}_L \bar{\mathbf{u}} = (\mathbb{M}_L - (1 - \theta)\Delta t \tilde{\mathbb{A}}^{n,u})\mathbf{u}^n$ 
      Compute  $\mathbf{f}_{k-1}^{n+1,u}$  using Zalesak's algorithm (5.12)–(5.15)
      Solve  $\mathbb{M}_L \tilde{\mathbf{u}} = \mathbb{M}_L \bar{\mathbf{u}} + \mathbf{f}_{k-1}^{n+1,u}$ 
      Solve  $(\mathbb{M}_L + \theta \Delta t \tilde{\mathbb{A}}_{k-1}^{n+1,u})\mathbf{u}_k^{n+1} = \mathbb{M}_L \tilde{\mathbf{u}}$ 
    end solve
    solve for  $\mathbf{v}_k^{n+1}$ 
      Solve  $\mathbb{M}_L \bar{\mathbf{v}} = (\mathbb{M}_L - (1 - \theta)\Delta t \tilde{\mathbb{A}}^{n,v})\mathbf{v}^n$ 
      Compute  $\mathbf{f}_{k-1}^{n+1,v}$  using Zalesak's algorithm (5.12)–(5.15)
      Solve  $\mathbb{M}_L \tilde{\mathbf{v}} = \mathbb{M}_L \bar{\mathbf{v}} + \mathbf{f}_{k-1}^{n+1,v}$ 
      Solve  $(\mathbb{M}_L + \theta \Delta t \tilde{\mathbb{A}}_{k-1}^{n+1,v})\mathbf{v}_k^{n+1} = \mathbb{M}_L \tilde{\mathbf{v}}$ 
    end solve
    solve for  $\mathbf{w}_k^{n+1}$ 
      Compute  $\mathbf{w}_k^{n+1}$  from (5.4)
    end solve
    solve for  $\mathbf{z}_k^{n+1}$ 
      Compute  $\mathbf{z}_k^{n+1}$  from (5.5)
    end solve
  end for
end for

```

---

nonlinear, both nonlinearities can be handled simultaneously using a fixed-point iteration. Moreover, the procedure explained above can also be used to obtain a high-resolution positivity-preserving approximation of  $v$ .

We summarize the procedure above in the following algorithm.

We iterate until the residual/difference between two successive solutions is less than a prescribed tolerance or until the maximum number of the iterations is reached, then set  $\mathbf{u}^{n+1} = \mathbf{u}_k^{n+1}$ ,  $\mathbf{v}^{n+1} = \mathbf{v}_k^{n+1}$ ,  $\mathbf{w}^{n+1} = \mathbf{w}_k^{n+1}$  and  $\mathbf{z}^{n+1} = \mathbf{z}_k^{n+1}$  and advance to the next time level. The system of the algebraic equations was solved using the direct solver UMFPAK<sup>19</sup> and for the implementation of our newly designed algorithm we used the deal.II library.

## 6. Numerical Simulations

In this section, we present numerical experiments for solving the model problem (1.1) with  $f(v) = \frac{v}{1+v}$  and  $g(u) = \frac{u}{1+u}$ , and compare our results with results from the literature<sup>2, 3, 6</sup> pertaining to related systems. The numerical simulations are computed on a square domain  $\Omega = [-6, 6]^2$  that is discretized uniformly using quadrilateral elements with five levels of refinements which creates 1089 degrees of

freedom, i.e. 33 grid points in each direction. In addition, conforming bilinear finite elements are used for all unknown variables. Furthermore, we use the time interval  $[0, T]$ ,  $T = 1000$ , with a time step  $\Delta t = 1.0$  in the Crank–Nicolson discretization method ( $\theta$ -scheme for  $\theta = 0.5$ ). The setting above is fixed through all computations, unless otherwise mentioned.

### 6.1. *Can the gangs separate?*

In this section, we investigate the possibility of segregation for the model problem (1.1). Thus, we consider different values of  $D_u$ ,  $D_v$ ,  $\chi_u$  and  $\chi_v$  in our simulations. The initial conditions used in the following examples are given by

$$\begin{aligned} u^0(x, y) &= 0.1 + e^{-(x-2)^2 - (y-2)^2}, & v^0(x, y) &= 0.1 + e^{-(x+2)^2 - (y+2)^2}, \\ w^0(x, y) &= 0, & z^0(x, y) &= 0, \end{aligned} \tag{6.1}$$

see for instance Fig. 1(a) for an illustration of  $u^0, v^0$ . In our figures, we use the color red when the gang density  $u$  presented in the discrete domain is larger than the gang density  $v$  by at least  $10^{-6}$  and dark blue color when the opposite inequality holds true. Locations where the difference between the gang densities is less than  $10^{-6}$  are displayed in dark purple. Similarly and again with a cutoff of  $10^{-6}$ , we use orange if the amount of graffiti density  $z$  is larger in a region, light blue for the opposite situation and light purple if the graffiti densities  $z$  and  $w$  are roughly the same.

Unlike as in the analytical part, we do not fix  $D_u, D_v, \chi_u, \chi_v$  but on the contrary are interested in their effect on the dynamics. However, if  $(u, v, w, z)$  is a solution of (1.1) with  $f = g = \text{id}$  and initial data  $(u^0, v^0, w^0, z^0)$ , then  $(Au, Bv, Bw, Az)$  solves (1.1) with initial data  $(Au^0, Bu^0, Bw^0, Az^0)$  and  $(\chi_u, \chi_v)$  replaced by  $(\frac{\chi_u}{B}, \frac{\chi_v}{A})$  for any  $A, B > 0$ . That is, instead of considering larger initial data one may likewise consider larger  $\chi_u, \chi_v$ , and one expects that for functions  $f$  and  $g$  with linear growth near 0 these modifications have similar effects as well — at least as long as the solutions remain comparatively small; if they become large this scaling argument may no longer be applicable. In particular, Theorem 1.2 suggests convergence toward homogeneous equilibria for small  $\chi_u, \chi_v$  (and is inconclusive for large  $\chi_u, \chi_v$ ).

#### 6.1.1. *Convergence toward constant steady state*

To begin with, we first consider a case where  $D_u = D_v = \chi_u = \chi_v = 0.25$ . In the top row of Fig. 1, we present the plot of the gang population densities  $u$  and  $v$  and in the bottom row their corresponding graffiti densities  $z$  and  $w$  computed using the standard Galerkin method (5.2)–(5.5) over time. Figure 1(a) corresponds to the gangs' initial conditions at  $t = 0$ ; evolving the time, we observe that the gangs start to spread inside the domain and mark their territories by spraying graffiti. By the time  $t = 5$ , the whole upper triangle part of the domain is completely dominated by the gang  $u$  and the lower triangle part is dominated by the gang  $v$ , this situation

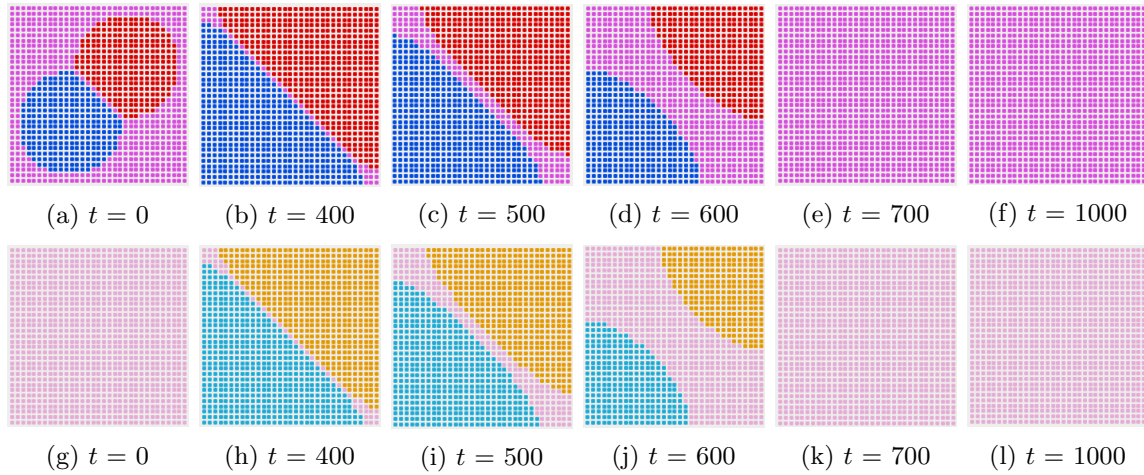


Fig. 1. (Color online) Numerical solutions for the model problem (1.1) obtained using the standard Galerkin method, at different time instant when  $D_u = D_v = \chi_u = \chi_v = 0.25$ . For the choice of colors, see the beginning of Sec. 6.1.

seems to continue till nearly  $t = 400$ , after that we can see that the dark purple starts to merge from the middle section and continues to grow wider to the point that covers the whole domain at  $t = 700$  and remains the same by the end of the time, which means that the same amount of the gang densities  $u$  and  $v$  are presented in each location of the domain. The same happens for their corresponding graffiti  $z$  and  $w$  in the bottom row, where there is no amount of graffiti at the beginning (at  $t = t_0$ ), then they gradually increase when the gangs start to mark their own territories and by the time  $t = 700$  each location is marked by the same amount of graffiti from each gang. It is evident from Fig. 1 that segregation does not happen in this case.

Before moving on to the next example, we examine this case more closely in Figs. 2 and 3, by taking a snapshot along the line  $y = x$  over time. Figure 2(a) shows the initial conditions at  $t = 0$ , when the time evolves it seems that the approximate solutions converge toward constant steady states already at  $t = 400$  in Fig. 2(c), but the close up in Figs. 3(a) and 3(b) shows that the gang density  $u$  and its corresponding graffiti  $z$  are slightly larger in the first half of the domain  $[-6, 0)$  and slightly smaller in the second half  $(0, 6]$  than the gang density  $v$  and

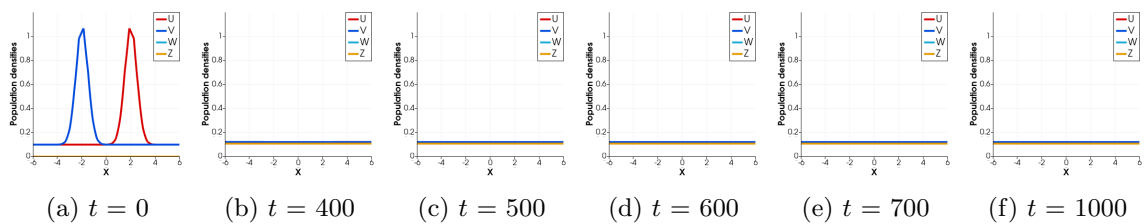


Fig. 2. The size of gang populations  $u, v$  and the amount of their corresponding graffiti  $z$  and  $w$  over time along the line  $y = x$  at different time instant  $t = 0, 400, 500, 600, 700, 1000$  when  $D_u = D_v = \chi_u = \chi_v = 0.25$ .

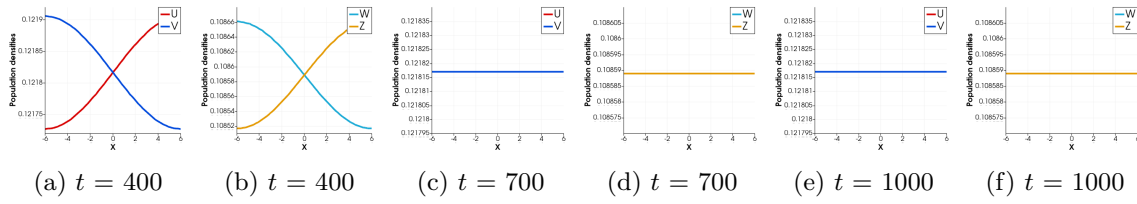


Fig. 3. Close up of populations densities along the line  $y = x$  at different time instant  $t = 400, 700, 1000$  when  $D_u = D_v = \chi_u = \chi_v = 0.25$ .

its corresponding graffiti  $w$ ; however, they tend to stabilize at around 0.121817 and 0.108588, respectively, at time  $t = 700$  and remain the same through the rest of the time, which corresponds well to the results shown in Fig. 1.

This case has been studied for (1.2) in Ref. 6, where the authors also show that the approximate solutions converge toward steady states; however they also report that segregation might not happen for that problem, at least not in the considered solution framework.

Next, we set  $D_u = D_v = 3$  and  $\chi_u = \chi_v = 0.25$  and again use the standard Galerkin method (5.2)–(5.5). Figure 4 shows that, when the time evolves gangs start to spread inside the domain and no accumulation seems to happen, the dark purple in the top row and light purple in the bottom row gradually cover the entire domain by the time  $t = 200$ , i.e. the gangs are evenly distributed in each location and produce the same amount of graffiti. It is clear from Fig. 4 that no separation happens when the problem is diffusion-dominated.

The snapshots of the results are presented along the line  $y = x$  in Figs. 5 and 6, where we observe that the gang densities and their corresponding graffiti converge toward constant steady states and seem to stabilize around 0.121818 and 0.10858, respectively, by the time  $t = 200$ .

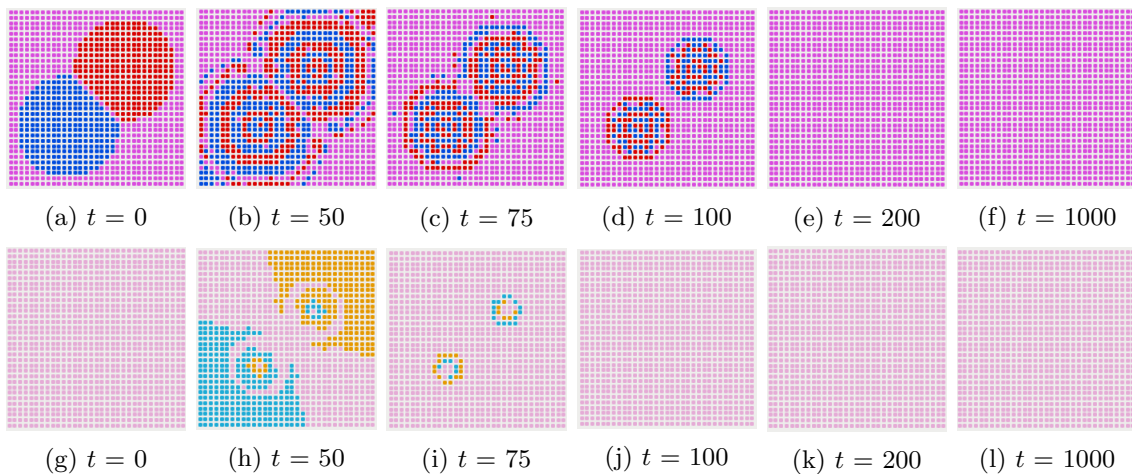


Fig. 4. (Color online) Numerical solutions for the model problem (1.1) obtained using the standard Galerkin method, at different time instant  $t = 0, 50, 75, 100, 200, 1000$  when  $D_u = D_v = 3.0$  and  $\chi_u = \chi_v = 0.25$ . For the choice of colors, see the beginning of Sec. 6.1.

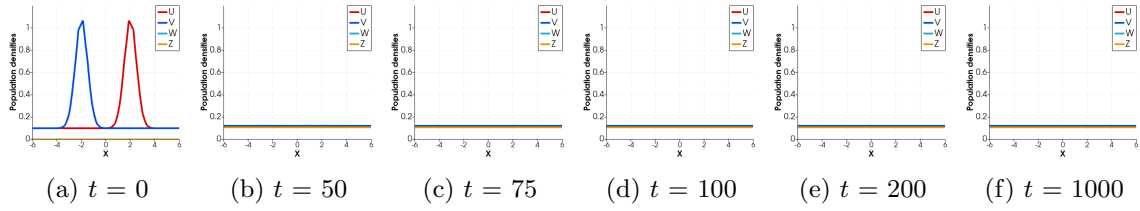


Fig. 5. The amount of gang densities  $u, v$  and their corresponding graffiti  $z$  and  $w$  over time along the line  $y = x$  at different time instant  $t = 0, 50, 75, 100, 200, 1000$  when  $D_u = D_v = 3.0$  and  $\chi_u = \chi_v = 0.25$ .

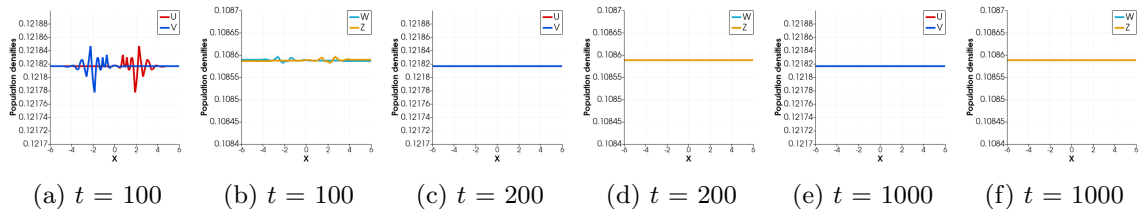


Fig. 6. Close up of population densities along the line  $y = x$  at different time instant  $t = 100, 200, 1000$  when  $D_u = D_v = 3.0$  and  $\chi_u = \chi_v = 0.25$ .

A similar behavior has been observed in Ref. 2 for an agent-based model: As long as a system parameter  $\beta$  corresponding to  $\chi_u, \chi_v$  in (1.1) is sufficiently small, the gangs completely mix inside the domain and the expected densities converge toward constant steady states.

### 6.1.2. Nonhomogeneous limit functions

For our next example, we consider  $D_u = D_v = 0.25$  and  $\chi_u = \chi_v = 3$ . As shown in Fig. 7, the standard Galerkin method (5.2)–(5.5) produces significant spurious oscillation in the entire domain which leads to negative nonphysical approximate solutions, and the simulation blows up at around  $t = 35$ . As explained in the previous section, as a remedy we apply the FEM-FCT scheme to reduce the oscillations and preserve the positivity of the solutions at all time. The top row of Fig. 8 shows the evolutionary movement of the gangs inside the domain. We observe that after a certain amount of time, red and blue clusters start to form, featuring some symmetries which track back to the symmetry of the initial data and the fact that the parameters are the same for both gangs. The same kind of pattern for graffiti densities emerge in the bottom row as each gang marks their own territory. The population densities seem to almost stabilize around  $t = 500$  and remain the same till the end of the time interval  $T = 1000$ .

Keeping the diffusion coefficient as before and changing the amount of convection coefficient to  $\chi_u = \chi_v = 10$ , a different segregation pattern can be seen in Fig. 10. Therefore, we conclude that larger amount of convection in the model leads to directional movement of the gangs and creates partially separated regions. The

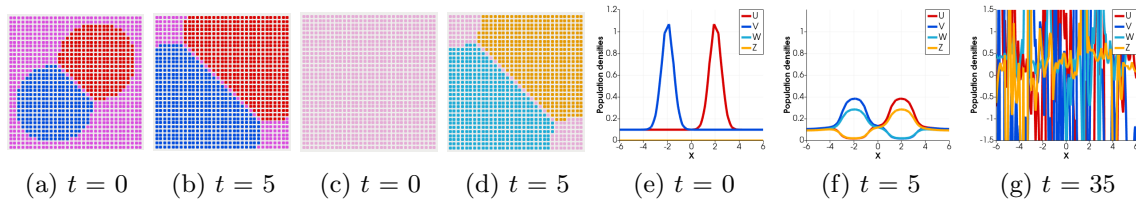


Fig. 7. (Colour online) Numerical solutions and their snapshots for the model problem (1.1) obtained using the standard Galerkin method, at different time instant  $t = 0, 5, 35$  when  $D_u = D_v = 0.25$  and  $\chi_u = \chi_v = 3.0$ .

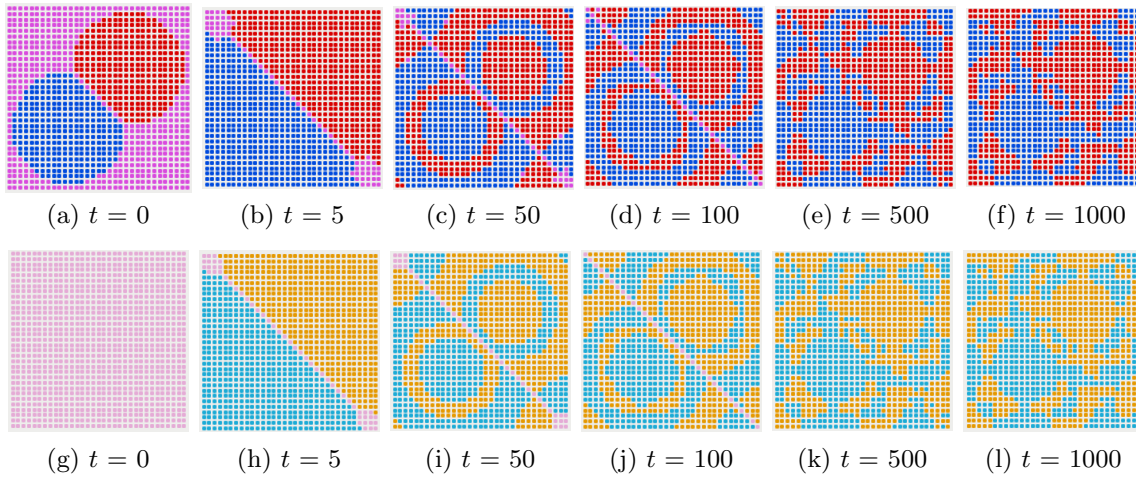


Fig. 8. (Color online) Numerical solutions for the model problem (1.1) obtained using the FEM-FCT method, at different time instant  $t = 0, 5, 50, 100, 500, 1000$  when  $D_u = D_v = 0.25$  and  $\chi_u = \chi_v = 3.0$ . For the choice of colors, see the beginning of Sec. 6.1.

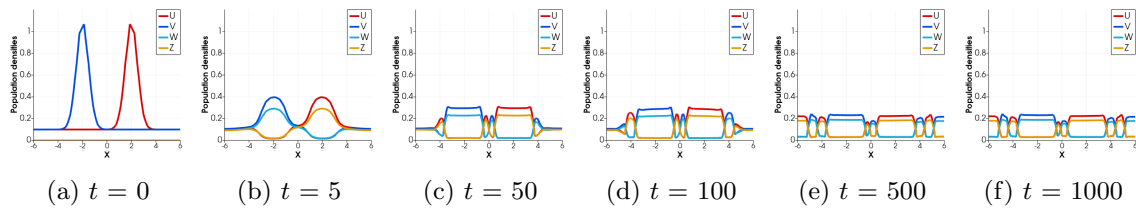


Fig. 9. The size of gang populations  $u, v$  and the amount of their corresponding graffiti  $z$  and  $w$  over time along the line  $y = x$  at different time instant  $t = 0, 5, 50, 100, 500, 1000$  when  $D_u = D_v = 0.25$   $\chi_u = \chi_v = 3.0$ .

snapshots of the results obtained with the FEM-FCT for both cases are displayed in Figs. 9 and 11, respectively.

Comparing our results with the stochastic simulation in the agent-based model in Ref. 2, we notice some similarities between our experiments' outcome and their reported results, namely that large values of  $\beta$  (the convection coefficient in their model) lead to a well-segregated phase. However, in each patch dominated by red or blue in Figs. 8 and 10, there is still some amount of the opposite group with much smaller — but still positive — density present (cf. Figs. 9 and 11), illustrating partial (as opposed to complete) segregation. Moreover, unlike as Ref. 2, we do not

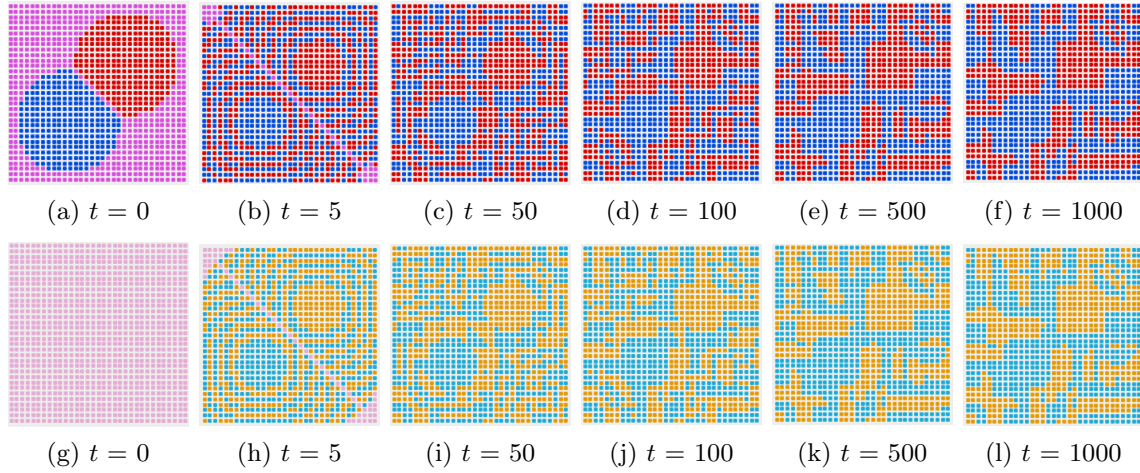


Fig. 10. (Color online) Numerical solutions for the model problem (1.1) obtained using the FEM-FCT method, at different time instant  $t = 0, 5, 50, 100, 500, 1000$  when  $D_u = D_v = 0.25$  and  $\chi_u = \chi_v = 10$ . For the choice of colors, see the beginning of Sec. 6.1.

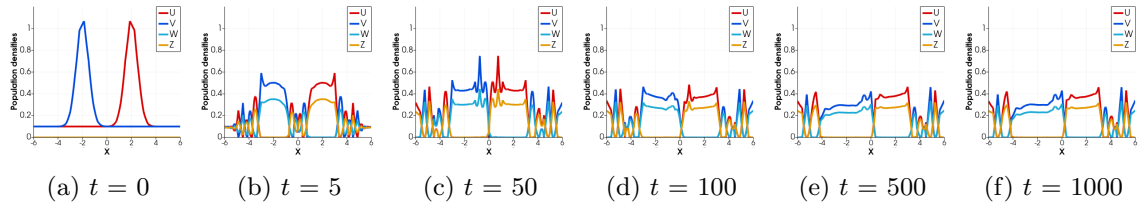


Fig. 11. The size of gang populations  $u, v$  and the amount of their corresponding graffiti  $z$  and  $w$  over time along the line  $y = x$  at different time instant  $t = 0, 5, 50, 100, 500, 1000$  when  $D_u = D_v = 0.25$  and  $\chi_u = \chi_v = 10$ .

observe any coarsening of the regions: Once territories are established, they are rather stable.

Next, we consider fixed diffusion rates as in the previous examples, i.e.  $D_u = D_v = 0.25$ , but different convection coefficients  $\chi_u = 2$  and  $\chi_v = 4$ , meaning that the second gang's movement is more affected by the first gang's graffiti than vice versa. As shown in Fig. 12, the gang with density  $v$  starts to cluster together very tightly in small patches and partially separates from the opposite group at around  $t = 200$  while the other gangs dominates on larger regions. That is, we observe that the segregation pattern for both groups is quite different here, in contrast to the previous examples.

The snapshots of concentration densities over time along the line  $y = x$  are displayed in Fig. 13. As it can be seen, the concentration density  $v$  and its corresponding graffiti  $w$  occupy less parts of the domain but with higher amount of densities (almost around 0.4 and 0.3 in their dominated regions, respectively), however, the gang population  $u$  spreads more freely in the wider range but with less density (nearly under 0.2). Of course, both gangs have the same total amount of mass throughout evolution since there are no source or sink terms in the model.

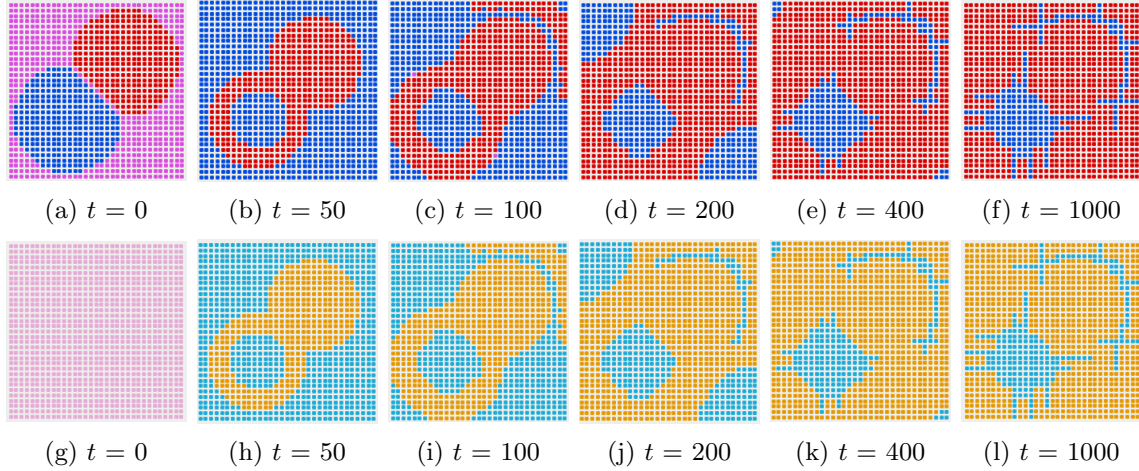


Fig. 12. (Color online) Numerical solutions for the model problem (1.1) obtained using the FEM-FCT method, at different time instant  $t = 0, 50, 100, 200, 400, 1000$  when  $D_u = D_v = 0.25$ ,  $\chi_u = 2.0$  and  $\chi_v = 4.0$ . For the choice of colors, see the beginning of Sec. 6.1.

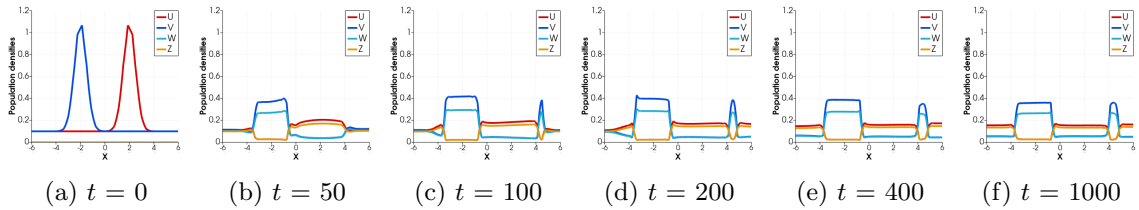


Fig. 13. The size of gang populations  $u, v$  and the amount of their corresponding graffiti  $z$  and  $w$  over time along the line  $y = x$  at different time instant  $t = 0, 50, 100, 200, 400, 1000$  when  $D_u = D_v = 0.25$ ,  $\chi_u = 2.0$  and  $\chi_v = 4.0$ .

In Sec. 5.1 in Ref. 3, a similar behavior, namely that the gang with the largest graffiti avoidance rate concentrates on smaller regions, has been observed for an agent-based model with three rivaling gangs.

### 6.1.3. An example of complete segregation

Finally, we turn to answer the question whether complete separation occurs at all or not; that is, whether the supports of the gang densities can become disjoint in the large-time limit. We show that this is at least possible in settings where the domain is divided into two parts in each of which one gang density is extremely small already at the initial time, and when additionally the diffusion rates are very low, the former prevents mixed amount of population and their interaction in the same regions at the beginning and the latter controls their spread inside the domain over time. For this reason, we consider the case where  $D_u = D_v = 0.01$  and  $\chi_u = \chi_v = 3$  and the initial conditions are given by

$$\begin{aligned}
 u^0(x, y) &= e^{-(x-3)^2 - (y-3)^2}, & v^0(x, y) &= e^{-(x+3)^2 - (y+3)^2}, \\
 w^0(x, y) &= 0, & z^0(x, y) &= 0,
 \end{aligned}
 \tag{6.2}$$



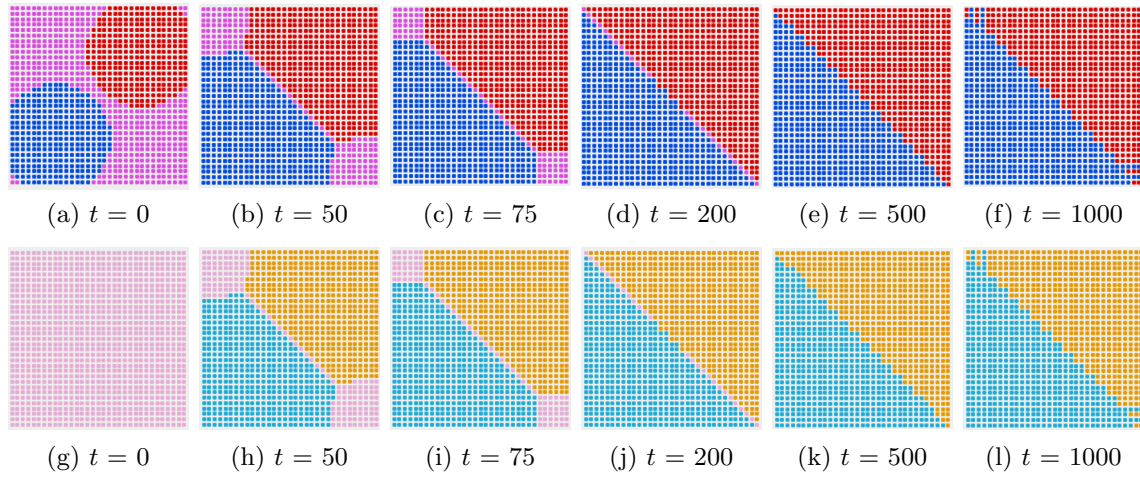


Fig. 14. (Color online) Numerical solutions for the model problem (1.1) obtained using the FEM-FCT method, at different time instant  $t = 0, 50, 75, 200, 500, 1000$  when  $D_u = D_v = 0.01$  and  $\chi_u = \chi_v = 3.0$ . For the choice of colors, see the beginning of Sec. 6.1.

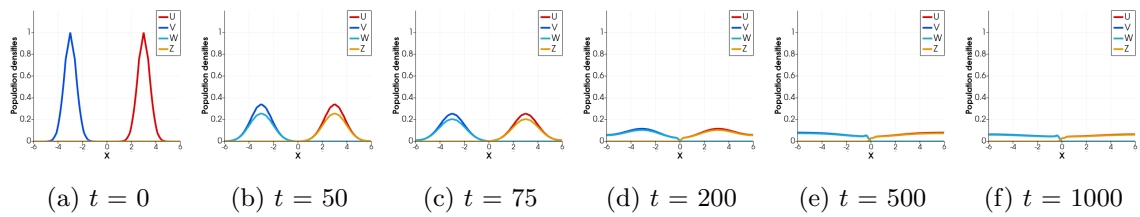


Fig. 15. The size of gang populations  $u, v$  and the amount of their corresponding graffiti  $z$  and  $w$  over time along the line  $y = x$  at different time instant  $t = 0, 50, 75, 200, 500, 1000$  when  $D_u = D_v = 0.01$  and  $\chi_u = \chi_v = 3.0$ .

which are displayed in Fig. 14(a). Compared to the initial data in (6.1), the functions in (6.2) miss the additive constant 0.1 and their maxima are further apart from each other. Therefore, the dark purple at the initial stage in Fig. 14(a) represents not only that both concentrations are (roughly) equal but additionally that they are both roughly zero.

From Fig. 14, we observe that the gangs stay totally separated over the time and there is no interaction between the two groups from beginning to the end. The snapshot of the result along the line  $y = x$  is displayed in Fig. 15. We also show the close up of population densities at different time instants along the line  $y = x$  in

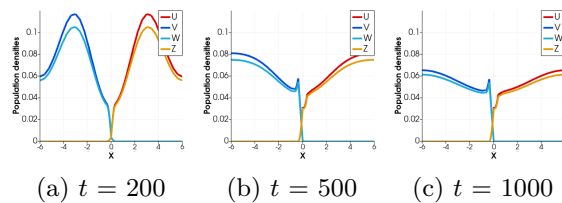


Fig. 16. Close up of populations densities along the line  $y = x$  at different time instant  $t = 200, 500, 1000$  when  $D_u = D_v = 0.01$  and  $\chi_u = \chi_v = 3.0$ .

Fig. 16, as can be seen no size of gang densities  $u$  appears in the regions occupied by the gang density  $v$  and vice versa.

### 6.2. Time step and mesh convergence study

Now, we examine the effect of time and mesh refinement, which is one of the standard numerical experiments used to validate approximate solutions. For this reason, as in the first example in Sec. 6.1.2 (see Figs. 8 and 9) we consider  $D_u = D_v = 0.25$ ,  $\chi_u = \chi_v = 3$  and, the initial conditions (6.1) over the time interval  $[0, 500]$ . To begin

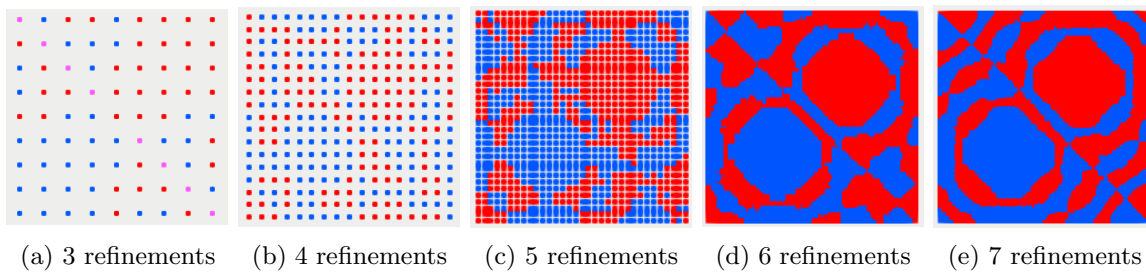


Fig. 17. (Color online) Numerical solutions  $u$  and  $v$  obtained using FEM-FCT scheme in different refinement levels at final time  $T = 500$  where  $D_u = D_v = 0.25$  and  $\chi_u = \chi_v = 3$ . For the choice of colors, see the beginning of Sec. 6.1.

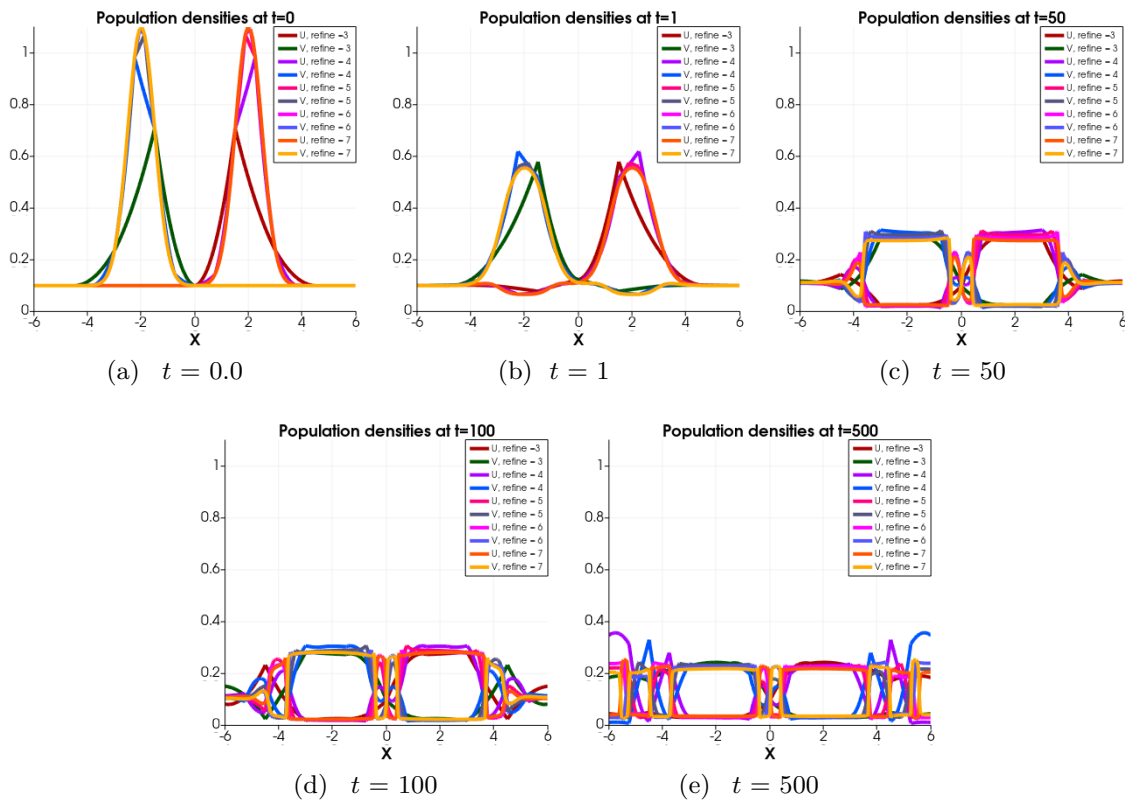


Fig. 18. The amount of gang concentration  $u$ ,  $v$  and graffiti densities  $w$  and  $z$  along the line  $y = x$  in different level of refinement at different time instants  $t = 0, 1, 50, 100, 500$  where  $D_u = D_v = 0.25$  and  $\chi_u = \chi_v = 3$ .

with, we keep all of the setting fixed as before and only coarsen and refine the mesh. Comparing the plots depicted in Fig. 17 for different levels of refinement at final time  $T = 500$ , we observe that the segregation patterns are almost the same with slight smoothness in the segregated patches boundaries when using the finer mesh. Therefore, it is clear that the formation of blue and red cluster inside the domain does not change significantly by changing the mesh size. The snapshots of the mesh convergence results along the line  $y = x$  are also displayed in Fig. 18, which shows the approximate solutions obtained at different refinement levels almost converge toward the same segregation states.

Next, preserving the five level of refinements as in all the previous examples and keep the other parameters fixed as before, we choose different time step sizes in our computation. Figure 19 shows that the segregation patterns are similar and

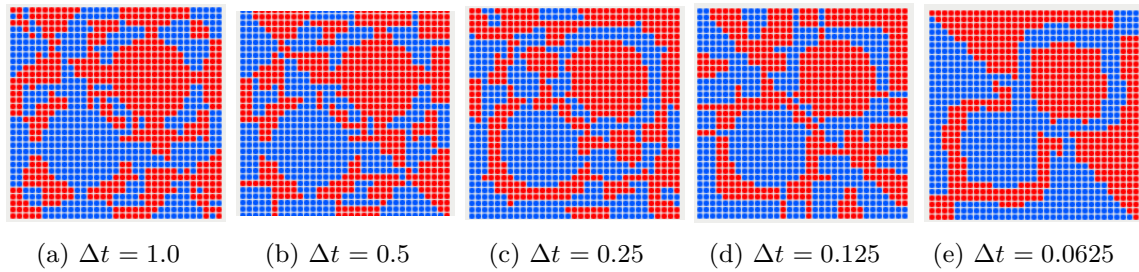


Fig. 19. (Color online) Numerical solutions  $u$  and  $v$  obtained using FEM-FCT scheme with different number of time steps  $\Delta t = 1.0, 0.5, 0.25, 0.125, 0.0625$  at final time  $T = 500$  where  $D_u = D_v = 0.25$  and  $\chi_u = \chi_v = 3$ . For the choice of colors, see the beginning of Sec. 6.1.

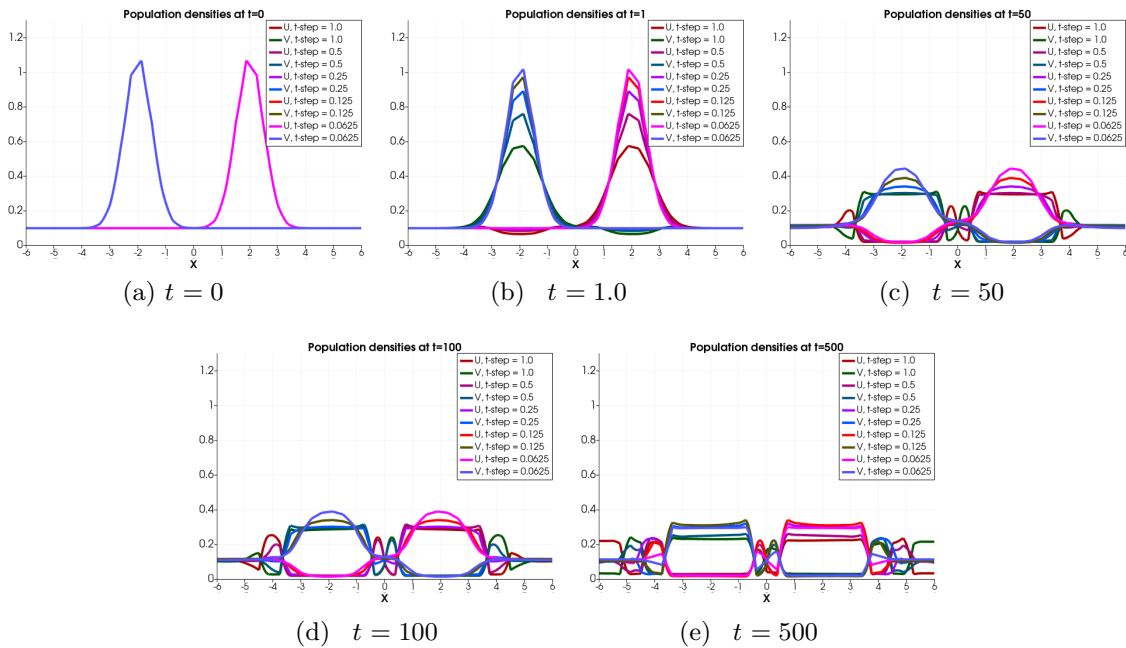


Fig. 20. The amount of gangs concentration  $u, v$  and graffiti densities  $z$  and  $w$  along the line  $y = x$  with different time steps at different time instants  $t = 0, 1, 50, 100, 500$  where  $D_u = D_v = 0.25$  and  $\chi_u = \chi_v = 3$ .


the differences between the approximate solutions obtained using FEM-FCT with different time steps  $\Delta t = 1.0, 0.5, 0.25, 0.125$ , and  $0.0625$  are very small. The snapshots of the results along the line  $y = x$  are depicted in Fig. 20, which shows the time convergence behavior.

We conclude that these convergence studies provide strong indication that our numerical experiments are reliable and that hence indeed both partial and complete separation may occur in (1.1).

## Acknowledgments

The work of Shahin Heydari was supported through the Grant SVV-2023-260711 of the Charles University and Grant No. 22-01591S of the Czech Science Foundation.

## ORCID

Mario Fuest  <https://orcid.org/0000-0002-8471-4451>

Shahin Heydari  <https://orcid.org/0009-0007-9128-5257>

## References

1. N. Alikakos,  $L^p$  bounds of solutions of reaction-diffusion equations, *Comm. Partial Differential Equations* **4** (1979) 827–868.
2. A. Alsenafi and A. B. T. Barbaro, A convection–diffusion model for gang territoriality, *Phys. A* **510** (2018) 765–786.
3. A. Alsenafi and A. B. T. Barbaro, A multispecies cross-diffusion model for territorial development, *Mathematics* **9** (2021) 1428.
4. D. Arndt, W. Bangerth, D. Davydov, T. Heister, L. Heltai, M. Kronbichler, M. Maier, J.-P. Pelteret, B. Turcksin and D. Wells, The deal.II finite element library: Design, features, and insights, *Comput. Math. Appl.* **81** (2021) 407–422.
5. D. Arndt, W. Bangerth, M. Feder, M. Fehling, R. Gassmüller, T. Heister, L. Heltai, M. Kronbichler, M. Maier, P. Munch, J.-P. Pelteret, S. Sticker, B. Turcksin and D. Wells, The deal.II library, version 9.4, *J. Numer. Math.* **30** (2022) 231–246.
6. A. B. T. Barbaro, N. Rodriguez, H. Yoldaş and N. Zamponi, Analysis of a cross-diffusion model for rival gangs interaction in a city, *Commun. Math. Sci.* **19** (2021) 2139–2175.
7. G. R. Barrenechea, V. John and P. Knobloch, Finite element methods respecting the discrete maximum principle for convection-diffusion equations, *SIAM Rev.* **66** (2024) 3–88.
8. N. Bellomo, D. Burini, G. Dosi, L. Gibelli, D. Knopoff, N. Outada, P. Terna and M. E. Virgillito, What is life? A perspective of the mathematical kinetic theory of active particles, *Math. Models Methods Appl. Sci.* **31** (2021) 1821–1866.
9. N. Bellomo, L. Gibelli, A. Quaini and A. Reali, Towards a mathematical theory of behavioral human crowds, *Math. Models Methods Appl. Sci.* **32** (2022) 321–358.
10. N. Bellomo, N. Outada, J. Soler, Y. Tao and M. Winkler, Chemotaxis and cross-diffusion models in complex environments: Models and analytic problems toward a multiscale vision, *Math. Models Methods Appl. Sci.* **32** (2022) 713–792.
11. D. Book, J. Boris and K. Hain, Flux-corrected transport II: Generalizations of the method, *J. Comput. Phys.* **18** (1975) 248–283.

12. M. Braukhoff and J. Lankeit, Stationary solutions to a chemotaxis-consumption model with realistic boundary conditions for the oxygen, *Math. Models Methods Appl. Sci.* **29** (2019) 2033–2062.
13. A. N. Brooks and T. J. R. Hughes, Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations, *Comput. Methods Appl. Mech. Engrg.* **32** (1982) 199–259.
14. W. K. Brown, Graffiti, identity and the delinquent gang, *Int. J. Offender Ther. Comp. Criminol.* **22** (1978) 46–48.
15. E. Burman, Consistent SUPG-method for transient transport problems: Stability and convergence, *Comput. Methods Appl. Mech. Engrg.* **199** (2010) 1114–1123.
16. J. W. Cahn and J. E. Hilliard, Free energy of a nonuniform system. I. Interfacial free energy, *J. Chem. Phys.* **28** (1958) 258–267.
17. L. Corrias, B. Perthame and H. Zaag, A chemotaxis model motivated by angiogenesis, *C. R. Math.* **336** (2003) 141–146.
18. L. Corrias, B. Perthame and H. Zaag, Global solutions of some chemotaxis and angiogenesis systems in high space dimensions, *Milan J. Math.* **72** (2004) 1–28.
19. T. A. Davis, Algorithm 832: UMFPACK V4.3 — An unsymmetric-pattern multifrontal method, *ACM Trans. Math. Software* **30** (2004) 196–199.
20. R. Ducasse, F. Santambrogio and H. Yoldaş, A cross-diffusion system obtained via (convex) relaxation in the JKO scheme, *Calc. Var. Partial Differential Equations* **62** (2023) 29.
21. X. Feng, X. Huang and K. Wang, Error estimate of unconditionally stable and decoupled linear positivity-preserving FEM for the chemotaxis-Stokes equations, *SIAM J. Numer. Anal.* **59** (2021) 3052–3076.
22. M. A. Fontelos, A. Friedman and B. Hu, Mathematical analysis of a model for the initiation of angiogenesis, *SIAM J. Math. Anal.* **33** (2002) 1330–1355.
23. L. P. Franca, C. Farhat, M. Lesoinne and A. Russo, Unusual stabilized finite element methods and residual free bubbles, *Internat. J. Numer. Methods Fluids* **27** (1998) 159–168.
24. L. P. Franca, G. Hauke and A. Masud, Revisiting stabilized finite element methods for the advective–diffusive equation, *Comput. Methods Appl. Mech. Engrg.* **195** (2006) 1560–1572.
25. A. Friedman and J. Tello, Stability of solutions of chemotaxis equations in reinforced random walks, *J. Math. Anal. Appl.* **272** (2002) 138–163.
26. M. Fuest, Blow-up profiles in quasilinear fully parabolic Keller–Segel systems, *Nonlinearity* **33** (2020) 2306–2334.
27. M. Fuest, Global solutions near homogeneous steady states in a multidimensional population model with both predator- and prey-taxis, *SIAM J. Math. Anal.* **52** (2020) 5865–5891.
28. M. Fuest, Global weak solutions to fully-cross diffusive systems with nonlinear diffusion and saturated taxis sensitivity, *Nonlinearity* **35** (2022) 608–657.
29. M. Fuest, Sh. Heydari, P. Knobloch, J. Lankeit and T. Wick, Global existence of classical solutions and numerical simulations of a cancer invasion model, *ESAIM Math. Model. Numer. Anal.* **57** (2023) 1893–1919.
30. S. Ganesan and L. Tobiska, Stabilization by local projection for convection–diffusion and incompressible flow problems, *J. Sci. Comput.* **43** (2010) 326–342.
31. V. Giunta, T. Hillen, M. Lewis and J. R. Potts, Local and global existence for nonlocal multispecies advection-diffusion models, *SIAM J. Appl. Dyn. Syst.* **21** (2022) 1686–1708.

32. Sh. Heydari, P. Knobloch and T. Wick, Flux-corrected transport stabilization of an evolutionary cross-diffusion cancer invasion model, *J. Comput. Phys.* **499** (2024) 112711.
33. X. Huang, X. Feng, X. Xiao and K. Wang, Fully decoupled, linear and positivity-preserving scheme for the chemotaxis–Stokes equations, *Comput. Methods Appl. Mech. Engrg.* **383** (2021) 113909.
34. T. J. R. Hughes, L. P. Franca and G. M. Hulbert, A new finite element formulation for computational fluid dynamics. VIII. The Galerkin/least-squares method for advective-diffusive equations, *Comput. Methods Appl. Mech. Engrg.* **73** (1989) 173–189.
35. V. John and P. Knobloch, On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part I — A review, *Comput. Methods Appl. Mech. Engrg.* **196** (2007) 2197–2215.
36. V. John, P. Knobloch and P. Korsemeier, On the solvability of the nonlinear problems in an algebraically stabilized finite element method for evolutionary transport-dominated equations, *Math. Comp.* **90** (2021) 595–611.
37. V. John and J. Novo, Error analysis of the SUPG finite element discretization of evolutionary convection-diffusion-reaction equations, *SIAM J. Numer. Anal.* **49** (2011) 1149–1176.
38. A. Jüngel, S. Portisch and A. Zurek, Nonlocal cross-diffusion systems for multi-species populations and networks, *Nonlinear Anal.* **219** (2022) 112800.
39. K. Kuto and Y. Yamada, Multiple coexistence states for a prey–predator system with cross-diffusion, *J. Differential Equations* **197** (2004) 315–348.
40. D. Kuzmin, Explicit and implicit FEM-FCT algorithms with flux linearization, *J. Comput. Phys.* **228** (2009) 2517–2534.
41. D. Kuzmin, Algebraic flux correction I. Scalar conservation laws, in *Flux-Corrected Transport: Principles, Algorithms, and Applications*, eds. D. Kuzmin, R. Löhner and S. Turek, 2nd edn. (Springer, 2012), pp. 145–192.
42. D. Kuzmin and S. Turek, Flux correction tools for finite elements, *J. Comput. Phys.* **175** (2002) 525–558.
43. O. A. Ladyženskaja, V. A. Solonnikov and N. N. Ural’ceva, *Linear and Quasi-Linear Equations of Parabolic Type*, Translations of Mathematical Monographs, Vol. 23 (Amer. Math. Soc., 1988).
44. Ph. Laurençot and B.-V. Matioc, Bounded weak solutions to the thin film Muskat problem via an infinite family of Liapunov functionals, *Trans. Amer. Math. Soc.* **375** (2022) 5963–5986.
45. Ph. Laurençot and B.-V. Matioc, Bounded weak solutions to a class of degenerate cross-diffusion systems, *Ann. Henri Lebesgue* **6** (2023) 847–874.
46. D. Ley and R. Cybriwsky, Urban graffiti as territorial markers, *Ann. Assoc. Am. Geogr.* **64** (1974) 491–505.
47. G. Li, Y. Tao and M. Winkler, Large time behavior in a predator-prey system with indirect pursuit-evasion interaction, *Discrete Contin. Dyn. Syst. Ser. B* **25** (2020) 4383–4396.
48. M. Mimura and K. Kawasaki, Spatial segregation in competitive interaction-diffusion equations, *J. Math. Biol.* **9** (1980) 49–64.
49. A. Mizukami and T. Hughes, A Petrov-Galerkin finite element method for convection-dominated flows: An accurate upwinding technique for satisfying the maximum principle, *Comput. Methods Appl. Mech. Engrg.* **50** (1985) 181–193.
50. L. Nirenberg, On elliptic partial differential equations, *Ann. Sc. Norm. Super. Pisa Cl. Sci. (3)* **13** (1959) 115–162.

51. J. R. Potts and M. A. Lewis, How memory of direct animal interactions can lead to territorial pattern formation, *J. R. Soc. Interface* **13** (2016) 20160059.
52. J. R. Potts and M. A. Lewis, Spatial memory and taxis-driven pattern formation in model ecosystems, *Bull. Math. Biol.* **81** (2019) 2725–2747.
53. P. Rybka and K.-H. Hoffmann, Convergence of solutions to Cahn-Hilliard equation, *Comm. Partial Differential Equations* **24** (1999) 1055–1077.
54. N. Shigesada, K. Kawasaki and E. Teramoto, Spatial segregation of interacting species, *J. Theor. Biol.* **79** (1979) 83–99.
55. R. Strehl, A. Sokolov, D. Kuzmin, D. Horstmann and S. Turek, A positivity-preserving finite element method for chemotaxis problems in 3D, *J. Comput. Appl. Math.* **239** (2013) 290–303.
56. R. Strehl, A. Sokolov, D. Kuzmin and S. Turek, A flux-corrected finite element method for chemotaxis problems, *Comput. Methods Appl. Math.* **10** (2010) 219–232.
57. M. Sulman and T. Nguyen, A positivity preserving moving mesh finite element method for the Keller–Segel chemotaxis model, *J. Sci. Comput.* **80** (2019) 649–666.
58. Y. Tao and M. Winkler, Boundedness in a quasilinear parabolic–parabolic Keller–Segel system with subcritical sensitivity, *J. Differential Equations* **252** (2012) 692–715.
59. Y. Tao and M. Winkler, Energy-type estimates and global solvability in a two-dimensional chemotaxis–haptotaxis model with remodeling of non-diffusible attractant, *J. Differential Equations* **257** (2014) 784–815.
60. Y. Tao and M. Winkler, A fully cross-diffusive two-component evolution system: Existence and qualitative analysis via entropy-consistent thin-film-type approximation, *J. Funct. Anal.* **281** (2021) 109069.
61. Y. Tao and M. Winkler, Existence theory and qualitative analysis for a fully cross-diffusive predator-prey system, *SIAM J. Math. Anal.* **54** (2022) 4806–4864.
62. Y. Tyutyunov, L. Titova and R. Arditi, A minimal model of pursuit-evasion in a predator-prey system, *Math. Model. Nat. Phenom.* **2** (2007) 122–134.
63. J. Wei and M. Winter, On the stationary Cahn–Hilliard equation: Interior spike solutions, *J. Differential Equations* **148** (1998) 231–267.
64. S. Wu, Global boundedness of a diffusive prey-predator model with indirect prey-taxis and predator-taxis, *J. Math. Anal. Appl.* **507** (2021) 125820.
65. X. Xu, Existence theorem for a partially parabolic cross-diffusion system, preprint (2023), arXiv:2304.04329.
66. S. T. Zalesak, Fully multidimensional flux-corrected transport algorithms for fluids, *J. Comput. Phys.* **31** (1979) 335–362.

# 5. Paper V

## 5.1 Introduction: Nonstandard finite difference scheme

In this chapter, we consider a nonstandard finite difference (NSFD) scheme, which was introduced nearly four decades ago by Mickens [244, 245, 246] to deal with numerical instabilities that occur when finite difference methods are used for solving differential equations. Since then, these methods has been widely used to obtain numerical solutions of vast variety of ordinary and partial differential equations which appear in different problems in science and engineering [247, 248]. It was shown that the NSFD schemes overpower the standard finite difference methods even when dealing with numerically challenging equations, such as convection-dominated transport problems with nonlinear reaction terms [249, 250, 251, 245], or Maxwell's equations [252, 245].

The basic idea behind these methods is as follows:

1. A nonlocal representation of the nonlinear terms should be used on computational grid instead of conventional local representation. Let us consider the logistic equation  $\frac{du}{dt} = u(1 - u)$ , then the idea is to evaluate the nonlinear term  $u^2$  at two (or more) different grid points, e.g.,

$$u^2 \rightarrow u_{k+1}u_k,$$

where

$$\lim_{h \rightarrow 0} u_{k+1}u_k = \lim_{h \rightarrow 0} u_k^2 = u^2,$$

or

$$u^2 \rightarrow 2(u_k^2) - u_{k+1}u_k.$$

2. Replacing the denominator functions in the discrete derivatives by more complicated functional forms than those used in the standard scheme. Considering again the logistic equation, the derivative is to be approximated as

$$\frac{du}{dt} \rightarrow \frac{u_{k+1} - u_k}{\psi(h)},$$

where

$$\psi(h) = h + O(h^2), \quad h \rightarrow 0,$$

or more generally as

$$\frac{du}{dt} = \lim_{h \rightarrow 0} \frac{u[t + \psi_1(h)] - u(t)}{\psi_2(h)},$$

where

$$\psi_i(h) = h + O(h^2), \quad h \rightarrow 0, \quad i = 1, 2,$$

examples of such functions that satisfy the assumptions above are:  $\psi(h) = \sin(h)$ ,  $\psi(h) = 1 - \exp(-h)$ .



Using the preceding steps, it is obvious that there is a lot of freedom in the construction of NSFD schemes, however, several important rules have been introduced over the years which play an important role in proper construction of these methods:

Rule 1: The orders of discrete derivatives have to be equal to the orders of the corresponding derivatives of the differential equations, violation of this rule leads to spurious oscillations in the approximate solutions.

Rule 2: The solutions of the NSFD discretization should preserve the properties of the solutions of the differential equations.

Rule 3: For differential equations with  $N \geq 3$  terms, it is useful to consider various sub-equations composed of  $M < N$  terms, then construct the finite difference schemes for each one of these sub-equations, and finally combine all together in a consistent way.

Rule 4: In the discretization of partial differential equations, there may exist functional relations between various space and time step-sizes, e.g.,  $\Delta t \propto \Delta x$ .

Rule 5: In the discretization of partial differential equations, different types of discrete derivatives may be required for the time and space variables.

Despite huge amount of practical applications of NSFD for numerical simulation of ordinary and partial differential equations and their huge success to overcome numerical instabilities over the years, the amount of theoretical studies to understand why they work so well and when they do is considerably poor and calls for further investigation.

## **5.2 A positive and elementary stable nonstandard explicit scheme for a mathematical model of the influenza disease**

This section is based on the paper entitled "A positive and elementary stable nonstandard explicit scheme for a mathematical model of the influenza disease", published in the Journal of Mathematics and Computers in Simulation.

In this paper, we considered a system of four strongly coupled nonlinear ordinary differential equations describing the spread of Influenza disease. The studied system is solved numerically using a method of nonstandard finite difference scheme following the Mickens rules [245, 244, 246], which are briefly described in Section 2. Given the description and the properties of the model in Section 3, it was proved that for positive initial conditions the solution of the system stays positive over time, which satisfies also the physics behind the model. Next, the strategy and formulation of the nonstandard discretization for solving the studied system was proposed in Section 4, it was shown that the proposed method is elementary stable and under certain assumption on the time-step function it is also positivity preserving. The properties of the proposed scheme were verified using some numerical results in Section 5.



## Original articles

# A positive and elementary stable nonstandard explicit scheme for a mathematical model of the influenza disease

Mohammad Mehdizadeh Khalsaraei<sup>a,\*</sup>, Ali Shokri<sup>a</sup>, Higinio Ramos<sup>b</sup>, Shahin Heydari<sup>c</sup><sup>a</sup> Faculty of Mathematical Science, University of Maragheh, Maragheh, Iran<sup>b</sup> Department of Applied Mathematics, Scientific Computing Group, University of Salamanca, Spain<sup>c</sup> Charles University, Faculty of Mathematics and Physics, Department of Numerical Mathematics, Czech Republic

Received 15 September 2020; received in revised form 10 November 2020; accepted 12 November 2020

Available online 20 November 2020

## Abstract

In this paper, a nonstandard explicit discretization strategy is considered to construct a new nonstandard finite difference scheme for solving a mathematical model of the influenza disease. The new proposed scheme has some interesting properties such as high accuracy and ease of implementation, as well as some preserving properties of the exact theoretical solution of the SIRC system, like positivity and elementary stability. These characteristics make it suitable for solving efficiently the proposed model. We provide some numerical comparisons to illustrate our results.

© 2020 International Association for Mathematics and Computers in Simulation (IMACS). Published by Elsevier B.V. All rights reserved.

MSC: 65L05; 03H05; 03H10

Keywords: Nonstandard finite difference; Positivity; Elementary stability; Conservation law

## 1. Introduction

Ordinary differential equations (ODEs) are used extensively in the modeling of many biological and physical applications. 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. Numerical methods based on the finite difference approximations [3,4,9,28], Taylor series expansion [29], and interpolation, such as Euler, Runge–Kutta and multistep methods [20–22,31], and some other methods [1,8,12,15,30–41], are widely used. Traditionally, important requirements in this context are, the investigation of the consistency of the discrete scheme with the original differential equation and linear stability analysis for problems with smooth solutions. These requirements are formulated to guarantee the convergence of the discrete solution to the exact one, but sometimes the essential qualitative properties of the solution are not transferred to the numerical solution. One way to tackle with this issue is to employ finite difference schemes that are nonstandard in the sense of Mickens' definition [18,25,27]. Nonstandard finite difference methods (NSFDs) in addition to the usual properties of the solutions such as consistency, stability and hence convergence, may also preserve essential properties of

\* Corresponding author.

E-mail addresses: [mehdizadeh@maragheh.ac.ir](mailto:mehdizadeh@maragheh.ac.ir) (M.M. Khalsaraei), [shokri@maragheh.ac.ir](mailto:shokri@maragheh.ac.ir) (A. Shokri), [higra@usal.es](mailto:higra@usal.es) (H. Ramos), [sh1990.heydari@gmail.com](mailto:sh1990.heydari@gmail.com) (S. Heydari).

<https://doi.org/10.1016/j.matcom.2020.11.013>

0378-4754/© 2020 International Association for Mathematics and Computers in Simulation (IMACS). Published by Elsevier B.V. All rights reserved.

the solutions, like positivity, boundedness, monotonicity and total variation diminishing [2,6,7,16–19,23–25,27]. In this paper, we propose a new NSFD scheme for approximating the solution of the influenza disease system. The proposed scheme enables us to solve the examined problem accurately. An important feature of the new scheme is the positivity preservation of the produced solutions, which is an essential property in this context. We also prove that the new scheme is elementary stable.

The rest of the paper is organized as follows. In Section 2, we provide some preliminaries and definitions, including that of non-standard finite difference methods for ODEs, and a review of the general influenza disease model. [14]. In Section 3, we propose the new scheme and investigate its positivity and elementary stability. In Section 4, we compare the results obtained from the new scheme with the ones obtained from the classical fourth order Runge–Kutta method (we call it RK4), ode45, ode15s, the NSFD scheme in [13] and the NSFD scheme in [26]. Finally, we end the paper with some conclusions in Section 5.

## 2. Preliminaries and definitions

In this section, we give a brief summary of the NSFD methods for the numerical solution of initial value problems for systems of ODEs that can be written in the autonomous form

$$\frac{d}{dt}y(t) = F(y(t)), \quad (t \geq 0), \quad y(t_0) = y_0, \tag{2.1}$$

where  $y(t)$  may be a single function or a vector function of length  $k$  mapping  $[t_0, T] \rightarrow \mathbb{R}^k$  and  $F$  is a single function or a vector function of length  $k$  mapping  $\mathbb{R}^k \rightarrow \mathbb{R}^k$ . By defining  $t_n = t_0 + n\Delta t$ , where  $\Delta t$  is a positive step size, the continuous differential equation (2.1) can be discretized as

$$\mathcal{D}_{\Delta t}y_n = \mathcal{F}_n(F, y_n), \tag{2.2}$$

where  $y_n \approx y(t_n)$ ,  $\mathcal{D}_{\Delta t}y_n$  represents the discretized version of  $\frac{d}{dt}y(t)$  and  $\mathcal{F}_n(F, y_n)$  approximates  $F(y(t))$  at time  $t_n$ . In the sequel, we will consider the definition of the nonstandard finite-difference methods given in [2].

**Definition 2.1** ([2]). The method given in (2.2) is called a nonstandard finite-difference method if at least one of the following conditions is met:

- In the discrete derivatives  $\mathcal{D}_{\Delta t}y_n$ , the traditional denominator  $\Delta t$  is replaced by a nonnegative function  $\varphi(\Delta t)$  such that

$$\varphi(\Delta t) = \Delta t + O(\Delta t^2) \text{ as } 0 < \Delta t \rightarrow 0, \tag{2.3}$$

for example:

$$\varphi(\Delta t) = 1 - \exp(-\Delta t), \quad \varphi(\Delta t) = \tanh(\Delta t).$$

- $F(y(t))$  is approximated in a nonlocal way, i.e., by a suitable function of several points of the mesh. For instance, the terms  $y$ ,  $y^2$  and  $y^3$  can be modeled as follows:

$$\begin{aligned} y &\approx ay_k + (1 - a)y_{k+1}; & y &\approx a(y_{k+1} + y_{k-1}) + (1 - 2a)y_k, & a &\in \mathbb{R}; \\ y^2 &\approx ay_k^2 + by_k y_{k+1}, & a + b &= 1, & a, b &\in \mathbb{R}; & y^2 &\approx y_k \left( \frac{y_{k+1} + y_{k-1}}{2} \right); \\ y^3 &\approx ay_k^3 + (1 - a)y_k^2 y_{k+1}, & a &\in \mathbb{R}, \end{aligned}$$

where  $y_{k+j}$  denotes an approximation of the true solution  $y(t_{k+j})$ .

**Definition 2.2.** Any constant value  $\tilde{y}_0$  satisfying  $F(\tilde{y}_0) = 0$  is called an equilibrium point (a fixed-point or a critical point) of the differential equation given in (2.1). The constant solutions of the discretized system are also called equilibrium points.

**Definition 2.3.** The finite difference method given in (2.2) is called elementary stable, if for any value of the step size  $\Delta t$ , the only equilibrium points are those of the differential system (2.1), and the linear stability property of each one is the same for both, the differential system and its discretized version.

**Table 1**  
 Descriptions and values of the parameters used in the system (3.1).

Description	Parameter	Value
Cross-immune period	$\gamma^{-1}$	2
Infectious period	$\alpha^{-1}$	$\frac{5}{365}$
Total immune period	$\delta^{-1}$	1
Per capita birth rate	$\mu$	$\frac{1}{50}$
Fraction of the exposed cross-immune individuals	$\sigma$	0.05

**Definition 2.4.** An equilibrium point  $\tilde{y}_0$  of (2.1) is linearly

- (i) stable iff  $|Re\lambda_j| < 1$  for all  $j$ ,
- (ii) unstable iff  $|Re\lambda_j| > 1$  for at least one  $j$ ,

where the  $\lambda_j$ 's are the eigenvalues of the Jacobian matrix of the system (2.1) evaluated at  $\tilde{y}_0$ .

### 3. A mathematical model of the influenza disease

In this section, we consider the mathematical model of the influenza disease, completely analyzed in [13,14,26], given in the form

$$\begin{aligned}
 \frac{dS(t)}{dt} &= \mu - \mu S(t) - \beta S(t)I(t) + \gamma C(t), \\
 \frac{dI(t)}{dt} &= \beta S(t)I(t) + \sigma\beta C(t)I(t) - (\mu + \alpha)I(t), \\
 \frac{dR(t)}{dt} &= (1 - \sigma)\beta C(t)I(t) + \alpha I(t) - (\mu + \delta)R(t), \\
 \frac{dC(t)}{dt} &= \delta R(t) - \beta C(t)I(t) - (\mu + \gamma)C(t), \\
 S(0) &= S_0, \quad I(0) = I_0, \quad R(0) = R_0, \quad C(0) = C_0,
 \end{aligned} \tag{3.1}$$

where  $S$ ,  $I$ ,  $R$  and  $C$  represent the proportion of susceptible, infective, recovered and cross-immune individuals at time  $t$ , respectively and  $\beta$  is the contact rate. The definitions of the other parameters present in the system (3.1) and the values used for them in this article can be found in Table 1. One of the main assumptions of this model is that the per capita birth rate is a constant  $\mu > 0$  and the birth rate is the same as death rate. It implies that  $S'(t) + I'(t) + R'(t) + C'(t) = 0$  (conservation law).

**Theorem 3.1.** The solution  $(S(t), E(t), I(t), R(t))$  of system (3.1) with positive initial condition is positive on  $[0, \infty)$ .

**Proof.** Assume the solution  $(S(t), I(t), R(t), C(t))$  with a positive initial condition exists and is unique on  $[0, b)$ , where  $0 < b \leq \infty$  (see [11]). Since

$$I'(t) = [\beta S(t) + \sigma\beta C(t) - (\mu + \alpha)] I(t),$$

then

$$I(t) = I(0) \exp \left[ \int_0^t [\beta S(\theta) + \sigma\beta C(\theta) - (\mu + \alpha)] d\theta \right] > 0.$$

So, for all  $t \in [0, b)$  we have  $I(t) > 0$ . Now, for all  $t \in [0, b)$ , one must have  $C(t) > 0$ . Otherwise, there will exist a  $t_1 \in (0, b)$  such that  $C(t_1) = 0$  and  $C(t) > 0$  in  $(0, t_1)$ . Thus, for any  $t \in [0, t_1)$ ,

$$\begin{aligned}
 S'(t) &= \mu - \mu S(t) - \beta S(t)I(t) + \gamma C(t) \\
 &\geq \mu - \mu S(t) - \beta S(t)I(t) \\
 &\geq -(\mu + \beta I(t))S(t).
 \end{aligned}$$

Hence, for all  $t \in (0, t_1)$ ,

$$S(t) \geq S(0) \exp \left[ \int_0^t -(\mu + \beta I(\theta)) d\theta \right] > 0.$$

Now since  $1 - \sigma \geq 0$ , for all  $t \in [0, t_1]$  we have

$$\begin{aligned} R'(t) &= (1 - \sigma)\beta C(t)I(t) + \alpha I(t) - (\mu + \delta)R(t) \\ &\geq \alpha I(t) - (\mu + \delta)R(t) \\ &\geq -(\mu + \delta)R(t). \end{aligned}$$

Then, for all  $t \in (0, t_1)$ ,

$$R(t) \geq R(0) \exp \left[ \int_0^t -(\mu + \delta) d\theta \right] > 0.$$

Therefore, for  $t \in [0, t_1]$  we can write

$$C'(t) = \delta R(t) - \beta C(t)I(t) - (\mu + \gamma)C(t) \geq -(\beta I(t) + (\mu + \gamma))C(t).$$

Hence, by using a comparison argument we obtain that

$$C(t) \geq C(0) \exp \left[ - \int_0^t (\beta I(\theta) + (\mu + \gamma)) d\theta \right] > 0,$$

and in particular, for  $t = t_1$  we get

$$C(t_1) \geq C(0) \exp \left[ - \int_0^{t_1} (\beta I(\theta) + (\mu + \gamma)) d\theta \right] > 0$$

which is a contradiction to  $C(t_1) = 0$ . So, for all  $t \in [0, b)$ ,  $C(t) > 0$ . Using similar procedures, one can show that  $R(t) > 0$  and  $S(t) > 0$  for all  $t \in [0, b)$ . On the other hand, we have

$$\frac{dN}{dt} = \mu - \mu N(t), \quad N(t) = S(t) + I(t) + R(t) + C(t), \tag{3.2}$$

whose exact solution is

$$N(t) = 1 + (N(0) - 1) e^{-\mu t} = 1 - e^{-\mu t} + N(0)e^{-\mu t}, \tag{3.3}$$

where  $N(0) = S(0) + I(0) + R(0) + C(0) > 0$ . We have that for  $t \in [0, b)$  it is

$$N(t) < 1 + N(0)e^{-\mu t} < 1 + N(0).$$

Thus,  $S(t), I(t), R(t), C(t)$  are bounded on  $[0, b)$  and we have that  $b = \infty$ . This completes the proof.  $\square$

Following Definition 2.2, the system (3.1) has the equilibrium points [5]:

- the disease-free equilibrium (DFE),  $E_0 = (1, 0, 0, 0)$ ,
- the positive endemic equilibrium (EE),  $E^* = (S^*, I^*, R^*, C^*)$ .

The stability of the these points is often described in terms of the *reproductive number* of the system. The reproductive number represents the number of secondary infections a primary infection generates on average over the course of its infectious period. The reproductive number for system (3.1) is,

$$R_0 = \frac{\beta}{\alpha + \mu},$$

and the stability of the equilibrium points is as follows:

- the disease free equilibrium,  $E_0$ , is asymptotically stable if  $R_0 < 1$  and is unstable if  $R_0 > 1$ ,
- the endemic equilibrium,  $E^*$ , is asymptotically stable if  $R_0 > 1$ .

#### 4. Description and properties of the numerical scheme

By using the strategy of the nonstandard discretizations, we propose a new scheme for (3.1) given by:

$$\begin{aligned}
 \frac{S_{i+1} - S_i}{\varphi(\Delta t)} &= \mu - \mu(2S_{i+1} - S_i) - \beta S_{i+1} I_i + \gamma C_i, \\
 \frac{I_{i+1} - I_i}{\varphi(\Delta t)} &= \beta S_{i+1} I_i + \sigma \beta C_i I_i - \mu(2I_{i+1} - I_i) - \alpha I_{i+1}, \\
 \frac{R_{i+1} - R_i}{\varphi(\Delta t)} &= (1 - \sigma) \beta C_i I_i + \alpha I_{i+1} - \mu(2R_{i+1} - R_i) - \delta R_{i+1}, \\
 \frac{C_{i+1} - C_i}{\varphi(\Delta t)} &= \delta R_{i+1} - \beta C_i I_i - \mu(2C_{i+1} - C_i) - \gamma C_i.
 \end{aligned}
 \tag{4.1}$$

The explicit form of (4.1) can be written as

$$S_{i+1} = \frac{(1 + \varphi(\Delta t)\mu)S_i + \varphi(\Delta t)\mu + \varphi(\Delta t)\gamma C_i}{1 + 2\varphi(\Delta t)\mu + \varphi(\Delta t)\beta I_i},
 \tag{4.2}$$

$$I_{i+1} = \frac{(1 + \varphi(\Delta t)\beta S_{i+1} + \varphi(\Delta t)\sigma \beta C_i + \varphi(\Delta t)\mu)I_i}{1 + 2\varphi(\Delta t)\mu + \varphi(\Delta t)\alpha},
 \tag{4.3}$$

$$R_{i+1} = \frac{(1 + \varphi(\Delta t)\mu)R_i + \varphi(\Delta t)(1 - \sigma)\beta C_i I_i + \varphi(\Delta t)\alpha I_{i+1}}{1 + 2\varphi(\Delta t)\mu + \varphi(\Delta t)\delta},
 \tag{4.4}$$

$$C_{i+1} = \frac{(1 + \varphi(\Delta t)\mu - \varphi(\Delta t)\beta I_i - \varphi(\Delta t)\gamma)C_i + \varphi(\Delta t)\delta R_{i+1}}{1 + 2\varphi(\Delta t)\mu}.
 \tag{4.5}$$

**Proposition 4.1.** *The new scheme (4.1) preserves the conservation law.*

**Proof.** It can be obtained by using induction. You have that  $S + I + R + C = 1$ , and thus for the initial values it is  $S_0 + I_0 + R_0 + C_0 = 1$ . Using the above after summing the left hand sides, and the right hand sides in (4.1) for  $i = 0$  you get  $S_1 + I_1 + R_1 + C_1 - 1 = 2\varphi\mu(1 - S_1 + I_1 + R_1 + C_1)$ , and thus  $S_1 + I_1 + R_1 + C_1 = 1$ . The inductive procedure results in  $S_{i+1} + I_{i+1} + R_{i+1} + C_{i+1} = 1$ , therefore the new scheme (4.1) preserves the conservation law.  $\square$

In the following, when  $\varphi(\Delta t) = \Delta t$ , the new method will be referred as NSFD- $\Delta t$ , and if  $\varphi(\Delta t)$  is different from  $\Delta t$ , the method will be referred as NSFD- $\varphi(\Delta t)$ .

**Theorem 4.2.** *The new proposed scheme (4.2)–(4.5) is elementary stable and for a chosen  $\varphi(\Delta t)$ , the sufficient condition for positivity is*

$$\varphi(\Delta t) \geq \frac{1}{\beta + \gamma - \mu}.$$

**Proof. Elementary stability:** The equilibrium points for the new proposed scheme are exactly the points  $E_0$  and  $E^*$  of the system (3.1). The Jacobian  $J$  of the scheme (4.1) has the form  $J(S_i, I_i, R_i, C_i) = [j_{mn}(S_i, I_i, R_i, C_i)]_{4 \times 4}$ , where

$$\begin{aligned}
 j_{11}(S_i, I_i, R_i, C_i) &= \frac{1 + \varphi(\Delta t)\mu}{1 + 2\varphi(\Delta t)\mu + \varphi(\Delta t)\beta I_i}, \\
 j_{12}(S_i, I_i, R_i, C_i) &= \frac{-\varphi(\Delta t)\beta[(1 + \varphi(\Delta t)\mu)S_i + \varphi(\Delta t)\mu + \varphi(\Delta t)\gamma C_i]}{(1 + 2\varphi(\Delta t)\mu + \varphi(\Delta t)\beta I_i)^2}, \\
 j_{13}(S_i, I_i, R_i, C_i) &= 0, \\
 j_{14}(S_i, I_i, R_i, C_i) &= \frac{\varphi(\Delta t)\gamma}{1 + 2\varphi(\Delta t)\mu + \varphi(\Delta t)\beta I_i}, \\
 j_{21}(S_i, I_i, R_i, C_i) &= \frac{\varphi(\Delta t)\beta I_i j_{11}}{1 + 2\varphi(\Delta t)\mu + \varphi(\Delta t)\alpha},
 \end{aligned}$$

$$\begin{aligned}
 j_{22}(S_i, I_i, R_i, C_i) &= \frac{1 + \varphi(\Delta t)\beta S_{i+1} + \varphi(\Delta t)\beta I_i j_{12} + \varphi(\Delta t)\sigma\beta C_i + \varphi(\Delta t)\mu}{1 + 2\varphi(\Delta t)\mu + \varphi(\Delta t)\alpha}, \\
 j_{23}(S_i, I_i, R_i, C_i) &= 0, \\
 j_{24}(S_i, I_i, R_i, C_i) &= \frac{\varphi(\Delta t)\beta\sigma I_i + \varphi(\Delta t)\beta I_i j_{11}}{1 + 2\varphi(\Delta t)\mu + \varphi(\Delta t)\alpha}, \\
 j_{31}(S_i, I_i, R_i, C_i) &= \frac{\varphi(\Delta t)\alpha j_{21}}{1 + \varphi(\Delta t)\delta + 2\varphi(\Delta t)\mu}, \\
 j_{32}(S_i, I_i, R_i, C_i) &= \frac{\varphi(\Delta t)(1 - \sigma)\beta C_i + \varphi(\Delta t)\alpha j_{22}}{1 + \varphi(\Delta t)\delta + 2\varphi(\Delta t)\mu}, \\
 j_{33}(S_i, I_i, R_i, C_i) &= \frac{\varphi(\Delta t)(1 - \sigma)\beta I_i + \varphi(\Delta t)\alpha j_{23}}{1 + \varphi(\Delta t)\delta + 2\varphi(\Delta t)\mu}, \\
 j_{34}(S_i, I_i, R_i, C_i) &= \frac{1 + \varphi(\Delta t)\mu + \varphi(\Delta t)\alpha j_{24}}{1 + \varphi(\Delta t)\delta + 2\varphi(\Delta t)\mu}, \\
 j_{41}(S_i, I_i, R_i, C_i) &= \frac{\varphi(\Delta t)\delta j_{31}}{1 + 2\varphi(\Delta t)\mu}, \\
 j_{42}(S_i, I_i, R_i, C_i) &= \frac{-\varphi(\Delta t)\beta C_i + \varphi(\Delta t)\delta j_{32}}{1 + 2\varphi(\Delta t)\mu}, \\
 j_{43}(S_i, I_i, R_i, C_i) &= \frac{1 + \varphi(\Delta t)\mu - \varphi(\Delta t)\gamma - \varphi(\Delta t)\beta I_i + \varphi(\Delta t)\delta j_{33}}{1 + 2\varphi(\Delta t)\mu}, \\
 j_{44}(S_i, I_i, R_i, C_i) &= \frac{\varphi(\Delta t)\delta j_{34}}{1 + 2\varphi(\Delta t)\mu}.
 \end{aligned}$$

By substituting  $(S_0, I_0, R_0, C_0) = (1, 0, 0, 0) = E_0$ , we have

$$J(E_0) = \begin{pmatrix} a_1 & a_2 & a_3 & a_4 \\ a_5 & a_6 & a_7 & a_8 \\ a_9 & a_{10} & a_{11} & a_{12} \\ a_{13} & a_{14} & a_{15} & a_{16} \end{pmatrix},$$

where

$$\begin{aligned}
 a_1 &= \frac{1 + \varphi(\Delta t)\mu}{1 + 2\varphi(\Delta t)\mu}, & a_2 &= \frac{-\varphi(\Delta t)\beta}{1 + 2\varphi(\Delta t)\mu}, & a_3 &= 0, \\
 a_4 &= \frac{\varphi(\Delta t)\gamma}{1 + 2\varphi(\Delta t)\mu}, & a_5 &= 0, & a_6 &= \frac{1 + \varphi(\Delta t)(\mu + \beta)}{1 + \varphi(\Delta t)(2\mu + \alpha)}, \\
 a_7 &= 0, & a_8 &= 0, & a_9 &= 0, \\
 a_{10} &= \frac{\varphi(\Delta t)\alpha(1 + \varphi(\Delta t)(\mu + \beta))}{(1 + \varphi(\Delta t)(2\mu + \delta))(1 + \varphi(\Delta t)(2\mu + \alpha))}, \\
 a_{11} &= \frac{1 + \varphi(\Delta t)\mu}{1 + \varphi(\Delta t)(2\mu + \delta)}, & a_{12} &= 0, & a_{13} &= 0, \\
 a_{14} &= \frac{\varphi(\Delta t)^2\delta\alpha(1 + \varphi(\Delta t)(\mu + \beta))}{(1 + \varphi(\Delta t)(2\mu + \alpha))(1 + \varphi(\Delta t)(2\mu + \delta))(1 + 2\varphi(\Delta t)\mu)}, \\
 a_{15} &= \frac{\varphi(\Delta t)\delta(1 + \varphi(\Delta t)\mu)}{(1 + \varphi(\Delta t)(2\mu + \delta))(1 + 2\varphi(\Delta t)\mu)}, \\
 a_{16} &= \frac{1 + \varphi(\Delta t)(\mu - \gamma)}{1 + 2\varphi(\Delta t)\mu}.
 \end{aligned}$$

The eigenvalues of  $J(E_0)$  are

$$\begin{aligned}
 \lambda_1 &= \frac{1 + \varphi(\Delta t)\mu}{1 + 2\varphi(\Delta t)\mu}, & \lambda_2 &= \frac{1 + \varphi(\Delta t)\mu + \varphi(\Delta t)\beta}{1 + 2\varphi(\Delta t)\mu + \varphi(\Delta t)\alpha}, \\
 \lambda_3 &= \frac{1 + \varphi(\Delta t)\mu}{1 + 2\varphi(\Delta t)\mu + \varphi(\Delta t)\delta}, & \lambda_4 &= \frac{1 + \varphi(\Delta t)\mu - \varphi(\Delta t)\gamma}{1 + 2\varphi(\Delta t)\mu}.
 \end{aligned}$$

**Table 2**

Qualitative behavior with respect to  $E_0$  of the schemes considered on the problem (3.1) with  $\beta = 50$  and different step sizes  $\Delta t$ ,  $T = 60$ .

$\Delta t$	ode45	RK4	Euler	NSFD – $\Delta t$	NSFD – $\varphi(\Delta t)$
0.01	Convergence	Convergence	Convergence	Convergence	Convergence
0.1	Convergence	Convergence	Divergence	Convergence	Convergence
1	Divergence	Divergence	Divergence	Convergence	Convergence
2	Divergence	Divergence	Divergence	Convergence	Convergence
3	Divergence	Divergence	Divergence	Convergence	Convergence
4	Divergence	Divergence	Divergence	Divergence	Convergence
5	Divergence	Divergence	Divergence	Divergence	Convergence
10	Divergence	Divergence	Divergence	Divergence	Convergence
50	Divergence	Divergence	Divergence	Divergence	Convergence
100	Divergence	Divergence	Divergence	Divergence	Convergence

It is clear that  $|\lambda_1| < 1$ ,  $|\lambda_3| < 1$ ,  $|\lambda_4| < 1$  and if  $R_0 < 1$  then  $|\lambda_2| < 1$  too, and therefore  $E_0 = (1, 0, 0, 0)$  is stable.

It is fair to say that for  $E^*$  we have no formal proof. But, the numerical results obtained by using the Math Toolbox software of MATLAB show that for any step-size  $\Delta t > 0$  the equilibrium point  $(S^*, I^*, R^*, C^*)$  is stable (see Figs. 1–3 and Tables 3–4). These results guarantee the dynamical consistency between system (3.1) and the numerical scheme (4.1) around all the equilibrium points. Therefore, the new proposed scheme (4.1) is elementary stable.

**Positivity:** With positivity, we mean that the component-wise non-negativity of the initial vector is preserved in time for the approximated solution. Assuming  $(S_0, I_0, R_0, C_0) \geq 0$ , since all of the parameters are positive then  $S_{i+1} > 0$  and  $I_{i+1} > 0$ . Also if  $\sigma < 1$ , then from (4.4) we have  $R_{i+1} > 0$ . Now for the positivity of  $C_{i+1}$ , it is sufficient to have

$$1 + \varphi(\Delta t)\mu - \varphi(\Delta t)\beta I_i - \varphi(\Delta t)\gamma \geq 0.$$

Since  $I_i \leq 1$  it is sufficient to have

$$1 + \varphi(\Delta t)\mu - \varphi(\Delta t)\beta - \varphi(\Delta t)\gamma \geq 0,$$

which is equivalent to

$$1 + \varphi(\Delta t)(\mu - \beta - \gamma) \geq 0,$$

and

$$\varphi(\Delta t) \geq \frac{1}{\beta + \gamma - \mu}, \quad \mu \leq \beta + \gamma.$$

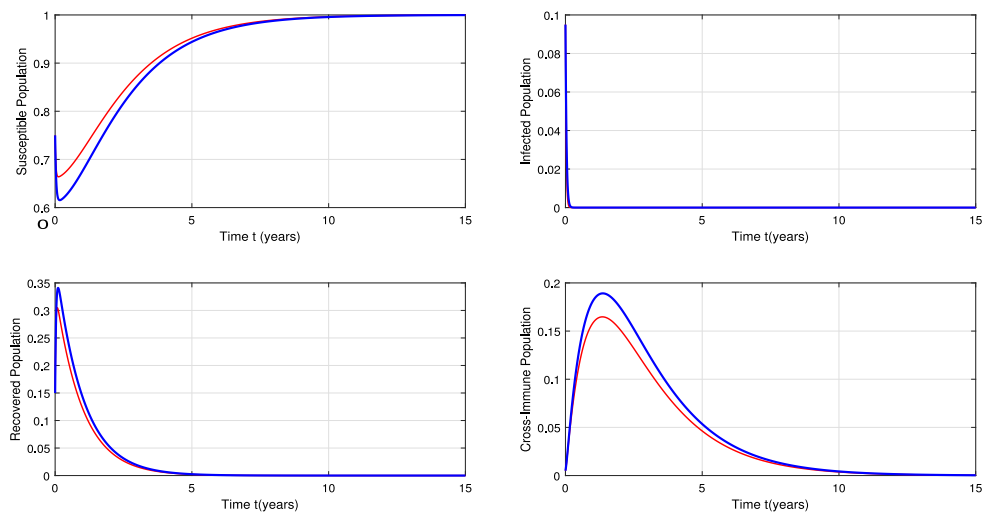
Therefore the new proposed scheme is positive and elementary stable and this completes the proof.  $\square$

### 5. Numerical results

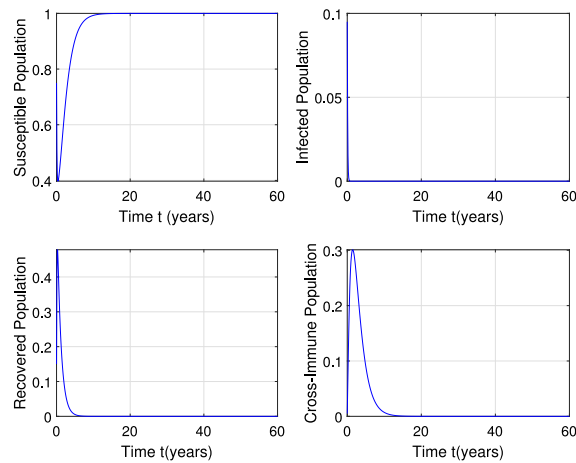
In this section, we present some numerical results to verify the properties of the proposed scheme and compare its performance with other methods available in the literature, namely, the RK4 method, ode45, ode15s, the NSFD method presented in [13] and the NSFD method presented in [26]. All the parameter values used in these simulations have been taken from [26] and we have considered  $\varphi(\Delta t) = \tanh(\Delta t)$ . For each experiment, the final value of the integration interval  $[t_0, T]$  is specified on the graphs or the corresponding table.

It can be seen in Fig. 1 that the proposed scheme and the RK4 method with  $\Delta t = 0.01$  preserve the stability of the equilibrium  $E_0$ . Furthermore, our new scheme converges to  $E_0$  for large step-sizes, as can be seen in Figs. 2–5. By increasing the step-size, we can observe that ode45, the RK4 and Euler methods diverge, whereas NSFD- $\Delta t$  converges for larger moderate values of  $\Delta t$  (until  $\Delta t = 4$ ) and the NSFD- $\varphi(\Delta t)$  scheme converges for all step sizes. Table 2 shows the qualitative behavior of the considered schemes for different values of the step size.

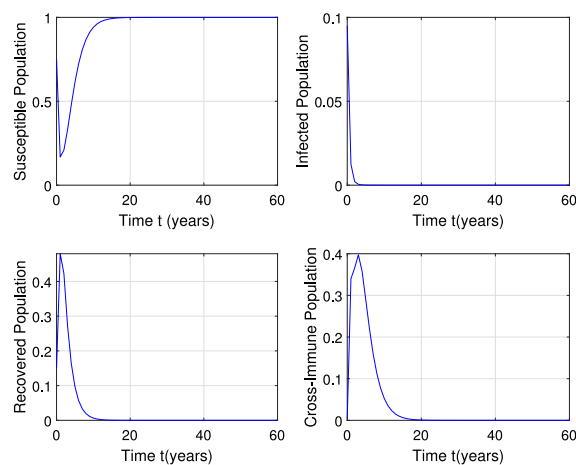




**Fig. 1.** Numerical results of the problem (3.1) by ode15s and the new scheme taking  $\Delta t = 0.01$  and initial values  $(S(0), I(0), C(0), R(0)) = (0.75, 0.095, 0.005, 0.15)$  with  $\beta = 50$  ( $R_0 < 1$ ).



**Fig. 2.** Numerical results of the problem (3.1) by the new scheme taking  $\Delta t = 0.1$  and initial values  $(S(0), I(0), C(0), R(0)) = (0.75, 0.095, 0.005, 0.15)$  with  $\beta = 50$  ( $R_0 < 1$ ).



**Fig. 3.** Numerical results of the problem (3.1) by the new scheme taking  $\Delta t = 1$  and initial values  $(S(0), I(0), C(0), R(0)) = (0.75, 0.095, 0.005, 0.15)$  with  $\beta = 50$  ( $R_0 < 1$ ).

**Table 3**

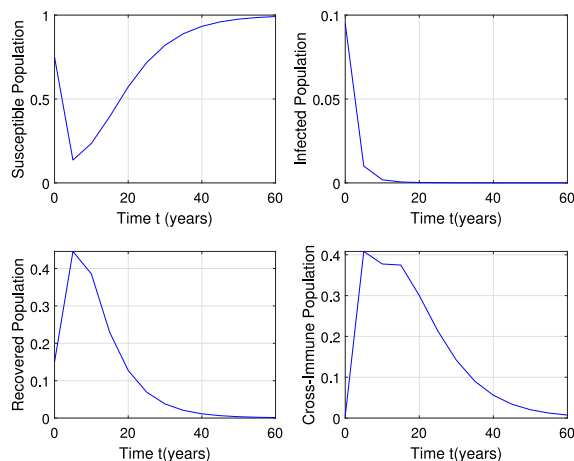
Qualitative behavior with respect to  $E^*$  of the schemes considered on the problem (3.1) for different step sizes with  $\beta = 100$ ,  $T = 45$ .

$\Delta t$	<i>ode45</i>	<i>RK4</i>	<i>Euler</i>	<i>NSFD</i> – $\Delta t$	<i>NSFD</i> – $\varphi(\Delta t)$
0.01	Divergence	Divergence	Divergence	Convergence	Convergence
0.1	Divergence	Divergence	Divergence	Convergence	Convergence
1	Divergence	Divergence	Divergence	Convergence	Convergence
2	Divergence	Divergence	Divergence	Convergence	Convergence
3	Divergence	Divergence	Divergence	Convergence	Convergence
4	Divergence	Divergence	Divergence	Divergence	Convergence
5	Divergence	Divergence	Divergence	Divergence	Convergence
10	Divergence	Divergence	Divergence	Divergence	Convergence
50	Divergence	Divergence	Divergence	Divergence	Convergence
100	Divergence	Divergence	Divergence	Divergence	Convergence

**Table 4**

Spectral radius of the Jacobian matrix with respect to  $E^*$  with  $\beta = 100$  and the parameter values in Table 1.

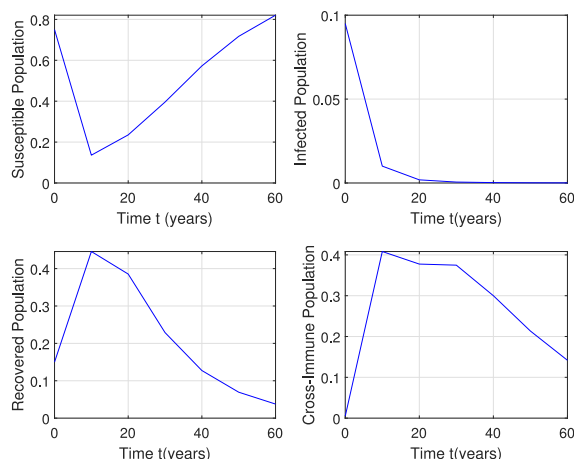
$\Delta t$	$\rho - NSFD - \Delta t$	$\rho - NSFD - \varphi(\Delta t)$
0.001	0.9999-Convergence	0.9999-Convergence
0.01	0.9997-Convergence	0.9997-Convergence
0.05	0.9989-Convergence	0.9989-Convergence
0.1	0.9979-Convergence	0.9980-Convergence
0.5	0.9901-Convergence	0.9921-Convergence
1	0.9806-Convergence	0.9875-Convergence
2	0.9628-Convergence	0.9831-Convergence
4	0.9308-Convergence	0.9831-Convergence
10	0.8567-Divergence	0.9806-Convergence
20	0.7771-Divergence	0.9806-Convergence
100	0.6760-Divergence	0.9806-Convergence



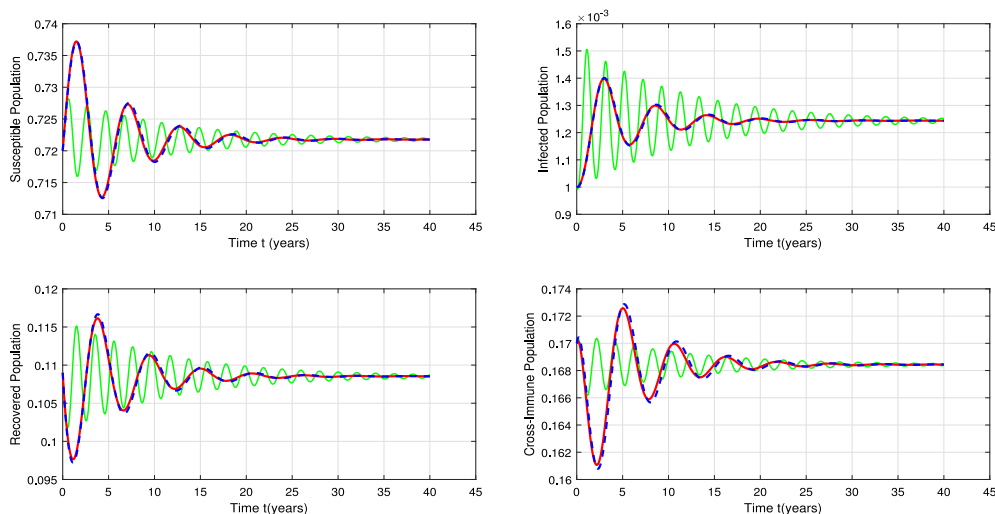
**Fig. 4.** Numerical results of the problem (3.1) by the new scheme taking  $\Delta t = 5$  and initial values  $(S(0), I(0), C(0), R(0)) = (0.75, 0.095, 0.005, 0.15)$  with  $\beta = 50$  ( $R_0 < 1$ ).

In the previous examples we have only emphasized the qualitative behavior of the solutions. It is obvious that the smaller the step size, the smaller the errors involved. In what follows we will show this aspect with respect to the equilibrium point  $E^*$ .

Fig. 6 shows that the new method preserves the stability of  $E^*$  for small step sizes. Similar behavior occurs for the NSFD method presented in [26] with  $\Delta t = 0.01$ , but when increasing the step size the scheme (4.1) converges to



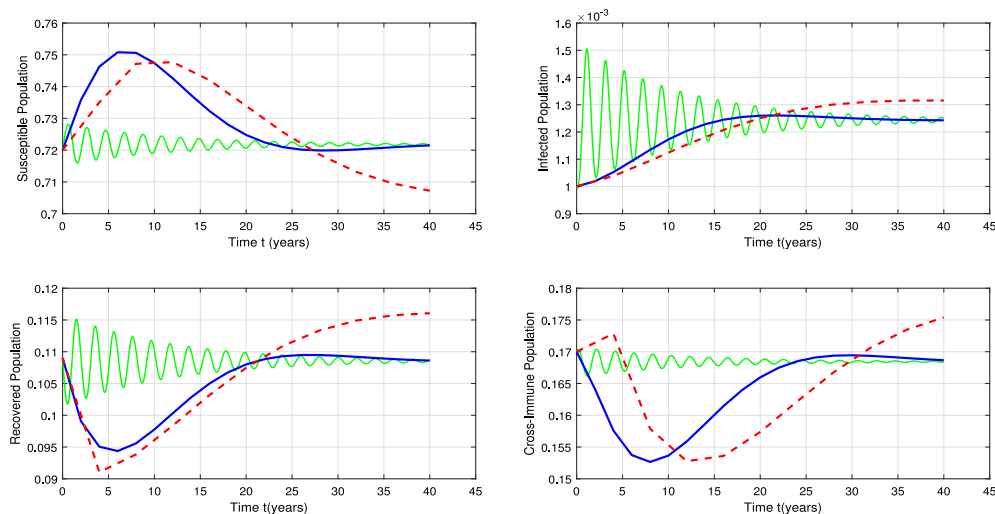
**Fig. 5.** Numerical results of the problem (3.1) by the new scheme taking  $\Delta t = 10$  and initial values  $(S(0), I(0), C(0), R(0)) = (0.75, 0.095, 0.005, 0.15)$  with  $\beta = 50$  ( $R_0 < 1$ ).



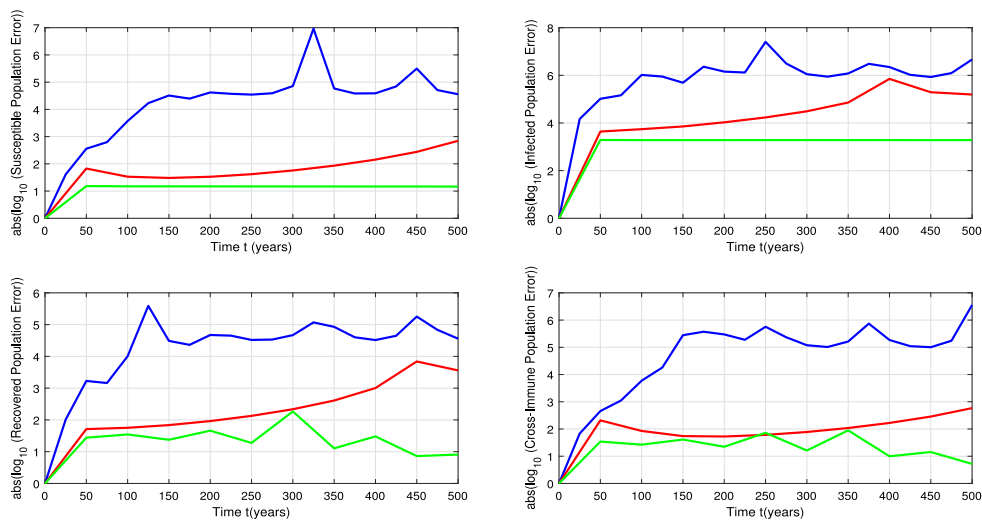
**Fig. 6.** Oscillatory behavior of the solution of problem (3.1) by ode15s (green line) taking  $\Delta t = 0.01$ , which is used as a reference solution. The proposed method in [13] (blue dashed line) and the new scheme (red line) present also this oscillatory behavior taking  $\Delta t = 0.1$ . The initial values are  $(S(0), I(0), C(0), R(0)) = (0.72, 0.001, 0.17, 0.109)$  with  $\beta = 100$  ( $R_0 > 1$ ). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the equilibrium  $E^*$  more accurately than the other methods. This can be seen in Fig. 7. Furthermore, the qualitative behavior of the considered schemes for different step sizes with respect to  $E^*$  is presented in Table 3. It can be seen that ode45, the RK4 and Euler methods diverge for all step-sizes but NSFD- $\varphi(\Delta t)$  is convergent. In Table 4 we observe that the spectral radius of the Jacobian matrix associated to the new scheme with respect to  $E^*$  are less than one showing that the scheme (4.1) is stable. Since we do not have an analytic solution for the nonlinear problem in (3.1), we use as a reference solution the one calculated with ODE15s method to represent our *true solution*. Figs. 8–11 include the absolute errors for different schemes and different values of  $\Delta t$ , showing that the proposed scheme is more accurate than the other methods.

Numerical simulations were developed for different representative values of  $R_0$  which can cover most of the possible realistic values. In Tables 3–4 we present some qualitative results and it can be observed that Scheme NSFD- $\varphi(\Delta t)$  converges to the equilibrium point for all the numerical simulations.



**Fig. 7.** Oscillatory behavior of the problem (3.1) by ode15s (green line) with  $\Delta t = 0.01$ , which is used as a reference solution. The proposed method in [13] (red dashed line) and the new scheme (blue line) present also this oscillatory behavior taking  $\Delta t = 2$  and initial values  $(S(0), I(0), C(0), R(0)) = (0.72, 0.001, 0.17, 0.109)$  with  $\beta = 100$  ( $R_0 > 1$ ). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

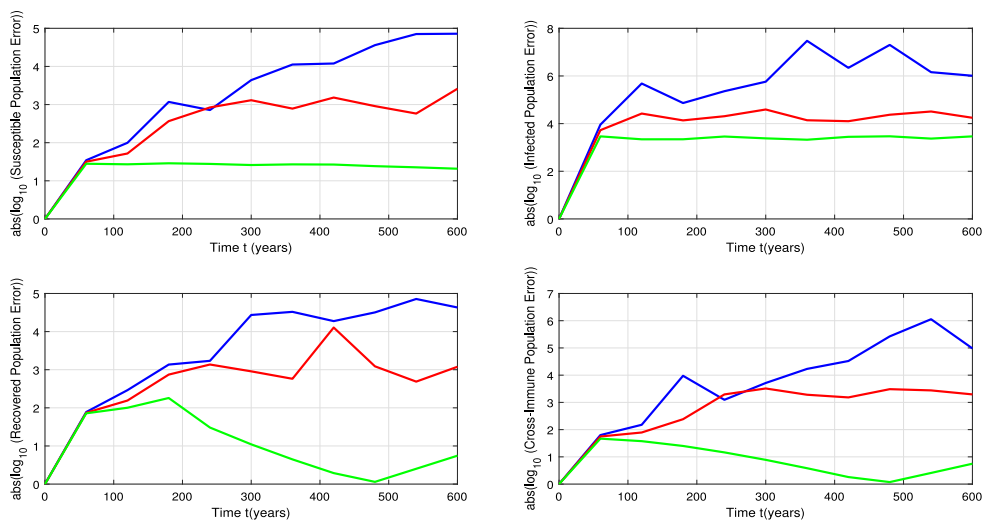


**Fig. 8.** Absolute errors for the problem (3.1) with  $\Delta t = 0.01$  by the new scheme (blue line), the proposed method in [13] (red line) and the proposed method in [26] (green line) taking  $\Delta t = 2$  and initial values  $(S(0), I(0), C(0), R(0)) = (0.72, 0.001, 0.17, 0.109)$  with  $\beta = 100$  ( $R_0 > 1$ ), using ode15s as a reference solution. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

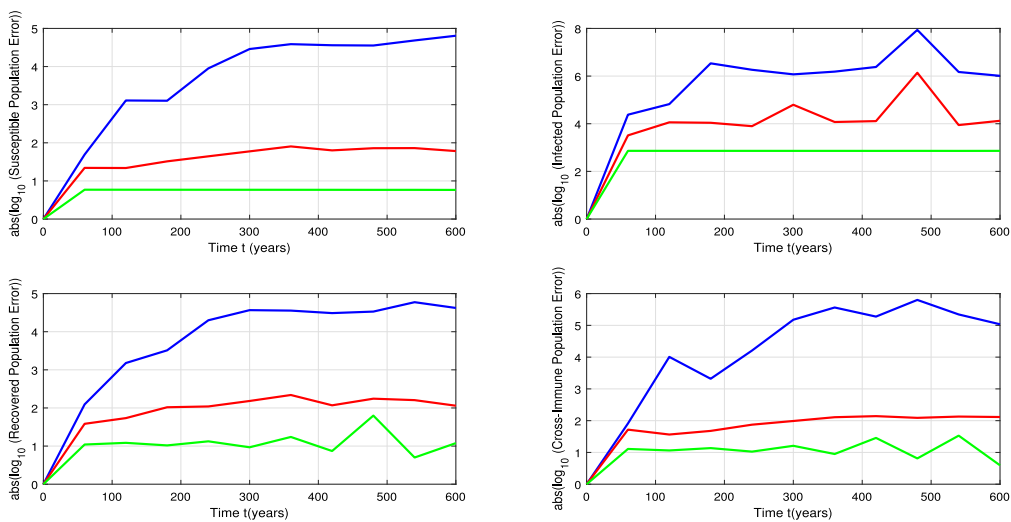
As in the work of Gumel et al. [10], convergence of the proposed scheme has not been proved but based on all the developed numerical simulations, it seems to be unconditionally convergent to the equilibrium  $E^*$  of the SIRC model.

**Conclusion**

In this article, a nonstandard discretization approach is applied to solve numerically the influenza disease model analyzed in [26]. The new proposed scheme preserves the stability of all equilibrium points and the positivity of



**Fig. 9.** Absolute errors of the problem (3.1) by the new scheme and the NSFD presented in [13] with  $\Delta t = 10$ , using ode15s as a reference solution with the initial values  $(S(0), I(0), C(0), R(0)) = (0.72, 0.001, 0.17, 0.109)$  and  $\beta = 100$ .

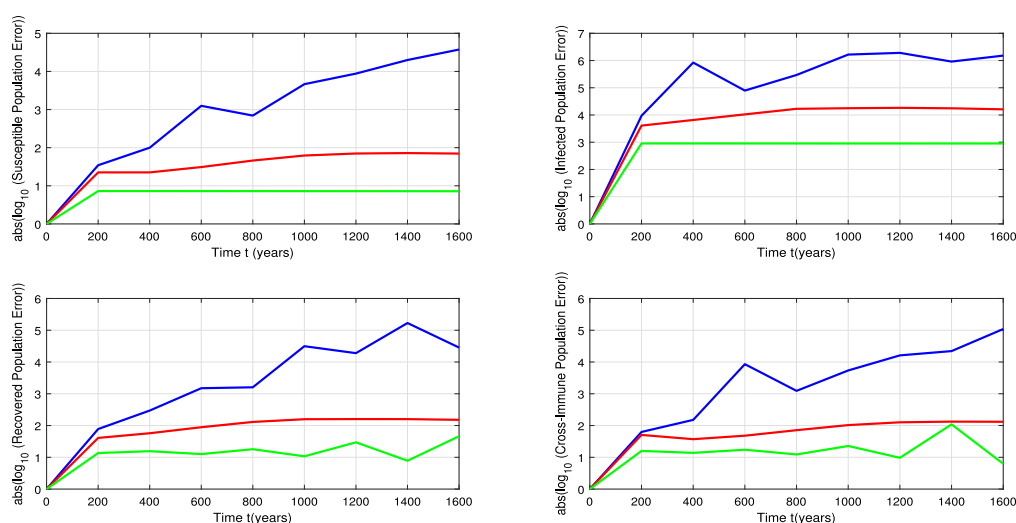


**Fig. 10.** Absolute errors of the problem (3.1) by the new scheme and the NSFD presented in [13] with  $\Delta t = 15$ , using ode15s as a reference solution with the initial values  $(S(0), I(0), C(0), R(0)) = (0.72, 0.001, 0.17, 0.109)$  and  $\beta = 100$ .

solutions. Compared with the RK4, ode15s, the NSFD method presented in [26] and the NSFD method presented in [13], we show that the proposed scheme improves the accuracy and presents a better qualitative behavior for large step-sizes.

**Acknowledgments**

The authors wish to thank the anonymous referees for the comments that greatly improved the manuscript.



**Fig. 11.** Absolute errors of the problem (3.1) by the new scheme and the NSFD presented in [13] with  $\Delta t = 100$ , using ode15s as a reference solution with the initial values  $(S(0), I(0), C(0), R(0)) = (0.72, 0.001, 0.17, 0.109)$  and  $\beta = 100$ .

## References

- [1] S. Abbas, M. Benchohra, N. Hamidi, J.J. Nieto, Hilfer and Hadamard fractional differential equations in Fréchet spaces, *TWMS J. Pure Appl. Math.* 10 (1) (2019) 102–116.
- [2] R. Anguelov, J.M.S. Lubuma, Contributions to the mathematics of the nonstandard finite difference method and applications, *Numer. Methods Partial Differential Equations* 17 (2001) 518–543.
- [3] A. Ashyralyev, D. Agirseven, R.P. Agarwal, Stability estimates for delay parabolic differential and difference equations, *Appl. Comput. Math.* 19 (2) (2020) 175–204.
- [4] A. Ashyralyev, A.S. Erdogan, S.N. Tekalan, An investigation on finite difference method for the first order partial differential equation with the nonlocal boundary condition, *Appl. Comput. Math.* 18 (3) (2019) 247–260.
- [5] R. Casagrandi, L. Bolzoni, S.A. Levin, V. Andreasen, The SIRC model and influenza A, *Math. Biosci.* 200 (2006) 152–169.
- [6] W. Chen, C. Wang, X. Wang, S.M. Wise, Positivity-preserving, energy stable numerical schemes for the Cahn–Hilliard equation with logarithmic potential, *J. Comput. Phys.* X3 (2019) 100031.
- [7] L. Dong, C. Wang, H.A. Zhang, Z. Zhang, A positivity-preserving, energy stable and convergent numerical scheme for the Cahn–Hilliard equation with a flory-huggins-degennes energy, *Commun. Math. Sci.* 17 (4) (2019) 921–939.
- [8] T. Gadjiev, S. Aliev, Sh. Galandarova, A priori estimates for solutions to Dirichlet boundary value problems for polyharmonic equations in generalized Morrey spaces, *TWMS J. Pure Appl. Math.* 9 (2) (2018) 231–242.
- [9] A. Golbabai, O. Nikan, M. Molavi-Arabshahi, Numerical approximation of time fractional advection–dispersion model arising from solute transport in rivers, *TWMS J. Pure Appl. Math.* 10 (1) (2019) 117–131.
- [10] A.B. Gumel, T.D. Loewena, P.N. Shivakumara, B.M. Sahaib, P. Yuc, M.L. Garbad, Numerical modelling of the perturbation of HIV-1 during combination anti-retroviral therapy, *Comput. Biol. Med.* 31 (2001) 287–301.
- [11] J.K. Hale, *Ordinary Differential Equations*, Wiley-Interscience, New York, 1969.
- [12] S. Harikrishnan, K. Kanagarajan, E.M. Elsayed, Existence and stability results for differential equations with complex order involving Hilfer fractional derivative, *TWMS J. Pure Appl. Math.* 10 (1) (2019) 94–101.
- [13] L. Jódar, R.J. Villanueva, A.J. Arenas, G. González, Nonstandard numerical methods for a mathematical model for influenza disease, *Math. Comput. Simulation* 79 (2008) 622–633.
- [14] M.M. Khader, Sweilam, A.M.S. Mahdy, N.K. Abdel Moniem, Numerical simulation for the fractional SIRC model and influenza A, *Appl. Math. Inf. Sci.* 8 (3) (2014) 1029–1036.
- [15] R.J. LeVeque, *Numerical Methods for Conservation Laws*, Birkhauser-Verlag, Basel, Boston, Berlin, 1992.
- [16] M. Mehdizadeh Khalsaraei, An improvement on the positivity results for 2-stage explicit Runge–Kutta methods, *J. Comput. Appl. Math.* 235 (1) (2010) 137–143.
- [17] M. Mehdizadeh Khalsaraei, Positivity of an explicit Runge–Kutta method, *Ain Shams Eng. J.* 6 (4) (2015) 1217–1223.
- [18] M. Mehdizadeh Khalsaraei, F. Khodadoosti, A new total variation diminishing implicit nonstandard finite difference scheme for conservation laws, *Comput. Methods Diff. Equ.* 2 (2014) 85–92.
- [19] M. Mehdizadeh Khalsaraei, F. Khodadoosti, Qualitatively stability of nonstandard 2-stage explicit Runge–Kutta methods of order two, *Compu. Math. Phys.* 56 (2) (2016) 235–242.
- [20] M. Mehdizadeh Khalsaraei, A. Shokri, A new explicit singularly P-stable four-step method for the numerical solution of second order IVPs, *Iran. J. Math. Chem.* 11 (1) (2020) 17–31.

- [21] M. Mehdizadeh Khalsaraei, A. Shokri, The new classes of high order implicit six-step P-stable multiderivative methods for the numerical solution of Schrödinger equation, *Appl. Comput. Math.* 19 (1) (2020) 59–86.
- [22] M. Mehdizadeh Khalsaraei, A. Shokri, M. Molayi, The new high approximation of stiff systems of first ordinary IVPs arising from chemical reactions by k-step L-stable hybrid methods, *Iran. J. Math. Chem.* 10 (2) (2019) 181–193.
- [23] M. Mehdizadeh Khalsaraei, R. Shokri Jahandizi, Positivity-preserving nonstandard finite difference schemes for simulation of advection-diffusion reaction equations, *Comput. Methods Diff. Equ.* 2 (4) (2014) 256–267.
- [24] M. Mehdizadeh Khalsaraei, R. Shokri Jahandizi, Efficient explicit nonstandard finite difference scheme with positivity-preserving property, *GU J. Sci.* 30 (1) (2017) 259–268.
- [25] R.E. Mickens, *Nonstandard Finite Difference Models of Differential Equations*, World Scientific, Singapore, 1994.
- [26] R.E. Mickens, Numerical integration of population models satisfying conservation laws: NSFD methods, *Biol. Dyn.* 1 (4) (2007) 1751–1766.
- [27] R.E. Mickens, P.M. Jordan, A positivity-preserving nonstandard finite difference scheme for the damped wave equation, *Numer. Methods Partial Differential Equations* 20 (2004) 639–649.
- [28] M.I. Modebei, R.B. Adeniyi, S.N. Jator, H. Ramos, A block hybrid integrator for numerically solving fourth-order initial value problems, *Appl. Math. Comput.* 346 (2019) 680–694.
- [29] Z. Odibat, Fractional power series solutions of fractional differential equations by using generalized Taylor series, *Appl. Comput. Math.* 19 (1) (2020) 47–58.
- [30] S. Qureshi, H. Ramos, L-stable explicit nonlinear method with constant and variable step-size formulation for solving initial value problems, *Int. J. Nonlinear Sci. Numer. Simul.* 19 (7–8) (2018) 741–751.
- [31] H. Ramos, Development of a new Runge–Kutta method and its economical implementation, *Comput. Math. Methods* 1 (2) (2019) e1016.
- [32] H. Ramos, P. Popescu, How many k-step linear block methods exist and which of them is the most efficient and simplest one? *Appl. Math. Comput.* 316 (2018) 296–309.
- [33] H. Ramos, M.A. Rufai, Third derivative modification of k-step block Falkner methods for the numerical solution of second order initial-value problems, *Appl. Math. Comput.* 333 (2018) 231–245.
- [34] H. Ramos, M.A. Rufai, A third-derivative two-step block Falkner-type method for solving general second-order boundary-value systems, *Math. Comput. Simulation* 165 (2019) 139–155.
- [35] H. Ramos, M.A. Rufai, Numerical solution of boundary value problems by using an optimized two-step block method, *Numer. Algorithms* 84 (1) (2020) 229–251.
- [36] H. Ramos, G. Singh, A tenth order A-stable two-step hybrid block method for solving initial value problems of ODEs, *Appl. Math. Comput.* 310 (2017) 75–88.
- [37] A. Shokri, M. Mehdizadeh Khalsaraei, A. Atashyar, A new two-step hybrid singularly P-stable method for the numerical solution of second-order IVPs with oscillating solutions, *Iran. J. Math. Chem.* 11 (2) (2020) 113–132.
- [38] A. Shokri, M. Tahmourasi, A new two-step Obrechhoff method with vanished phase-lag and some of its derivatives for the numerical solution of radial Schrödinger equation and related IVPs with oscillating solutions, *Iran. J. Math. Chem.* 8 (2) (2017) 137–159.
- [39] M.F. Simões Patrício, M. Patrício, H. Ramos, Extrapolating for attaining high precision solutions for fractional partial differential equations, *Fract. Calc. Appl. Anal.* 21 (6) (2018) 1506–1523.
- [40] M.F. Simões Patrício, H. Ramos, M. Patrício, Solving initial and boundary value problems of fractional ordinary differential equations by using collocation and fractional powers, *J. Comput. Appl. Math.* 354 (2019) 348–359.
- [41] G. Singh, H. Ramos, An optimized two-step hybrid block method formulated in variable step-size mode for integrating  $y = f(x, y, y')$  numerically, *Numer. Math. Theory Methods Appl.* 12 (2) (2019) 640–660.

# Conclusion and outlook

## Conclusion

In this thesis we mainly presented our results on investigating various cross-diffusion systems from both theoretical and numerical point of view. A very important aspect of these type of the systems is the presence of the cross-diffusion term(s), which often complicates the analytical and numerical investigations. We started this work with a brief introduction to the definition and some studied background for the cross-diffusion systems. Next, we mentioned that the strength of the cross-diffusion term(s) in the system can strongly affect the behavior of their approximate solutions, which usually leads to spurious oscillations or even blow-up in the discrete solutions whenever the cross-diffusion term(s) are dominant and standard non-adaptive discretization techniques are used. Since, this process resembles the convection-dominated regime in the convection-diffusion-reaction equations for which a huge amount of the stabilization methods are available, hence we briefly recalled some of these methods in Chapter 1. Next, we studied a chemotaxis-type cross-diffusion system modeling a cancer invasion in Chapter 2, for which we established theoretical proofs, numerical algorithms and numerical simulations. The main aspect of the considered system was lack of spatial regularity in both second and third equations which made the considered system more challenging to deal with both from the theoretical and numerical view points. In the theoretical part, making use of parabolic regularity theory, the existence of global classical solutions was shown in two- and three- dimensional bounded domains. The proof was established provided that the second and third equations in the system were at least regularized in time. In the numerical part, the numerical stability of the system was investigated. It was shown that, spurious oscillation and numerical blow-up occurs in the cross-diffusion-dominated regime whenever the standard Galerkin finite element method is used along with  $\theta$ -scheme for discretization in space and time, respectively. Lastly, the theoretical results were supported by various numerical experiments in two- and three-dimensions utilizing finite element library Deal.II. In Chapter 3, we considered the haptotaxis counterpart of the aforementioned system for which the techniques considered in the previous chapter were no longer applicable and proving the existence of the solution from theoretical point of view is still an open question and calls for further investigation. Hence, we addressed this point by means of numerical schemes. In this regard, a high-resolution nonlinear stabilized finite element flux-corrected transport method was employed for spatial discretization combined with  $\theta$ -method for temporal discretization. Making use of Brouwer's fixed point theorem it was proved that both the nonlinear scheme and the linearized system used in the fixed-point iterations are solvable and positivity-preserving. Several numerical simulations were carried out in two dimensions to demonstrate the performance of the proposed methods, where it was observed that the usage of high-resolution scheme simply allows high-accuracy for smooth regions and good oscillation diminishing in non-regular regime. In the next chapter, Chapter 4, we considered a system consisting of a double cross-diffusion terms modeling two rivaling gangs. The main feature of this system was the presence of cross-diffusion



term in two different equations in the system which made the analytical and numerical investigations even more challenging. From the analytical point of view, we proved that there exist a global, bounded classical solution. Moreover, we showed that for sufficiently small initial data these solutions converge toward homogeneous steady states, however, it was shown that obtaining such a results for large data seemed to be very difficult theoretically. Once again we addressed this difficulty by means of numerical methods. In this regard, we employed stabilized finite element flux-corrected transport method along with  $\theta$ -scheme for spatial and temporal discretization, respectively. We proved that the proposed method is positivity preserving and satisfies the discrete maximum principle under certain assumptions. Utilizing finite element library Deal.II, we presented various numerical experiments to support our theoretical and numerical results. Lastly, we finished this thesis in Chapter 5 by investigating a strongly coupled nonlinear ordinary differential equation modeling influenza disease, where a new nonstandard finite difference method was employed to solve the system under consideration. It was proved that the proposed method is positivity-preserving and also elementary stable. The results were supported by providing some numerical simulations utilizing MATLAB.

## Outlook

As mentioned in the introduction numerical analysis and simulation of cross-diffusion systems are considerably rare compared to the theoretical analysis. Thus, in the following we present several ideas that are not addressed in this work and can naturally be considered for possible future developments in this regard:

- Studying more complicated model problems such as: systems which are consist of not only one cross-diffusion term in a single equation in the system but several cross-diffusion terms appearing in the same equation, systems consisting of cross-diffusion terms along with compressible/incompressible Navier-Stokes equations, or systems that combine fluid-structure interaction and cross-diffusion,
- Considering more complicated domains with moving boundaries,
- Exploring and developing numerical analysis and theory in more details for various model problems,
- Estimate the errors of the approximate solutions to cross-diffusion systems and derive a refinement process in cross-diffusion-dominated regime. Desirably developing robust a posteriori error estimators for the nonlinear finite element flux-corrected transport method in the FCT norm which are preferably independent of the choice of the limiter,
- Software development and implementation, for example, parallelization programming for the goal of efficient computation in the higher dimensional domain,
- Examining other iteration methods to treat the nonlinearities in the system, for example, Newton-like methods or a combination of Newton and fixed-

point approaches in order to be able to switch between the two schemes whenever it is necessary,

- Employing high-order stable time integrators such as multi-step or strong stability preserving Runge-Kutta methods. Utilizing adaptive time stepping techniques instead of uniform time stepping used in this work.

# Bibliography

- [1] L. Onsager and R. M. Fuoss. Irreversible processes in electrolytes. Diffusion, conductance and viscous flow in arbitrary mixtures of strong electrolytes. *The Journal of Physical Chemistry*, 36(11):2689–2778, 2002.
- [2] J. B. Duncan and H. L. Toor. An experimental study of three component gas diffusion. *journal of the American Institute of Chemical Engineers*, 8(1):38–41, 1962.
- [3] N. Shigesada, K. Kawasaki, and E. Teramoto. Spatial segregation of interacting species. *Journal of Theoretical Biology*, 79(1):83–99, 1979.
- [4] E. H. Kerner. Further considerations on the statistical mechanics of biological associations. *The Bulletin of Mathematical Biophysics*, 21:217–255, 1959.
- [5] A. R. Anderson and M. A. Chaplain. Continuous and discrete mathematical models of tumor-induced angiogenesis. *Bulletin of Mathematical Biology*, 60(5):857–899, 1998.
- [6] I. N. Figueiredo, C. Leal, G. Romanazzi, B. Engquist, and P. N. Figueiredo. A convection-diffusion-shape model for aberrant colonic crypt morphogenesis. *Computing and Visualization in Science*, 14:157–166, 2011.
- [7] T. L. Jackson and H. M. Byrne. A mathematical model to study the effects of drug resistance and vasculature on the response of solid tumors to chemotherapy. *Mathematical Biosciences*, 164(1):17–38, 2000.
- [8] L. Zhang, M. Yang, and M. Jiang. Mathematical modeling for convection-enhanced drug delivery. *Procedia Engineering*, 29:268–274, 2012.
- [9] E. F. Keller and L. A. Segel. Model for chemotaxis. *Journal of Theoretical Biology*, 30(2):225–234, 1971.
- [10] E. F. Keller and L. A. Segel. Initiation of slime mold aggregation viewed as an instability. *Journal of Theoretical Biology*, 26(3):399–415, 1970.
- [11] L. Chen and A. Jüngel. Analysis of a multidimensional parabolic population model with strong cross-diffusion. *SIAM Journal on Mathematical Analysis*, 36(1):301–322, 2004.
- [12] A. Jüngel and N. Zamponi. Qualitative behavior of solutions to cross-diffusion systems from population dynamics. *Journal of Mathematical Analysis and Applications*, 440(2):794–809, 2016.
- [13] B. Bozzini, D. Lacitignola, C. Mele, and I. Sgura. Coupling of morphology and chemistry leads to morphogenesis in electrochemical metal growth: a review of the reaction-diffusion approach. *Acta Applicandae Mathematicae*, 122:53–68, 2012.

- [14] K. J. Painter. Continuous models for cell migration in tissues and applications to cell sorting via differential chemotaxis. *Bulletin of Mathematical Biology*, 71:1117–1147, 2009.
- [15] D. E. Woodward, R. Tyson, M. R. Myerscough, J. D. Murray, E. O. Budrene, and H. C. Berg. Spatio-temporal patterns generated by salmonella typhimurium. *Biophysical Journal*, 68(5):2181–2189, 1995.
- [16] P. Domschke, D. Trucu, A. Gerisch, and M. A. Chaplain. Mathematical modelling of cancer invasion: implications of cell adhesion variability for tumour infiltrative growth patterns. *Journal of Theoretical Biology*, 361:41–60, 2014.
- [17] A. Gerisch and M. A. Chaplain. Robust numerical methods for taxis–diffusion–reaction systems: applications to biomedical problems. *Mathematical and Computer Modelling*, 43(1–2):49–75, 2006.
- [18] Y. S. Choi, Z. Huan, and R. Lui. Global existence of solutions of a strongly coupled quasilinear parabolic system with applications to electrochemistry. *Journal of Differential Equations*, 194(2):406–432, 2003.
- [19] L. Chen and A. Jüngel. Analysis of a parabolic cross-diffusion population model without self-diffusion. *Journal of Differential Equations*, 224(1):39–59, 2006.
- [20] T. Lepoutre, M. Pierre, and G. Rolland. Global well-posedness of a conservative relaxed cross-diffusion system. *SIAM Journal on Mathematical Analysis*, 44(3):1674–1693, 2012.
- [21] R. Redlinger. Existence of the global attractor for a strongly coupled parabolic system arising in population dynamics. *Journal of Differential Equations*, 118:219–252, 1995.
- [22] A. Jüngel. The boundedness-by-entropy method for cross-diffusion systems. *Nonlinearity*, 28(6):1963, 2015.
- [23] Y. Lou, W. M. Ni, and Y. Wu. On the global existence of a cross-diffusion system. *Discrete and Continuous Dynamical Systems*, 4:193–204, 1998.
- [24] Z. Wen and S. Fu. Global solutions to a class of multi-species reaction-diffusion systems with cross-diffusions arising in population dynamics. *Journal of Computational and Applied Mathematics*, 230(1):34–43, 2009.
- [25] X. Chen and A. Jüngel. Weak–strong uniqueness of renormalized solutions to reaction–cross-diffusion systems. *Mathematical Models and Methods in Applied Sciences*, 29(02):237–270, 2019.
- [26] M. Pierre. Global existence in reaction-diffusion systems with control of mass: a survey. *Milan Journal of Mathematics*, 78:417–255, 2010.
- [27] A. Jüngel. *Entropy methods for diffusive partial differential equations*. Springer, Berlin, 2016.

- [28] P. Degond, S. Génieys, and A. Jüngel. A system of parabolic equations in nonequilibrium thermodynamics including thermal and electrical effects. *Journal de Mathématiques Pures et Appliquées*, 76(10):991–1015, 1997.
- [29] A. Jüngel. The boundedness-by-entropy method for cross-diffusion systems. *Nonlinearity*, 28(6):1963, 2015.
- [30] M. Burger, M. Di Francesco, J.F. Pietschmann, and B. Schlake. Nonlinear cross-diffusion with size exclusion. *SIAM Journal on Mathematical Analysis*, 42:2842–2871, 2010.
- [31] C. Choquet, C. Rosier, and L. Rosier. Well posedness of general cross-diffusion systems. *Journal of Differential Equations*, 300:386–425, 2021.
- [32] D. Le. Some maximum principles for cross diffusion systems. *Preprint*, page arXiv:2304.08262., 2023.
- [33] M. Rascle and C. Ziti. Finite time blow-up in some models of chemotaxis. *Journal of Mathematical Biology*, 33:388–414, 1995.
- [34] T. Xiang. Finite time blow-up in the higher dimensional parabolic-elliptic-ODE minimal chemotaxis-haptotaxis system. *Journal of Differential Equations*, 336:44–72, 2022.
- [35] M. Fuest. Blow-up profiles in quasilinear fully parabolic Keller–Segel systems. *Nonlinearity*, 33(5):2306, 2020.
- [36] T. Cieślak and M. Winkler. Finite-time blow-up in a quasilinear system of chemotaxis. *Nonlinearity*, 21(5):1057, 2008.
- [37] K. Trivisa and F. Weber. A convergent explicit finite difference scheme for a mechanical model for tumor growth. *ESAIM: Mathematical Modelling and Numerical Analysis*, 51(1):35–62, 2017.
- [38] M. K. Kolev, M. N. Koleva, and L. G. Vulkov. An unconditional positivity-preserving difference scheme for models of cancer migration and invasion. *Mathematics*, 10(1):131, 2022.
- [39] H. Murakawa. Error estimates for discrete-time approximations of nonlinear cross-diffusion systems. *SIAM Journal on Numerical Analysis*, 52(2):955–974, 2014.
- [40] B. Andreianov, M. Bendahmane, and R. Ruiz-Baier. Analysis of a finite volume method for a cross-diffusion model in population dynamics. *Mathematical Models and Methods in Applied Sciences*, 21(02):307–344, 2011.
- [41] G. Zhou and N. Saito. Finite volume methods for a Keller–Segel system: discrete energy, error estimates and numerical blow-up analysis. *Numerische Mathematik*, 135(1):265–311, 2017.
- [42] A. Chertock and A. Kurganov. A second-order positivity preserving central-upwind scheme for chemotaxis and haptotaxis models. *Numerische Mathematik*, 111:169–205, 2008.

- [43] J. W. Barrett and J. F. Blowey. Finite element approximation of a nonlinear cross-diffusion population model. *Numerische Mathematik*, 98:195–221, 2004.
- [44] R. Strehl. *Advanced numerical treatment of chemotaxis driven PDEs in mathematical biology*. Doctoral dissertation. Universitätsbibliothek Dortmund, 2013.
- [45] X. Zheng, S. M. Wise, and V. Cristini. Nonlinear simulation of tumor necrosis, neo-vascularization and tissue invasion via an adaptive finite-element/level-set method. *Bulletin of Mathematical Biology*, 67:211–259, 2005.
- [46] N. Kolbe, J. Kat’uchová, N. Sfakianakis, N. Hellmann, and M. Lukáčová-Medvid’ová. A study on time discretization and adaptive mesh refinement methods for the simulation of cancer invasion: The urokinase model. *Applied Mathematics and Computation*, 273:353–376, 2016.
- [47] A. A. I. Quiroga, D. Fernandez, G. A. Torres, and C. V. Turner. Adjoint method for a tumor invasion PDE-constrained optimization problem in 2D using adaptive finite element method. *Applied Mathematics and Computation*, 270:358–368, 2015.
- [48] A. Sokolov, R. Ali, and S. Turek. An AFC-stabilized implicit finite element method for partial differential equations on evolving-in-time surfaces. *Journal of Computational and Applied Mathematics*, 289:101–115, 2015.
- [49] M. Sulman and T. Nguyen. A positivity preserving moving mesh finite element method for the Keller–Segel chemotaxis model. *Journal of Scientific Computing*, 80:649–666, 2019.
- [50] X. Huang, X. Xiao, J. Zhao, and X. Feng. An efficient operator-splitting FEM-FCT algorithm for 3D chemotaxis models. *Engineering with Computers*, 36:1393–1404, 2020.
- [51] R. Strehl, A. Sokolov, D. Kuzmin, and S. Turek. A flux-corrected finite element method for chemotaxis problems. *Computational Methods in Applied Mathematics*, 10(2):219–232, 2010.
- [52] A. Sokolov, R. Strehl, and S. Turek. Numerical simulation of chemotaxis models on stationary surfaces. *Discrete and Continuous Dynamical Systems - Series B*, 18(10):2689–2704, 2013.
- [53] G. DE Vahl Davis and G. D. Mallinson. An evaluation of upwind and central difference approximations by a study of recirculating flow. *Computers and Fluids*, 4(1):29–43, 1976.
- [54] P. M. Gresho and R. L. Lee. Don’t suppress the wiggles—they’re telling you something! *Computers and Fluids*, 9(2):223–253, 1981.
- [55] B. P. Leonard. A survey of finite differences of opinion on numerical muddling of the incomprehensible defective convection equation. *Finite Element Methods for Convection Dominated Flows*, 34:1–10, 1979.

- [56] H. G. Roos, M. Stynes, and L. Tobiska. *Robust numerical methods for singularly perturbed differential equations*. Springer-Verlag, Berlin, 2008.
- [57] K. E. Barrett. Finite element analysis for flow between rotating discs using exponentially weighted basis functions. *International Journal for Numerical Methods in Engineering*, 11(12):1809–1817, 1977.
- [58] P.J. Roache. *Computational fluid dynamics*. Computational Fluid Dynamics, Albuquerque: Hermosa, 1976.
- [59] I. Christie, D. F. Griffiths, A. R. Mitchell, and O. C. Zienkiewicz. Finite element methods for second order differential equations with significant first derivatives. *International Journal for Numerical Methods in Engineering*, 10(6):1389–1396, 1976.
- [60] J. C. Heinrich, P. S. Huyakorn, O. C. Zienkiewicz, and A. Mitchell. An'upwind'finite element scheme for two-dimensional convective transport equation. *International Journal for Numerical Methods in Engineering*, 11(1):131–143, 1977.
- [61] J. C. Heinrich and O. C. Zienkiewicz. Quadratic finite element schemes for two-dimensional convective-transport problems. *International Journal for Numerical Methods in Engineering*, 11(12):1831–1844, 1977.
- [62] M. Tabata. A finite element approximation corresponding to the upwind finite differencing. *Memoirs of Numerical Mathematics*, 4:47–63, 1977.
- [63] G. Ciarlet Philippe. *The finite element method for elliptic problems*. North-Holland, Amsterdam, 1978.
- [64] T.J. Hughes. A multidimensional upwind scheme with no crosswind diffusion. finite element methods for convection dominated flows. *AMD*, 34, 1979.
- [65] T. J. Hughes. A theoretical framework for Petrov-Galerkin methods with discontinuous weighting functions: Application to the streamline-upwind procedure. *Finite Element in Fluids*, 4:Chapter–3, 1982.
- [66] A. N. Brooks and T. J. Hughes. Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations. *Computer Methods in Applied Mechanics and Engineering*, 32(1–3):199–259, 1982.
- [67] M. Stynes and L. Tobiska. Necessary L2-uniform convergence conditions for difference schemes for two-dimensional convection-diffusion problems. *Computers and Mathematics with Applications*, 29(4):45–53, 1995.
- [68] Y. T. Shih and H. C. Elman. Modified streamline diffusion schemes for convection-diffusion problems. *Computer Methods in Applied Mechanics and Engineering*, 174(1–2):137–151, 1999.

- [69] N. Madden and M. Stynes. Linear enhancements of the streamline diffusion method for convection-diffusion problems. *Computers and Mathematics with Applications*, 32(10):29–42, 1996.
- [70] N. Madden and M. Stynes. Efficient generation of oriented meshes for solving convection–diffusion problems. *International Journal for Numerical Methods in Engineering*, 40(3):565–576, 1997.
- [71] P. Knobloch. On the choice of the SUPG parameter at outflow boundary layers. *Advances in Computational Mathematics*, 31:369–389, 2009.
- [72] V. John, P. Knobloch, and S. B. Savescu. A posteriori optimization of parameters in stabilized methods for convection-diffusion problems—Part I. *Computer Methods in Applied Mechanics and Engineering*, 200(41–44):2916–2929, 2011.
- [73] V. John, P. Knobloch, and U. Wilbrandt. A posteriori optimization of parameters in stabilized methods for convection-diffusion problems—Part II. *Journal of Computational and Applied Mathematics*, 428:115–167, 2023.
- [74] L. P. Franca, S. L. Frey, and T. J. Hughes. Stabilized finite element methods: I. Application to the advective-diffusive model. *Computer Methods in Applied Mechanics and Engineering*, 95(2):253–276, 1992.
- [75] H.G. Roos. Robust numerical methods for singularly perturbed differential equations: a survey covering 2008-2012. *International Scholarly Research Notices*, 1:379547, 2012.
- [76] B. Fischer, A. Ramage, D. J. Silvester, and A. J. Wathen. On parameter choice and iterative convergence for stabilized discretizations of advection–diffusion problems. *Computer Methods in Applied Mechanics and Engineering*, 179(1–2):179–195, 1999.
- [77] P. Knobloch. On the definition of the SUPG parameter. *Electronic Transactions on Numerical Analysis*, 32:76–89, 2008.
- [78] V. Jhon and P. Knobloch. On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part I—A review. *Computer Methods in Applied Mechanics and Engineering*, 196(17–20):2197–2215, 2007.
- [79] C. Johnson, U. Navert, and J. Pitkaranta. Finite element methods for linear hyperbolic problems. *Computer Methods in Applied Mechanics and Engineering*, 45:285–312, 1984.
- [80] U. Navert. *A finite element method for convection-dominated problems*. Ph.D. Thesis. Department of Computer Science, Chalmers University of Technology, Goteborg, Sweden, 1982.
- [81] T. J. Hughes and M. Mallet. A new finite element formulation for computational fluid dynamics: III. the generalized streamline operator for multidimensional advective-diffusive systems. *Computer Methods in Applied Mechanics and Engineering*, 58(3):305–328, 1986.



- [82] T. J. Hughes, L. P. Franca, and M. Balestra. A new finite element formulation for computational fluid dynamics: V. Circumventing the Babuška-Brezzi condition: A stable Petrov-Galerkin formulation of the Stokes problem accommodating equal-order interpolations. *Computer Methods in Applied Mechanics and Engineering*, 59(1):85–99, 1986.
- [83] L. P. Franca and S. L. Frey. Stabilized finite element methods: II. the incompressible Navier-Stokes equations. *Computer Methods in Applied Mechanics and Engineering*, 99(2–3):209–233, 1992.
- [84] T. Tezduyar and T. Hughes. Finite element formulations for convection dominated flows with particular emphasis on the compressible Euler equations. In *21st Aerospace Sciences Meeting*, page 125, 1983.
- [85] T. J. Hughes and T. Tezduyar. Finite element methods for first-order hyperbolic systems with particular emphasis on the compressible Euler equations. *Computer Methods in Applied Mechanics and Engineering*, 45(1–3):217–284, 1984.
- [86] T. J. Hughes, M. Mallet, and M. Akira. A new finite element formulation for computational fluid dynamics: II. Beyond SUPG. *Computer Methods in Applied Mechanics and Engineering*, 54(3):341–355, 1986.
- [87] T. J. Hughes and M. Mallet. A new finite element formulation for computational fluid dynamics: IV. A discontinuity-capturing operator for multidimensional advective-diffusive systems. *Computer Methods in Applied Mechanics and Engineering*, 58(3):329–336, 1986.
- [88] G. J. Le Beau and T. E. Tezduyar. Finite element computation of compressible flows with the SUPG formulation. *Advances in Finite Element Analysis in Fluid Dynamics ASME199*, 123:21–27, 1991.
- [89] T. J. Hughes, L. P. Franca, and G. M. Hulbert. A new finite element formulation for computational fluid dynamics: VIII. the Galerkin/least-squares method for advective-diffusive equations. *Computer Methods in Applied Mechanics and Engineering*, 73(2):173–189, 1989.
- [90] F. Shakib, T. J. Hughes, and Z. Johan. A new finite element formulation for computational fluid dynamics: X. The compressible Euler and Navier-Stokes equations. *Computer Methods in Applied Mechanics and Engineering*, 89(1–3):141–219, 1991.
- [91] I. Harari and T. J. Hughes. What are C and h?: Inequalities for the analysis and design of finite element methods. *Computer Methods in Applied Mechanics and Engineering*, 97(2):157–192, 1992.
- [92] D. N. Arnold, F. Brezzi, and M. Fortin. A stable finite element for the Stokes equations. *Calcolo*, 21(4):337–344, 1984.
- [93] L. P. Franca and R. Stenberg. Error analysis of Galerkin least squares methods for the elasticity equations. *SIAM Journal on Numerical Analysis*, 28(6):1680–1697, 1991.

- [94] C. Baiocchi, F. Brezzi, and L. P. Franca. Virtual bubbles and Galerkin-least-squares type methods (Ga.L.S). *Computer Methods in Applied Mechanics and Engineering*, 105(1):125–141, 1993.
- [95] R. Kumar and B. H. Dennis. A least-squares Galerkin split finite element method for compressible Navier-Stokes equations. *In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 48999:147–157, 2009.
- [96] J. Bonvin, M. Picasso, and R. Stenberg. GLS and EVSS methods for a three-field Stokes problem arising from viscoelastic flows. *Computer Methods in Applied Mechanics and Engineering*, 190(29–30):3893–3914, 2001.
- [97] I. Harari and T. J. Hughes. Galerkin/least-squares finite element methods for the reduced wave equation with non-reflecting boundary conditions in unbounded domains. *Computer Methods in Applied Mechanics and Engineering*, 98(3):411–454, 1992.
- [98] A. Masud and T. J. Hughes. A space-time Galerkin/least-squares finite element formulation of the Navier-Stokes equations for moving domain problems. *Computer Methods in Applied Mechanics and Engineering*, 146(1–2):91–126, 1997.
- [99] C. Johnson. *Numerical solution of partial differential equations by the finite element method*. Courier Corporation, 2012.
- [100] G. Lube. Stabilized Galerkin finite element methods for convection dominated and incompressible flow problems. *Banach Center Publications*, 29(1):85–104, 1994.
- [101] L. P. Franca and E. G. D. Do Carmo. The Galerkin gradient least-squares method. *Computer Methods in Applied Mechanics and Engineering*, 74(1):41–54, 1989.
- [102] I. Harari and T. J. Hughes. Stabilized finite element methods for steady advection—diffusion with production. *Computer Methods in Applied Mechanics and Engineering*, 115(1–2):165–191, 1994.
- [103] L. P. Franca and C. Farhat. Bubble functions prompt unusual stabilized finite element methods. *Computer Methods in Applied Mechanics and Engineering*, 123(1–4):299–308, 1995.
- [104] M. Lesoinne, C. Farhat, and L. P. Franca. Unusual stabilized finite element methods for second order linear differential equations. *Proceedings of the Ninth International Conference on Finite Elements in Fluids - New Trends and Applications*, 1995.
- [105] L. P. Franca and F. Valentin. On an improved unusual stabilized finite element method for the advective-reactive–diffusive equation. *Computer Methods in Applied Mechanics and Engineering*, 190(13–14):1785–1800, 2000.

- [106] L. P. Franca, C. Farhat, M. Lesoinne, and A. Russo. Unusual stabilized finite element methods and residual free bubbles. *International Journal for Numerical Methods in Fluids*, 27(1–4):159–168, 1998.
- [107] F. Brezzi, M. O. Bristeau, L. P. Franca, M. Mallet, and G. Roge. A relationship between stabilized finite element methods and the Galerkin method with bubble functions. *Computer Methods in Applied Mechanics and Engineering*, 96(1):117–129, 1992.
- [108] F. Brezzi and A. Russo. Choosing bubbles for advection-diffusion problems. *Mathematical Models and Methods in Applied Sciences*, 4(04):571–587, 1994.
- [109] F. Brezzi, D. Marini, and E. Suli. Residual-free bubbles for advection-diffusion problems: the general error analysis. *Numerische Mathematik*, 85(1):31–47, 2000.
- [110] A. Russo. Bubble stabilization of finite element methods for the linearized incompressible Navier-Stokes equations. *Computer Methods in Applied Mechanics and Engineering*, 132(3–4):335–343, 1996.
- [111] L. P. Franca and A. Russo. Deriving upwinding, mass lumping and selective reduced integration by residual-free bubbles. *Applied Mathematics Letters*, 9(5):83–88, 1996.
- [112] L. P. Franca and A. Russo. Mass lumping emanating from residual-free bubbles. *Computer Methods in Applied Mechanics and Engineering*, 142(3–4):353–360, 1997.
- [113] L. P. Franca and A. Russo. Unlocking with residual-free bubbles. *Computer Methods in Applied Mechanics and Engineering*, 142(3–4):361–364, 1997.
- [114] R. Becker and M. Braack. A finite element pressure gradient stabilization for the Stokes equations based on local projections. *Calcolo*, 38(4):173–199, 2001.
- [115] S. Ganesan and L. Tobiska. Stabilization by local projection for convection–diffusion and incompressible flow problems. *Journal of Scientific Computing*, 43:326–342, 2010.
- [116] P. Knobloch. On the application of local projection methods to convection–diffusion-reaction problems. In *BAIL 2008-Boundary and Interior Layers: Proceedings of the International Conference on Boundary and Interior Layers-Computational and Asymptotic Methods*, pages 183–194, 2009.
- [117] G. Matthies, P. Skrzypacz, and L. Tobiska. Stabilization of local projection type applied to convection-diffusion problems with mixed boundary conditions. *Electronic Transactions on Numerical Analysis*, 32:90–105, 2008.
- [118] M. Braack, E. Burman, V. John, and G. Lube. Stabilized finite element methods for the generalized Oseen problem. *Computer Methods in Applied Mechanics and Engineering*, 196(4–6):853–866, 2007.

- [119] R. Becker and M. Braack. A two-level stabilization scheme for the Navier-Stokes equations. *Numerical Mathematics and Advanced Applications*, pages 123–130, 2004.
- [120] R. Codina. Analysis of a stabilized finite element approximation of the Oseen equations using orthogonal subscales. *Applied Numerical Mathematics*, 58:264–283, 2008.
- [121] M. Braack and G. Lube. Finite elements with local projection stabilization for incompressible flow problems. *Journal of Computational Mathematics*, pages 116–147, 2009.
- [122] P. Knobloch. On a variant of the local projection method stable in the SUPG norm. *Kybernetika*, 45(4):634–645, 2009.
- [123] P. Knobloch. A generalization of the local projection stabilization for convection-diffusion-reaction equations. *SIAM Journal on Numerical Analysis*, 48(2):659–680, 2010.
- [124] P. Knobloch and L. Tobiska. On the stability of finite-element discretizations of convection–diffusion–reaction equations. *IMA Journal of Numerical Analysis*, 31(1):147–164, 2011.
- [125] G. Matthies, P. Skrzypacz, and L. Tobiska. A unified convergence analysis for local projection stabilizations applied to the Oseen problem. *ESAIM: Mathematical Modelling and Numerical Analysis*, 41(4):713–742, 2007.
- [126] S. Ganesan, G. Matthies, and L. Tobiska. Local projection stabilization of equal order interpolation applied to the Stokes problem. *Mathematics of Computation*, 77(264):2039–2060, 2008.
- [127] M. Braack and E. Burman. Local projection stabilization for the Oseen problem and its interpretation as a variational multiscale method. *SIAM Journal on Numerical Analysis*, 43(6):2544–2566, 2006.
- [128] G. Rapin, G. Lube and J. 2008 Löwe. Applying local projection stabilization to inf-sup stable elements. *Numerical Mathematics and Advanced Applications: Proceedings of ENUMATH 2007*, pages 521–525, 2009.
- [129] J. L. Guermond. Stabilization of Galerkin approximations of transport equations by subgrid modeling. *ESAIM: Mathematical Modelling and Numerical Analysis*, 33(6):1293–1316, 1999.
- [130] T. J. Hughes, L. Mazzei, and K. E. Jansen. Large eddy simulation and the variational multiscale method. *Computing and Visualization in Science*, 3:47–59, 2000.
- [131] V. John and S. Kaya. A finite element variational multiscale method for the Navier-Stokes equations. *SIAM Journal on Scientific Computing*, 26(5):1485–1503, 2005.

- [132] V. Gravemeier. The variational multiscale method for laminar and turbulent flow. *Archives of Computational Methods in Engineering*, 13:249–324, 2006.
- [133] P. Knobloch. Local projection method for convection-diffusion-reaction problems with projection spaces defined on overlapping sets. *Numerical Mathematics and Advanced Applications ENUMATH 2009*, pages 497–505, 2010.
- [134] T. Douglas, J. Dupont. Interior penalty procedures for elliptic and parabolic Galerkin methods, in: R. Glowinski, J.L. Lions (Eds.),. *Computing Methods in Applied Sciences*, pages 207–216, 1976.
- [135] E. Burman and P. Hansbo. Edge stabilization for Galerkin approximations of convection–diffusion-reaction problems. *Computer Methods in Applied Mechanics and Engineering*, 193(15–16):1437–1453, 2004.
- [136] E. Burman and P. Hansbo. Edge stabilization for the generalized Stokes problem: a continuous interior penalty method. *Computer Methods in Applied Mechanics and Engineering*, 195(19–22):2393–2410, 2006.
- [137] A. Bonito and E. Burman. A continuous interior penalty method for viscoelastic flows. *SIAM Journal on Scientific Computing*, 30(3):1156–1177, 2008.
- [138] E. Burman, M. A. Fernandez, and P. Hansbo. Continuous interior penalty finite element method for Oseen’s equations. *SIAM Journal on Numerical Analysis*, 44(3):1248–1274, 2006.
- [139] E. Burman and A. Ern. Continuous interior penalty hp-finite element methods for advection and advection-diffusion equations. *Mathematics of Computation*, 76(259):1119–1140, 2007.
- [140] A. Mizukami and T. J. Hughes. A Petrov-Galerkin finite element method for convection-dominated flows: an accurate upwinding technique for satisfying the maximum principle. *Computer Methods in Applied Mechanics and Engineering*, 50(2):181–193, 1985.
- [141] P. Knobloch. Improvements of the Mizukami–Hughes method for convection–diffusion equations. *Computer Methods in Applied Mechanics and Engineering*, 196(1–3):579–594, 2006.
- [142] P. Knobloch. Numerical solution of convection-diffusion equations using upwinding techniques satisfying the discrete maximum principle. *In Proceedings of Czech-Japanese Seminar in Applied Mathematics*, 3:69–76, 2005.
- [143] P. Knobloch. Application of the Mizukami–Hughes method to bilinear finite elements. *In Proceedings of Czech–Japanese Seminar in Applied Mathematics*, 6:137–147, 2006.
- [144] P. Knobloch. Numerical solution of convection–diffusion equations using a nonlinear method of upwind type. *Journal of Scientific Computing*, 43:454–470, 2010.

- [145] J. G. Rice and R. J. Schnipke. A monotone streamline upwind finite element method for convection-dominated flows. *Computer Methods in Applied Mechanics and Engineering*, 48:313–327, 1985.
- [146] H. Kanayama. Discrete models for salinity distribution in a bay: Conservation laws and maximum principle. *Theoretical and Applied Mechanics*, 28:559–579, 1980.
- [147] T. Ikeda. *Maximum principle in finite element models for convection-diffusion phenomena*, volume 4 of *Lecture Notes in Numerical and Applied Analysis*. North-Holland, Amsterdam, 1983.
- [148] T. E. Tezduyar and Y. J. Park. Discontinuity-capturing finite element formulations for nonlinear convection-diffusion-reaction equations. *Computer Methods in Applied Mechanics and Engineering*, 59:307–325, 1986.
- [149] A. C. Galeao and E. G. D. Do Carmo. A consistent approximate upwind Petrov-Galerkin method for convection-dominated problems. *Computer Methods in Applied Mechanics and Engineering*, 68(1):83–95, 1988.
- [150] E. G. D. Do Carmo and A. C. Galeao. Feedback Petrov-Galerkin methods for convection-dominated problems. *Computer Methods in Applied Mechanics and Engineering*, 88(1):1–16, 1991.
- [151] C. Johnson. Adaptive finite element methods for diffusion and convection problems. *Computer Methods in Applied Mechanics and Engineering*, 82(1–3):301–322, 1990.
- [152] R. C. Almeida and R. S. Silva. A stable Petrov-Galerkin method for convection-dominated problems. *Computer Methods in Applied Mechanics and Engineering*, 140(3–4):291–304, 1997.
- [153] P. A. B. De Sampaio and A. L. G. D. A. Coutinho. A natural derivation of discontinuity capturing operator for convection-diffusion problems. *Computer Methods in Applied Mechanics and Engineering*, 190(46–47):6291–6308, 2001.
- [154] E. G. D. Do Carmo and G. B. Alvarez. A new upwind function in stabilized finite element formulations, using linear and quadratic elements for scalar convection–diffusion problems. *Computer Methods in Applied Mechanics and Engineering*, 193(23–26):2383–2402, 2004.
- [155] P. Lukáš and P. Knobloch. Adaptive techniques in SOLD methods. *Applied Mathematics and Computation*, 319:24–30, 2018.
- [156] K. Eriksson and C. Johnson. Adaptive streamline diffusion finite element methods for stationary convection-diffusion problems. *Mathematics of Computation*, 60(201):167–188, 1993.
- [157] C. Johnson, A. H. Schatz, and L. B. Wahlbin. Crosswind smear and pointwise errors in streamline diffusion finite element methods. *Mathematics of Computation*, 49(179):25–38, 1987.

- [158] R. Codina. A discontinuity-capturing crosswind-dissipation for the finite element solution of the convection-diffusion equation. *Computer Methods in Applied Mechanics and Engineering*, 49(179):325–342, 1993.
- [159] T. Knopp, G. Lube, and G. Rapin. Stabilized finite element methods with shock capturing for advection-diffusion problems. *Computer Methods in Applied Mechanics and Engineering*, 191(27–28):2997–3013, 2002.
- [160] E. Burman and A. Ern. Nonlinear diffusion and discrete maximum principle for stabilized Galerkin approximations of the convection–diffusion–reaction equation. *Computer Methods in Applied Mechanics and Engineering*, 191(35):3833–3855, 2002.
- [161] V. John and P. Knobloch. A computational comparison of methods diminishing spurious oscillations in finite element solutions of convection-diffusion equations. *Programs and Algorithms of Numerical Mathematics*, pages 122–136, 2006.
- [162] V. John and P. Knobloch. On discontinuity-capturing methods for convection–diffusion equations, in: Bermu´dez de Castro, A. and Go´mez, D. and Quintela, P. and Salgado, P. (Eds.). *Numerical Mathematics and Advanced Applications, Proceedings of ENUMATH*, pages 336–344, 2006.
- [163] V. John and P. Knobloch. On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part II–Analysis for P1 and Q1 finite elements. *Computer Methods in Applied Mechanics and Engineering*, 197(21–24):1997–2014, 2008.
- [164] E. Burman and A. Ern. Stabilized Galerkin approximation of convection-diffusion-reaction equations: discrete maximum principle and convergence. *Mathematics of Computation*, 74(252):1637–1652, 2005.
- [165] R. Löhner, K. Morgan, J. Peraire, and M. Vahdati. Finite element flux-corrected transport (FEM–FCT) for the Euler and Navier–Stokes equations. *International Journal for Numerical Methods in Fluids*, 7(10):1093–1109, 1987.
- [166] P. Arminjon and A. Dervieux. Construction of TVD-like artificial viscosities on two-dimensional arbitrary FEM grids. *Journal of Computational Physics*, 106(1):176–198, 1993.
- [167] J. P. Boris and D. L. Book. Flux-corrected transport. I. SHASTA, a fluid transport algorithm that works. *Journal of Computational Physics*, 11(1):38–69, 1973.
- [168] S. T. Zalesak. Fully multidimensional flux-corrected transport algorithms for fluids. *Journal of Computational Physics*, 31(3):335–362, 1979.
- [169] D. Kuzmin and S. Turek. High-resolution FEM-TVD schemes based on a fully multidimensional flux limiter. *Journal of Computational Physics*, 198(1):131–158, 2004.

- [170] D. Kuzmin and M. Möller. Algebraic flux correction I. Scalar conservation laws. Flux-corrected Transport. *Principles, Algorithms, and Applications* (D. Kuzmin, R. Löhner and S. Turek eds), pages 155–206, 2005.
- [171] D. Kuzmin. On the design of general-purpose flux limiters for finite element schemes. I. Scalar convection. *Journal of Computational Physics*, 219(2):513–531, 2006.
- [172] D. Kuzmin. Algebraic flux correction for finite element discretizations of coupled systems. *Computational Methods for Coupled Problems in Science and Engineering II, CIMNE*, pages 653–656, 2007.
- [173] D. Kuzmin. On the design of algebraic flux correction schemes for quadratic finite elements. *Journal of Computational and Applied Mathematics*, 218(1):79–87, 2008.
- [174] D. Kuzmin. Explicit and implicit FEM-FCT algorithms with flux linearization. *Journal of Computational Physics*, 228(7):2517–2534, 2009.
- [175] D. Kuzmin. Linearity-preserving flux correction and convergence acceleration for constrained Galerkin schemes. *Journal of Computational and Applied Mathematics*, 236(9):2317–2337, 2012.
- [176] Richard S. Varga. *Matrix Iterative Analysis*, volume 27 of *expanded ed.*, in: *Springer Series in Computational Mathematics expanded ed.*, in: *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2000.
- [177] V. John and P. Knobloch. On algebraically stabilized schemes for convection–diffusion–reaction problems. *Numerische Mathematik*, 152(3):553–585, 2022.
- [178] G. R. Barrenechea, V. John, and P. Knobloch. Analysis of algebraic flux correction schemes. *SIAM Journal on Numerical Analysis*, 54(4):2427–2451, 2016.
- [179] G. R. Barrenechea, V. John, and P. Knobloch. An algebraic flux correction scheme satisfying the discrete maximum principle and linearity preservation on general meshes. *Mathematical Models and Methods in Applied Sciences*, 27(03):525–548, 2017.
- [180] P. Knobloch. An algebraically stabilized method for convection–diffusion–reaction problems with optimal experimental convergence rates on general meshes. *Numerical Algorithms*, 94(2):547–580, 2023.
- [181] G. R. Barrenechea, E. Burman, and F. Karakatsani. Edge-based nonlinear diffusion for finite element approximations of convection–diffusion equations and its relation to algebraic flux-correction schemes. *Numerische Mathematik*, 135:521–545, 2017.
- [182] A. Jha and V. John. A study of solvers for nonlinear AFC discretizations of convection-diffusion equations. *Computers and Mathematics with Applications*, 78(9):3117–3138, 2019.



- [183] G. R. Barrenechea, V. John, P. Knobloch, and R. Rankin. A unified analysis of algebraic flux correction schemes for convection-diffusion equations. *SeMA Journal*, 75:655–685, 2018.
- [184] P. Knobloch. On the discrete maximum principle for algebraic flux correction schemes with limiters of upwind type. In *Boundary and Interior Layers, Computational and Asymptotic Methods BAIL 2016*, pages 129–139, 2017.
- [185] P. Knobloch. A linearity preserving algebraic flux correction scheme of upwind type satisfying the discrete maximum principle on arbitrary meshes. *Numerical Mathematics and Advanced Applications ENUMATH 2017*, 126:909–918, 2019.
- [186] P. Knobloch. A new algebraically stabilized method for convection–diffusion–reaction equations. *Numerical Mathematics and Advanced Applications ENUMATH 2019*, 139:605–613, 2021.
- [187] D. Kuzmin. Monolithic convex limiting for continuous finite element discretizations of hyperbolic conservation laws. *Computer Methods in Applied Mechanics and Engineering*, 361:112804, 2020.
- [188] P. Knobloch, D. Kuzmin, and A. Jha. Well-balanced convex limiting for finite element discretizations of steady convection-diffusion-reaction equations. *Preprint*, page arXiv:2401.03964., 2024.
- [189] A. Jha, O. Pártl, N. Ahmed, and D. Kuzmin. An assessment of solvers for algebraically stabilized discretizations of convection–diffusion–reaction equations. *Journal of Numerical Mathematics*, 31(2):79–103, 2023.
- [190] A. Jha and V. John. On basic iteration schemes for nonlinear AFC discretizations. *Proceedings of BAIL 2018, in: WIAS Preprint 2533, Weierstrass Institute for Applied Analysis and Stochastics*, pages 113–128, 2020.
- [191] G. R. Barrenechea, V. John, and P. Knobloch. Some analytical results for an algebraic flux correction scheme for a steady convection–diffusion equation in one dimension. *IMA Journal of Numerical Analysis*, 35(4):1729–1756, 2015.
- [192] V. John, P. Knobloch, and J. Novo. Finite elements for scalar convection-dominated equations and incompressible flow problems: a never ending story? *Computing and Visualization in Science*, 19:47–63, 2018.
- [193] A. Jha. A residual based a posteriori error estimators for AFC schemes for convection-diffusion equations. *Computers and Mathematics with Applications*, 97:86–99, 2021.
- [194] A. Jha, V. John, and P. Knobloch. Adaptive grids in the context of algebraic stabilizations for convection-diffusion-reaction equations. *SIAM Journal on Scientific Computing*, 45(4):B564–B589, 2023.

- [195] G. R. Barrenechea, V. John, and P. Knobloch. Finite element methods respecting the discrete maximum principle for convection-diffusion equations. *SIAM Review*, 66(1):3–88, 2024.
- [196] V. John, P. Knobloch, and O. Pártl. A numerical assessment of finite element discretizations for convection-diffusion-reaction equations satisfying discrete maximum principles. *Computational Methods in Applied Mathematics*, 23(4):969–988, 2023.
- [197] A.C. Galeaao and E.G. Do Carmo. A consistent approximate upwind Petrov–Galerkin method for convection-dominated problems. *Computer Methods in Applied Mechanics and Engineering*, 88:83–95, 1988.
- [198] A.C. Galeaao and E.G. Do Carmo. Feedback Petrov–Galerkin methods for convection-dominated problems. *Computer Methods in Applied Mechanics and Engineering*, 88:1–16, 1991.
- [199] J. Donea. A Taylor–Galerkin method for convective transport problems. *International Journal for Numerical Methods in Engineering*, 20(1):101–119, 1984.
- [200] J. Peraire. *A finite element method for convection dominated flows*. Ph.D. Thesis. University of Wales, Swansea, 1986.
- [201] J. L. Guermond, R. Pasquetti, and B. Popov. Entropy viscosity method for nonlinear conservation laws. *Journal of Computational Physics*, 230(11):4248–4267, 2011.
- [202] T. J. Hughes. Multiscale phenomena: Green’s functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods. *Computer Methods in Applied Mechanics and Engineering*, 127(1–4):387–401, 1995.
- [203] R. Codina. Stabilization of incompressibility and convection through orthogonal sub-scales in finite element methods. *Computer Methods in Applied Mechanics and Engineering*, 190(13–14):1579–1599, 2000.
- [204] R. Codina and J. Blasco. Analysis of a stabilized finite element approximation of the transient convection-diffusion-reaction equation using orthogonal subscales. *Computing and Visualization in Science*, 4:1579–1599, 2002.
- [205] M. Stynes. Steady-state convection–diffusion problems. *Acta Numerica*, Cambridge University Press, pages 445–508, 2005.
- [206] M. Augustin, A. Caiazzo, A. Fiebach, J. Fuhrmann, V. John, A. Linke, and R. Umla. An assessment of discretizations for convection-dominated convection-diffusion equations. *Computer Methods in Applied Mechanics and Engineering*, 200(47–48):3395–3409, 2011.
- [207] V. John and E. Schmeyer. Finite element methods for time-dependent convection–diffusion–reaction equations with small diffusion. *Computer Methods in Applied Mechanics and Engineering*, 198(3–4):475–494, 2008.

- [208] P. B. Bochev, M. D. Gunzburger, and J. N. Shadid. Stability of the SUPG finite element method for transient advection-diffusion problems. *Computer Methods in Applied Mechanics and Engineering*, 193(23–26):2301–2323, 2004.
- [209] F. Shakib and T. J. Hughes. A new finite element formulation for computational fluid dynamics: IX. Fourier analysis of space-time Galerkin/least-squares algorithms. *Computer Methods in Applied Mechanics and Engineering*, 87(1):35–58, 1991.
- [210] G. Lube and G. Rapin. Residual-based stabilized higher-order FEM for advection-dominated problems. *Computer Methods in Applied Mechanics and Engineering*, 195(33–36):4124–4138, 2006.
- [211] G. Hauke. A simple subgrid scale stabilized method for the advection–diffusion–reaction equation. *Computer Methods in Applied Mechanics and Engineering*, 191(27–28):2925–2947, 2002.
- [212] F. Ilinca and J. F. Héту. Galerkin gradient least-squares formulations for transient conduction heat transfer. *Computer Methods in Applied Mechanics and Engineering*, 191(27–28):3073–3097, 2002.
- [213] J. De Frutos, B. García-Archilla, and J. Novo. Stabilization of Galerkin finite element approximations to transient convection-diffusion problems. *SIAM Journal on Numerical Analysis*, 48(3):953–979, 2010.
- [214] J. De Frutos, B. García-Archilla, and J. Novo. Accurate approximations to time-dependent nonlinear convection–diffusion problems. *IMA Journal of Numerical Analysis*, 30(4):1137–1158, 2010.
- [215] E. Burman. Consistent SUPG-method for transient transport problems: Stability and convergence. *Computer Methods in Applied Mechanics and Engineering*, 199(17–20):1114–1123, 2010.
- [216] V. John and J. Novo. Error analysis of the SUPG finite element discretization of evolutionary convection-diffusion-reaction equations. *SIAM Journal on Numerical Analysis*, 49(3):1149–1176, 2011.
- [217] M. C. Hsu, Y. Bazilevs, V. M. Calo, T. E. Tezduyar, and T. Hughes. Improving stability of stabilized and multiscale formulations in flow simulations at small time steps. *Computer Methods in Applied Mechanics and Engineering*, 199(13–16):828–840, 2010.
- [218] J. Donea, L. Quartapelle, and V. Selmin. An analysis of time discretization in the finite element solution of hyperbolic problems. *Journal of Computational Physics*, 70(2):463–499, 1987.
- [219] A. Huerta and J. Donea. Time-accurate solution of stabilized convection–diffusion–reaction equations: I—time and space discretization. *Communications in Numerical Methods in Engineering*, 18(8):565–573, 2002.

- [220] A. Huerta, B. Roig, and J. Donea. Time-accurate solution of stabilized convection–diffusion–reaction equations: II—accuracy analysis and examples. *Communications in Numerical Methods in Engineering*, 18(8):575–584, 2002.
- [221] J. Donea, B. Roig, and A. Huerta. High-order accurate time-stepping schemes for convection-diffusion problems. *Computer Methods in Applied Mechanics and Engineering*, 182(3–4):249–275, 2000.
- [222] P. M. Gresho and R. L. Sani. *Incompressible flow and the finite element method. Volume 1: Advection-diffusion and isothermal laminar flow*. Osi.Gov, 1998.
- [223] G. Lube and D. Weiss. Stabilized finite element methods for singularly perturbed parabolic problems. *Applied Numerical Mathematics*, 17(4):431–459, 1995.
- [224] P. Knobloch. Error estimates for a nonlinear local projection stabilization of transient convection-diffusion-reaction equations. *Discrete and Continuous Dynamical Systems-S*, 8(5):901, 2015.
- [225] S. Srivastava and S. Ganesan. Local projection stabilization with discontinuous Galerkin method in time applied to convection dominated problems in time-dependent domains. *BIT Numerical Mathematics*, 60:481–507, 2020.
- [226] N. Ahmed, G. Matthies, L. Tobiska, and H. Xie. Discontinuous Galerkin time stepping with local projection stabilization for transient convection–diffusion–reaction problems. *Computer Methods in Applied Mechanics and Engineering*, 200(21–22):1747–1756, 2011.
- [227] D. Kuzmin, S. Basting, and J. N. Shadid. Linearity-preserving monotone local projection stabilization schemes for continuous finite elements. *Computer Methods in Applied Mechanics and Engineering*, 322:23–41, 2017.
- [228] Y. Bazilevs, V. M. Calo, T. E. Tezduyar, and T. J. Hughes. YZB discontinuity capturing for advection-dominated processes with application to arterial drug delivery. *International Journal for Numerical Methods in Fluids*, 54(6–8):593–608, 2007.
- [229] E. G. D. Do Carmo and G. B. Alvarez. A new stabilized finite element formulation for scalar convection–diffusion problems: the streamline and approximate upwind/Petrov–Galerkin method. *Computer Methods in Applied Mechanics and Engineering*, 192(31–32):3379–3396, 2003.
- [230] V. John and J. Novo. On (essentially) non-oscillatory discretizations of evolutionary convection-diffusion equations. *Journal of Computational Physics*, 231(4):1570–1586, 2012.
- [231] V. John and E. Schmeier. On finite element methods for 3d time-dependent convection-diffusion-reaction equations with small diffusion. *Proceedings of the International Conference on Boundary and Interior Layers-Computational and Asymptotic Methods*, pages 173–181, 2009.

- [232] D. Kuzmin, M. Möller, and S. Turek. High-resolution FEM–FCT schemes for multidimensional conservation laws. *Computer Methods in Applied Mechanics and Engineering*, 193(45–47):4915–4946, 2004.
- [233] V. John, P. Knobloch, and P. Korschmeier. On the solvability of the nonlinear problems in an algebraically stabilized finite element method for evolutionary transport-dominated equations. *Mathematics of Computation*, 90(328):595–611, 2021.
- [234] V. John and P. Knobloch. Existence of solutions of a finite element flux-corrected-transport scheme. *Applied Mathematics Letters*, 115:106932, 2021.
- [235] A. Jha and N. Ahmed. Analysis of flux corrected transport schemes for evolutionary convection-diffusion-reaction equations. *Preprint*, page arXiv:2103.04776., 2021.
- [236] C.A. Henao, A. L. Coutinho, and L. P. Franca. A stabilized method for transient transport equations. *Computational Mechanics*, 46:199–204, 2010.
- [237] J. De Frutos and J. Novo. Bubble stabilization of linear finite element methods for nonlinear evolutionary convection–diffusion equations. *Computer Methods in Applied Mechanics and Engineering*, 197(45–48):3988–3999, 2008.
- [238] T. J. Hughes. Multiscale phenomena: Green’s functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods. *Computer Methods in Applied Mechanics and Engineering*, 127(1–4):387–401, 1995.
- [239] J. Douglas, Jr and T. F. Russell. Numerical methods for convection-dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures. *SIAM Journal on Numerical Analysis*, 19(5):871–885, 1982.
- [240] O. C. Zienkiewicz and R. Codina. A general algorithm for compressible and incompressible flow—Part I. the split, characteristic-based scheme. *International Journal for Numerical Methods in Fluids*, 28(8–9):869–885, 1995.
- [241] R. Codina. Comparison of some finite element methods for solving the diffusion-convection-reaction equation. *Computer Methods in Applied Mechanics and Engineering*, 156(1–4):185–210, 1998.
- [242] A. J. Perumpanani, J. A. Sherratt, J. Norbury, and H. M. Byrne. A two parameter family of travelling waves with a singular barrier arising from the modelling of extracellular matrix mediated cellular invasion. *Physica D: Nonlinear Phenomena*, 126(3–4):145–159, 1999.
- [243] A. Alsenafi and A. B. Barbaro. A convection–diffusion model for gang territoriality. *Physica A: Statistical Mechanics and its Applications*, 510:765–786, 2018.

- [244] R. E. Mickens. *Nonstandard finite difference models of differential equations*. World Scientific, 1994.
- [245] R. E. Mickens. *Applications of nonstandard finite difference schemes*. World Scientific, 2000.
- [246] R. E. Mickens. *Nonstandard finite difference schemes: methodology and applications*. World Scientific, 2020.
- [247] M. Mehdizadeh-Khalsaraei, Sh. Heydari, L. Davari Algoo, and R. Shokri Jahandizi. Positivity preserving and elementary stable nonstandard finite difference scheme for the predator-prey model. *International Journal of Mathematics and Computation*, 28(1):48–57, 2017.
- [248] M. Mehdizadeh-Khalsaraei, L. Davari Algoo, and Sh. Heydari. A family of explicit nonstandard finite difference schemes with positivity property for MSEIR models. *International Journal of Nonlinear Science*, 23(2):116–122, 2017.
- [249] M. Mehdizadeh-Khalsaraei, Sh. Heydari, and L. D. Algoo. Positivity preserving nonstandard finite difference schemes applied to cancer growth model. *Journal of Cancer Treatment and Research*, 4(4):27–33, 2016.
- [250] H. V. Kojouharov and B. M. Chen. Nonstandard methods for the convective transport equation with nonlinear reactions. *Numerical Methods for Partial Differential Equations: An International Journal*, 14(4):467–485, 1998.
- [251] H. V. Kojouharov and B. M. Chen. Nonstandard methods for the convective-dispersive transport equation with nonlinear reactions. *Numerical Methods for Partial Differential Equations: An International Journal*, 15(6):617–624, 1999.
- [252] J. B. Cole. High accuracy solution of Maxwell’s equations using nonstandard finite differences. *Computers in Physics*, 11(3):287–292, 1997.

# List of publications

## Journals

- Fuest, M., Heydari, Sh.: A cross-diffusion system modeling rivaling gangs: global existence of bounded solutions and FCT stabilization for numerical simulation, *Mathematical Models and Methods in Applied Sciences(M3AS)*, DOI: 10.1142/S0218202524500349, 2024.
- Heydari, Sh., Knobloch, P., Wick, T.: Flux-corrected transport stabilization of an evolutionary cross-diffusion cancer invasion model, *Journal of Computational Physics*, 499, Art. No. 112711, 2024.
- Fuest, M., Heydari, Sh., Knobloch, P., Lankeit, J., and Wick, Th.: Global existence of classical solutions and numerical simulations of a cancer invasion model, *Mathematical Modeling and Numerical Analysis (ESAIM: M2AN)*, 57(4):1893–1919, 2023.
- Mehdizadeh-Khalsaraei, M., Shokri, A., Ramos, H., Heydari, Sh.: A positive and elementary stable nonstandard explicit scheme for a mathematical model of the influenza disease, *Journal of Mathematics and Computers in Simulation*, 182,397-410, 2021.

## Conference proceedings

- Heydari, Sh., Knobloch, P.: Solvability and numerical solution of a cross-diffusion cancer invasion model, (*To appear in the proceedings of the ENU-MATH conference 2023*), 2024.