NUMERICAL ANALYSIS OF DDFV SCHEMES FOR SEMICONDUCTORS ENERGY-TRANSPORT MODELS

Marianne Bessemoulin-Chatard, Giulia Lissoni, Hélène Mathis

NUMERICAL ANALYSIS OF DDFV SCHEMES FOR SEMICONDUCTORS ENERGY-TRANSPORT MODELS

MARIANNE BESSEMOULIN-CHATARD, GIULIA LISSONI, AND HÉLÈNE MATHIS

Abstract. This article addresses the construction and the numerical analysis of implicit Discrete Duality Finite Volume schemes for a semiconductors' energy-transport model. The considered energy-transport model is presented in its scaled version as well as in a symmetrized form which involves entropy variables. We propose implicit in time numerical schemes for both the original system and its symmetrized form. As in the continuous framework, the numerical analysis is based on the reformulation of the PDE system using the set of entropic variables. The equivalence of both schemes allows to establish a discrete entropy inequality and consequently *a priori* estimates. As a by-product, existence of solutions to the schemes is proved by means of a Leray-Schauder argument. Numerical evidences allow to compare the performances of both schemes on the test case of a 2D ballistic diode.

Key-words. Energy-transport model; Discrete Duality Finite Volumes; Discrete entropy method.

2020 MCS. 65M08, 65M12, 35K20

Contents

The authors were partially funded by the Centre Henri Lebesgue (ANR-11-LABX-0020-01) and ANR Project MoHyCon (ANR-17-CE40-0027-01).

In this article, we propose and analyze a finite volume discretization for the energy-transport model. This macroscopic model can be obtained as a diffusion approximation of the semiconductor Boltzmann equation [3, 14, 25], and describes the flow of electrons in semiconductor devices, taking into account diffusive, electrical and thermal effects. This two-moments model is more accurate than the widely used drift-diffusion model derived from the kinetic equation by employing only the zeroth-order moment [29, 27]. It allows to model hot electron effects, which is not the case with the drift-diffusion model.

These quasilinear parabolic equations have been analysed in [12] where the existence of weak solutions is proven. The proof relies on the symmetrization of the system by means of the entropic variables. From this reformulation the authors obtain successively an entropy inequality and *a priori* estimates from which they deduce existence of solutions using a fixed point theorem. The long time asymptotic towards thermal equilibrium is also proven and the question of uniqueness is addressed in [24].

Several numerical methods have been proposed to approximate the energy-transport system during the two last decades. Mixed finite element methods have been developed for the stationary problem, in 1D with exponential fitting [13], and then in 2D [23] and 3D [19]. Concerning finite difference discretizations, an high-order scheme for the 1D stationary problem has been presented by Fournié [18], as well as a Scharfetter-Gummel type scheme for the 1D and 2D transient problem in [30]. In the finite volume framework, Chainais and Peng [9] have introduced a one-dimensional finite volume scheme for the transient energy-transport model. Then a 2D discrete duality finite volume scheme has been presented in [8]. All these papers address the construction of several numerical schemes, as well as their numerical validation, but no numerical analysis is provided. Following the analysis of the continuous energy transport system conducted in [12], a finite volume scheme with two point-flux approximation is constructed and analysed in [6, 5]. Using the entropic change of variables and similar techniques coming from the continuous framework, the authors establish an entropy inequality, which allows to exhibit discrete *a priori* estimates and to state the existence of solutions to the scheme by means of a Leray-Schauder fixed point theorem.

In the present article, we generalize the idea of the finite volume scheme constructed in [6, 5] to the Discrete Duality Finite Volume ("DDFV" for short) framework,that will result different from the one presented in [8]. This class of finite volume methods allows to consider very general meshes, which is not the case with two-point flux approximation requiring admissible meshes, that is meshes satisfying a quite constraining orthogonality condition (see for example [16]). It appears that for physically realistic test cases, the difference in thickness between different parts of the simulation domain may require the use of local mesh refinements (see for example [23]). In this context, it may be difficult to construct an admissible mesh, which is one of the motivations for using a DDFV scheme. The introduction of this type of finite volume method dates back to [21, 15], for the study of the Laplace equation on a large class of 2D meshes including nonconformal and distorted meshes.

Contrary to classical two-point flux approximation finite volume schemes, DDFV schemes require unknowns on both vertices and centers of primal control volumes and allow to build two-dimensional discrete gradient and divergence operators being in duality in a discrete sense. This kind of construction has two main advantages: as already mention, it first allows to consider general meshes, and also to reconstruct and mimic at the discrete level the dual properties of the continuous differential operators.

The article is organized as follows. In Section 1, we introduce the considered energy-transport model and detail rigorously its rescaling. Then the entropy change of variables is presented, leading to the symmetrization of the energy-transport model. Section 2 concerns the DDFV framework and the construction of the implicit DDFV schemes for both the scaled energy-transport model and its symmetrized form. We only recall the main ingredients of the DDFV framework which are essential for the numerical analysis of the schemes, detailed in Section 3. The key ingredient is to make use of the entropy change of variables to prove the equivalence of both schemes. It is then possible to exhibit a discrete entropy inequality, from which are deduced *a priori* estimates, as it is done in the continuous case [12]. Existence of a solution to the schemes is then proved by means of a Leray-Schauder theorem. Section 4 presents a comparison of the numerical performances of both implicit DDFV schemes on a 2D test case. Emphasis is given to the implementation of the overall method, in particular the Newton method which is employed to solve the nonlinear problem at each time iteration, the adaptated time stepping technique to ensure the convergence of the Newton method, as well as some DDFV features and the considered distorded meshes.

1. The model

1.1. **The energy-transport model and its scaling.** The energy-transport system consists in two continuity equations for the electron density ρ_1 and the internal energy density ρ_2, coupled with a Poisson equation for the electrical potential V. The electron and energy densities ρ_1 and ρ_2 are functions of the quantity $\mathbf{u} = (u_1, u_2)$ defined by

$$u_1 = \frac{\mu}{T}, \quad u_2 = -\frac{1}{T}, \tag{1}$$

where μ is the chemical potential and T the temperature.

The problem is set on an open bounded subset of $\mathbb{R}^2$, which corresponds to the geometry of the semiconductor device. The observation time range is $[0, t_{\max}]$, with $t_{\max} > 0$. The energy transport model reads in $[0, t_{\max}] \times \Omega$ as

$$\begin{cases} \partial_t \rho_1(\mathbf{u}) + \frac{1}{q}\mathrm{div}\,(J_1(\mathbf{u})) = 0, \\ \partial_t \rho_2(\mathbf{u}) + \mathrm{div}\,(J_2(\mathbf{u})) = \nabla V \cdot J_1(\mathbf{u}) + W(\mathbf{u}), \\ -\epsilon_s \Delta V = q\,(C(\mathbf{x}) - \rho_1(\mathbf{u}))\,, \end{cases} \tag{2}$$

where J_1 is the current density of electrons, J_2 is the current density of energy, $W(\mathbf{u})$ is the energy relaxation term, $\nabla V \cdot J_1$ is the Joule heating term and $C(\mathbf{x})$ is the given doping profile.

The current densities J_1 and J_2 are given by:

$$\begin{aligned} J_1(\mathbf{u}) &= -qL_{11}(\mathbf{u})\left(\nabla\left(\frac{q}{k_B}u_1\right)+\frac{q}{k_B}u_2\nabla V\right)-\frac{q}{k_B}L_{12}(\mathbf{u})\nabla u_2, \\ J_2(\mathbf{u}) &= -L_{21}(\mathbf{u})\left(\nabla\left(\frac{q}{k_B}u_1\right)+\frac{q}{k_B}u_2\nabla V\right)-\frac{1}{k_B}L_{22}(\mathbf{u})\nabla u_2, \end{aligned} \tag{3}$$

where the quantities $L_{ij}(\mathbf{u})$ form a matrix $L(\mathbf{u}) = (L_{ij}(\mathbf{u}))_{1\leq i,j\leq 2}$.

Some physical constants are present in the model (2): q denotes the elementary charge of electrons, ϵ_s the permittivity constant of the material and k_B the Boltzmann constant.

Let us now bring system (2) into a scaled and dimensionless form. We introduce μ_0 the mobility constant, τ_0 the energy relaxation time, T_0 the ambient temperature and U_T the thermal voltage at T_0 defined by $qU_T = k_BT_0$. Let also C_m be the maximal value of the doping profile and ℓ the diameter of the considered device. Following [9, 22], we introduce the scaled quantities, indicated by the $\tilde{\cdot}$ symbol:

$$\begin{aligned} x &\longrightarrow \ell\tilde{x}, & t &\longrightarrow \frac{\ell^2}{U_T\mu_0}\tilde{t}, \\ \rho_1 &\longrightarrow C_m\tilde{\rho}_1, & J_1 &\longrightarrow \frac{q\mu_0U_TC_m}{\ell}\tilde{J}_1, \\ \rho_2 &\longrightarrow qU_TC_m\tilde{\rho}_2, & J_2 &\longrightarrow \frac{q\mu_0U_T^2C_m}{\ell}\tilde{J}_2, \\ V &\longrightarrow U_T\tilde{V}, & W &\longrightarrow \frac{q\mu_0U_T^2C_m}{\ell^2}\tilde{W}, \\ C &\longrightarrow C_m\tilde{C}, & T &\longrightarrow T_0\tilde{T}. \end{aligned} \tag{4}$$

Droping the $\tilde{\cdot}$ notation for sake of readibility, one obtains the following scaled energy transport model in $[0, t_{\max}]\times\Omega$:

$$\begin{cases} \partial_t\rho_1(\mathbf{u}) + \operatorname{div}\,(J_1(\mathbf{u})) = 0, \\ \partial_t\rho_2(\mathbf{u}) + \operatorname{div}\,(J_2(\mathbf{u})) = \nabla V\cdot J_1(\mathbf{u}) + W(\mathbf{u}), \\ -\lambda^2\Delta V = C(\mathbf{x}) - \rho_1(\mathbf{u}), \end{cases} \tag{5}$$

where $\lambda^2 = \dfrac{\epsilon_sU_T}{qC_m\ell^2}$ is the scaled Debye length and the scaled current densities are given by:

$$\begin{aligned} J_1(\mathbf{u}) &= -L_{11}(\mathbf{u})(\nabla u_1 + u_2\nabla V) - L_{12}(\mathbf{u})\nabla u_2, \\ J_2(\mathbf{u}) &= -L_{21}(\mathbf{u})(\nabla u_1 + u_2\nabla V) - L_{22}(\mathbf{u})\nabla u_2. \end{aligned} \tag{6}$$

System (5) is supplemented with initial condition $\mathbf{u}(t=0) = \mathbf{u}_{init}$, and with mixed boundary conditions on the domain boundary $\partial\Omega$. In particular, we assume that the domain boundary is decomposed as $\partial\Omega = \Gamma_D\cup\Gamma_N$, with $\Gamma_D\cap\Gamma_N = \emptyset$ and the measure of Γ_D, denoted $m(\Gamma_D)$, is positive. The part Γ_D corresponds to the Ohmic contacts, where we impose Dirichlet boundary conditions:

$$u_1 = g_1, \quad u_2 = g_2, \quad V = h \quad \text{on } [0,t_{\max}]\times\Gamma_D. \tag{7}$$

The boundary conditions g_1, g_2 and h do not depend on time and are assumed to be traces of functions (still denoted g_1, g_2 and h) globally defined on Ω such that

$$g_1,\, g_2 \in H^1(\Omega), \quad h\in H^1(\Omega)\cap L^\infty(\Omega). \tag{8}$$

The part Γ_N corresponds to insulating boundary segments, where zero-flux boundary conditions are applied:

$$J_1 \cdot \mathbf{n} = J_2 \cdot \mathbf{n} = \nabla V \cdot \mathbf{n} = 0 \quad \text{on } [0, t_{\max}] \times \Gamma_N, \tag{9}$$

where $\mathbf{n}$ denotes the outward normal to the domain Ω.

Adpoting the framework presented in [12], some assumptions are required on system (5).

Assumptions 1.

i. The function $\boldsymbol{\rho} = (\rho_1, \rho_2)$ is strongly monotone in the sense that there exists a constant $C_0 > 0$ such that

$$(\boldsymbol{\rho}(\mathbf{u}) - \boldsymbol{\rho}(\mathbf{v})) \cdot (\mathbf{u} - \mathbf{v}) \geq C_0 |\mathbf{u} - \mathbf{v}|^2, \quad \forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^2.$$

We also assume that $\boldsymbol{\rho}$ derives from a potential, that is there exists $\chi \in \mathcal{C}^1(\mathbb{R}^2, \mathbb{R})$ strictly convex such that $\boldsymbol{\rho} = \nabla_{\mathbf{u}} \chi$.

ii. The matrix $L(\mathbf{u})$ is symmetric, uniformly positive definite.

iii. The boundary condition g_2 is negative and constant on Γ_D.

iv. The energy relaxation term W is such that for all $\mathbf{u} \in \mathbb{R}^2$ and $g_2 < 0$:

$$W(\mathbf{u})(u_2 - g_2) \leq 0.$$

Under these assumptions, Jüngel and co-authors have proved in [12, 24] existence and uniqueness of solutions to system (5). The cornerstone of these proofs is the derivation of a relative entropy inequality, relative to the boundary condition, obtained under the assumptions i., iii. and iv.. In practice, Assumptions 1 are not always satisfied in physical applications. Notably assumption ii. on the matrix $L(\mathbf{u})$ is quite restrictive. However existence results for positive semi-definite diffusion matrices are established in [17, 20] for the stationary model, and in [10, 11] for the transient system with data close to the thermal equilibrium. More recently, existence of solutions in simplified degenerate cases has been proved, in [26] for a model with a simplified temperature equation and in [31] for vanishing eletric field (avoiding the coupling with the Poisson equation).

1.2. **The symmetrized system.** The key point in the analysis of the energy-transport model (5) , used extensively in [12], is to deal with its symmetrized form. To do so, we introduce another set of variables, called *entropic variables* or *electrochemical potentials*, $\mathbf{w} = (w_1, w_2)$ defined as

$$w_1 = u_1 + u_2 V, \qquad w_2 = u_2. \tag{10}$$

The energy-transport system (5) is then equivalent to

$$\begin{cases} \partial_t b_1(\mathbf{w}, V) + \operatorname{div}\left(I_1(\mathbf{w}, V)\right) = 0, \\ \partial_t b_2(\mathbf{w}, V) + \operatorname{div}\left(I_2(\mathbf{w}, V)\right) = -\partial_t V b_1(\mathbf{w}, V) + \bar{W}(\mathbf{w}, V), \\ -\lambda^2 \Delta V = C(\mathbf{x}) - b_1(\mathbf{w}, V), \end{cases} \tag{11}$$

where

$$\begin{cases} b_1(\mathbf{w}, V) = \rho_1(\mathbf{u}), \\ b_2(\mathbf{w}, V) = \rho_2(\mathbf{u}) - V \rho_1(\mathbf{u}), \end{cases} \tag{12}$$

and $\bar{W}(\mathbf{w}, V) = W(\mathbf{u})$.

The symmetrized currents densities read

$$\begin{cases} I_1(\mathbf{w},V) = -D_{11}(\mathbf{w},V)\nabla w_1 - D_{12}(\mathbf{w},V)\nabla w_2, \\ I_2(\mathbf{w},V) = -D_{21}(\mathbf{w},V)\nabla w_1 - D_{22}(\mathbf{w},V)\nabla w_2, \end{cases} \tag{13}$$

where the diffusion coefficients $D_{ij}(\mathbf{w},V)$ form the symmetric matrix $D(\mathbf{w},V) = (D_{ij}(\mathbf{w},V))_{1\leq i,j\leq 2}$ defined by

$$\begin{aligned} D_{11}(\mathbf{w},V) &= L_{11}(\mathbf{u}), \\ D_{12}(\mathbf{w},V) &= D_{21}(\mathbf{w},V) = L_{12}(\mathbf{u}) - VL_{11}(\mathbf{u}), \\ D_{22}(\mathbf{w},V) &= L_{22}(\mathbf{u}) - 2VL_{12}(\mathbf{u}) + V^2L_{11}(\mathbf{u}). \end{aligned} \tag{14}$$

Another convenient expression of the diffusion matrix is

$$D(\mathbf{w},V) = {}^tPL(\mathbf{u})P \quad \text{with } P = \begin{pmatrix} 1 & -V \\ 0 & 1 \end{pmatrix}, \tag{15}$$

from which we deduce that the diffusion matrix $D(\mathbf{w},V)$ is symmetric and positive definite since $L(\mathbf{u})$ is, according to Assumption ii.. With an abuse of notations and up to a change of variables, we can write a relation between the fluxes I_i and J_i, $i = 1,2$ as

$$\begin{aligned} I_1 &= J_1, \\ I_2 &= J_2 - VJ_1. \end{aligned} \tag{16}$$

This symmetrized system (11) is supplemented with initial condition $\mathbf{w}(t = 0) = \mathbf{w}_{init}$ and with mixed boundary conditions which are deduced from the initial condition $\mathbf{u}_{init}$ and the boundary conditions (7)–(9) for the scaled energy-transport model (5) by means of the change of variables (10), leading to

$$\begin{cases} w_{1,init} = u_{1,init} + u_{2,init}V_{init}, \quad w_{2,init} = u_{2,init} \quad \text{on } \Omega, \\ w_1 = \bar{g}_1 := g_1 + g_2h, \quad w_2 = \bar{g}_2 := g_2, \quad V = h \quad \text{on } [0,t_{\max}]\times\Gamma_D, \\ I_1\cdot\mathbf{n} = I_2\cdot\mathbf{n} = \nabla V\cdot\mathbf{n} = 0 \quad \text{on } [0,t_{\max}]\times\Gamma_N. \end{cases} \tag{17}$$

Equipped with these two equivalent systems (5) and (11), an entropy estimate can be derived and allows to prove existence, uniqueness and regularity of solutions to the systems [12, 24]. Our aim is now to mimic this analysis to the discrete framework. It requires to construct numerical schemes for the energy-transport system (5) and for the symmetrized system (11) in such a way that they are equivalent. Such a construction have been done in [5] for two-point flux finite volume schemes. It allows to recover successively a discrete entropy inequality, *a priori* estimates on the discrete solutions and finally to prove the existence of solutions to the schemes by a fixed point theorem. The purpose of the following sections is do adapt such a construction to the DDFV framework, allowing to deal with general meshes as well as anisotropic configurations.

2. Numerical schemes

2.1. The DDFV framework.

The development of DDFV methods goes back to the study of the Laplace equation on general meshes [15, 21] and have since been largely improved to adapt to more general PDEs in 3D configurations [2, 1]. The method relies on the proper definition of discrete gradient and divergence operators in order to preserve in the discrete setting their duality property. To do so, the discretization requires to manipulate different meshes, namely the primal and dual

meshes for the unknowns at the centers and vertices of primal cells and the diamond mesh, on which the discrete gradients are defined.

The purpose of this section is to recall the main features of the DDFV framework presented in [2, 28], while focusing on the tools needed in the construction and the analysis of the DDFV schemes for the scaled energy-transport model (5) and the symmetrized model (11).

For a complete review of the method and details on the duality properties, the reader is refered to [28, 1].

2.1.1. *Meshes.* To define the DDFV scheme, we need to introduce three meshes: the primal and dual meshes, on which the unknowns $(\mathbf{u}, V)$ or $(\mathbf{w}, V)$ will be approximated, and the diamond mesh on which the discrete gradients will be defined.

The primal mesh, denoted $\overline{\mathfrak{M}}$ and represented on Figure 1-left, is composed of the interior mesh $\mathfrak{M}$, consisting of open disjoints polygons K, and the set $\partial\mathfrak{M}$ of boundary edges considered as degenerate primal cells. We associate to each cell $K \in \overline{\mathfrak{M}}$ a point x_K, corresponding to the center of gravity of the cell if $K \in \mathfrak{M}$ or the midpoint of the boundary edge if $K \in \partial\mathfrak{M}$.

From this primal mesh, let us now build the associated dual mesh, represented in Figure 1-middle. To any vertex x_{K^*} of the primal mesh, we associate a dual cell K^* obtained by joining the centers of the primal control volumes sharing x_{K^*} as a vertex. We distinguish the interior dual mesh, for which $x_{K^*} \in \Omega$, denoted by $\mathfrak{M}^*$, and the boundary dual mesh, for which x_{K^*} belongs to $\partial\Omega$, denoted by $\partial\mathfrak{M}^*$. The dual mesh is finally defined as $\overline{\mathfrak{M}^*} = \mathfrak{M}^* \cup \partial\mathfrak{M}^*$.

For all neighboring primal cells K and L in $\overline{\mathfrak{M}}$, we suppose that $\bar{\partial K} \cap \bar{\partial L}$ is a segment, corresponding to an edge of the primal mesh $\mathfrak{M}$, denoted by $\sigma = K|L$. We denote by $\mathcal{E}$ the set of such edges. Similarly, we define the set $\mathcal{E}^*$ of the edges of the dual mesh $\overline{\mathfrak{M}^*}$. For all couple of edges $(\sigma, \sigma^*) \in \mathcal{E} \times \mathcal{E}^*$ such that $\sigma = K|L = (x_{K^*}, x_{L^*})$ and $\sigma^* = K^*|L^* = (x_K, x_L)$, we define the quadrilateral $\mathcal{D}_{\sigma,\sigma^*}$ (or simply $\mathcal{D}$ if there is no ambiguity) whose diagonals are σ and σ^*. This quadrilateral $\mathcal{D}_{\sigma,\sigma^*}$ is called a diamond cell and is represented in Figure 1-right. If $\sigma \subset \partial\Omega$ is a boundary edge, the diamond $\mathcal{D}_{\sigma,\sigma^*}$ degenerates into a triangle. The set of all the diamond cells defines the diamond mesh, denoted by $\mathfrak{D}$. We distinguish in $\mathfrak{D}$ the subset $\mathfrak{D}_{int}$ of all the interior diamonds and the subset $\mathfrak{D}_{ext}$ of all the boundary diamonds. We also split $\mathfrak{D}_{ext}$ into $\mathfrak{D}_{ext,N}$ the subset of diamonds $\mathcal{D}_{\sigma,\sigma^*}$ such that $\sigma \subset \Gamma_N$ and $\mathfrak{D}_{ext,D}$ the subset of diamonds $\mathcal{D}_{\sigma,\sigma^*}$ such that $\sigma \subset \Gamma_D$. Remark that we have a bijection between the diamonds $\mathcal{D} \in \mathfrak{D}$ and the edges $\mathcal{E}$ of the primal mesh. Similarly, there is a bijection between the diamonds $\mathcal{D} \in \mathfrak{D}$ and the edges $\mathcal{E}^*$ of the dual mesh. In summary, the DDFV mesh is made of $\mathcal{T} = (\overline{\mathfrak{M}}, \overline{\mathfrak{M}^*})$ and $\mathfrak{D}$ (refer to Figure 1 for an illustration).

For a diamond $\mathcal{D} = \mathcal{D}_{\sigma,\sigma^*}$ whose vertices are $(x_{K^*}, x_L, x_{L^*}, x_K)$, we define $x_\mathcal{D}$ its center, $m_\mathcal{D}$ its measure, m_σ the length of the primal edge σ, m_{σ^*} the length of the dual edge σ^*, and $\alpha_\mathcal{D}$ the angle between (x_K, x_L) and (x_{K^*}, x_{L^*}). We also need to define two direct basis $(\boldsymbol{\tau}_{K^*L^*}, \mathbf{n}_{\sigma K})$ and $(\mathbf{n}_{\sigma^* K^*}, \boldsymbol{\tau}_{KL})$, where $\mathbf{n}_{\sigma K}$ is the unit vector normal to σ oriented from K to L, $\mathbf{n}_{\sigma^* K^*}$ is the unit vector normal to σ^* oriented from K^* to L^*, $\boldsymbol{\tau}_{K^*L^*}$ is the unit tangent vector to σ oriented from x_{K^*} to x_{L^*}, and $\boldsymbol{\tau}_{KL}$ is the unit tangent vector to σ^* oriented from x_K to x_L. All these notations are summarized in Figure 2-left for an interior diamond cell and in Figure 2-right for a boundary cell.

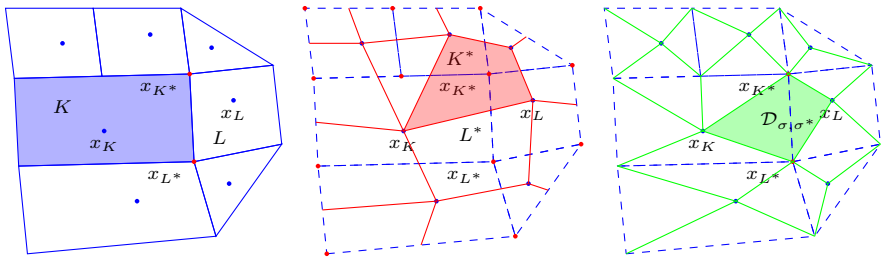

FIGURE 1. DDFV meshes on a nonconformal mesh: primal mesh $\mathfrak{M} \cup \partial\mathfrak{M}$ (blue), dual mesh $\mathfrak{M}^* \cup \partial\mathfrak{M}^*$ (red) and diamond mesh $\mathfrak{D}$ (green).

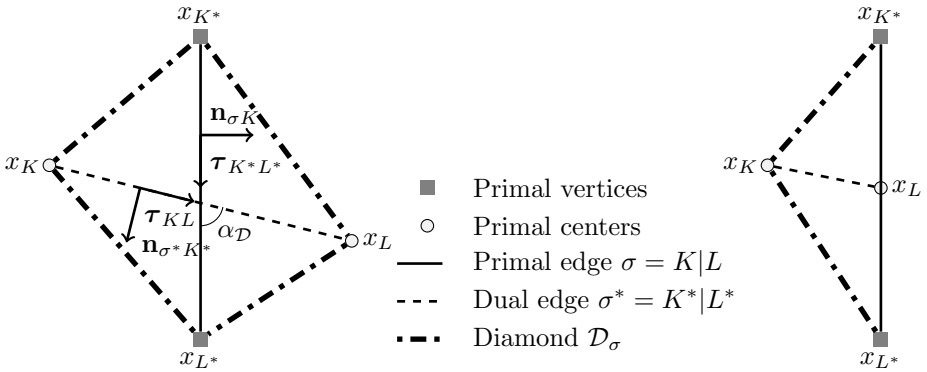

FIGURE 2. A diamond $\mathcal{D} = \mathcal{D}_{\sigma,\sigma^*}$, on the interior (left) and on the boundary (right).

Concluding the introduction of notations, for each primal cell $K \in \overline{\mathfrak{M}}$ (*resp.* dual cell $K^* \in \overline{\mathfrak{M}^*}$), we define m_K its measure (*resp.* m_{K^*}) and $\mathcal{D}_K$ the set of diamonds $\mathcal{D}_{\sigma,\sigma^*} \in \mathfrak{D}$ such that $m(\mathcal{D}_{\sigma,\sigma^*} \cap K) > 0$ (*resp.* $\mathcal{D}_{K^*}$).

Since we will have to use at some point a discrete Poincaré inequality (see for example [4]), we will be led to assume some regularity hypotheses on the mesh. First, the diamonds cannot be flat in the sense that there exists an angle $\alpha_{\mathcal{T}} \in \left]0, \frac{\pi}{2}\right]$ such that

$$|\sin(\alpha_{\mathcal{D}})| \geq \sin(\alpha_{\mathcal{T}}) \quad \forall \mathcal{D} \in \mathfrak{D}. \tag{18}$$

Moreover we assume that there exists $\zeta > 0$ such that:

$$\begin{aligned} \sum_{\mathcal{D}_{\sigma,\sigma^*} \in \mathcal{D}_K} m_\sigma m_{\sigma^*} &\leq \frac{m_K}{\zeta} \quad \forall K \in \mathfrak{M}, \\ \sum_{\mathcal{D}_{\sigma,\sigma^*} \in \mathcal{D}_{K^*}} m_\sigma m_{\sigma^*} &\leq \frac{m_{K^*}}{\zeta} \quad \forall K^* \in \overline{\mathfrak{M}^*}. \end{aligned} \tag{19}$$

2.1.2. *Unknowns and boundary conditions.* We now turn to the sets of discrete unknowns. We denote $\mathbb{R}^{\mathcal{T}}$ the linear space of scalar fields which are constant on

the primal and dual control volumes. The space $(\mathbb{R}^2)^{\mathfrak{D}}$ refers to the linear space of vector fields constant on the diamonds. We then have

$$v_{\mathcal{T}} \in \mathbb{R}^{\mathcal{T}} \Longleftrightarrow v_{\mathcal{T}} = \left((v_K)_{K\in\overline{\mathfrak{M}}}, (v_{K^*})_{K^*\in\overline{\mathfrak{M}^*}}\right),$$
$$\boldsymbol{\xi}_{\mathfrak{D}} \in (\mathbb{R}^2)^{\mathfrak{D}} \Longleftrightarrow \boldsymbol{\xi}_{\mathfrak{D}} = (\boldsymbol{\xi}_{\mathcal{D}})_{\mathcal{D}\in\mathfrak{D}}.$$

As the scaled energy-transport model is concerned, considering a solution $(\mathbf{u}, V)$ of (5), the approximations $(u_i)_{\mathcal{T}}$, $i = 1, 2$, of $\mathbf{u}$ and $V_{\mathcal{T}}$ of V are set on the primal and dual meshes. Similarly, for a solution $(\mathbf{w}, V)$ of the symmetrized system (11), the approximated values $(w_i)_{\mathcal{T}}$, $i = 1, 2$, of $\mathbf{w}$ and $V_{\mathcal{T}}$ of V are set on the primal and dual cells. On the other hand, the space $(\mathbb{R}^2)^{\mathfrak{D}}$ is dedicated to the approximate gradients of the considered quantities.

Since we have to handle mixed boundary conditions, we define two subsets of the boundary mesh, see Figure 3. As the Dirichlet domain is concerned, the primal and dual boundary meshes are

$$\partial\mathfrak{M}_D = \{K \in \partial\mathfrak{M} : x_K \in \Gamma_D\},$$
$$\partial\mathfrak{M}^*_D = \{K^* \in \partial\mathfrak{M}^* : x_{K^*} \in \Gamma_D\}.$$

For the Neumann boundary domain, we define two subsets

$$\partial\mathfrak{M}_N = \{K \in \partial\mathfrak{M} : x_K \in \Gamma_N\},$$
$$\partial\mathfrak{M}^*_N = \{K^* \in \partial\mathfrak{M}^* : x_{K^*} \in \Gamma_N \setminus \Gamma_D\}.$$

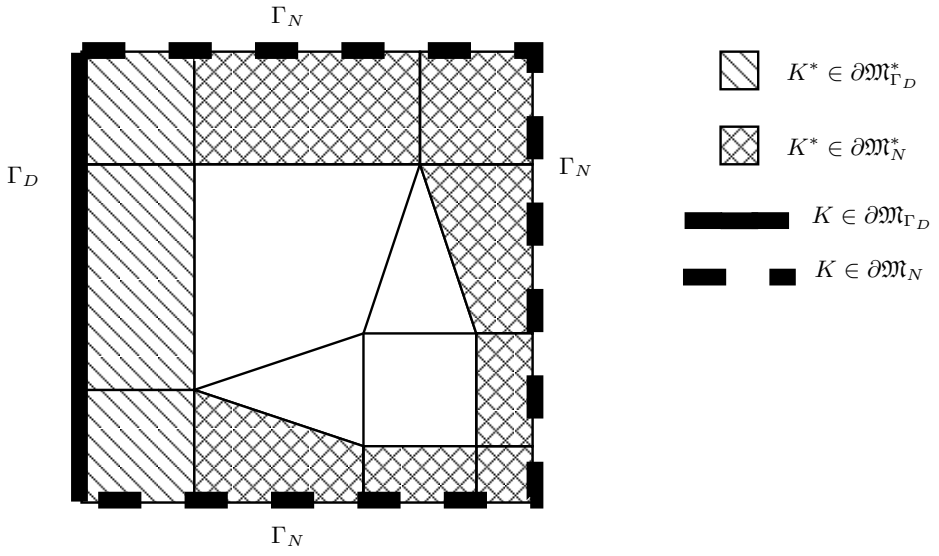

FIGURE 3. Domain with mixed boundary conditions.

2.1.3. *Discrete functional setting.* We now define scalar products on the approximation spaces $\mathbb{R}^{\mathcal{T}}$ and $(\mathbb{R}^2)^{\mathfrak{D}}$. On the primal and dual meshes $\mathcal{T}$, the scalar product is given by

$$[[u_{\mathcal{T}}, v_{\mathcal{T}}]]_{\mathcal{T}} = \frac{1}{2}\left(\sum_{K\in\mathfrak{M}} m_K u_K v_K + \sum_{K^*\in\overline{\mathfrak{M}^*}} m_{K^*} u_{K^*} v_{K^*}\right), \quad \forall u_{\mathcal{T}}, v_{\mathcal{T}} \in \mathbb{R}^{\mathcal{T}},$$

and the associated norm is

$$\|u_{\mathcal{T}}\|_{\mathcal{T}} = [[u_{\mathcal{T}}, u_{\mathcal{T}}]]_{\mathcal{T}}^{\frac{1}{2}}, \quad \forall u_{\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}.$$

On the diamond mesh $\mathfrak{D}$, the scalar product is

$$(\boldsymbol{\xi}_{\mathfrak{D}}, \boldsymbol{\Phi}_{\mathfrak{D}})_{\mathfrak{D}} = \sum_{\mathcal{D}_{\sigma,\sigma^*} \in \mathfrak{D}} m_{\mathcal{D}} \, \boldsymbol{\xi}_{\mathcal{D}} \cdot \boldsymbol{\Phi}_{\mathcal{D}}, \quad \forall \boldsymbol{\xi}_{\mathfrak{D}}, \, \boldsymbol{\Phi}_{\mathfrak{D}} \in (\mathbb{R}^2)^{\mathfrak{D}},$$

to which we associate the norm

$$\|\boldsymbol{\xi}_{\mathfrak{D}}\|_{\mathfrak{D}} = (\boldsymbol{\xi}_{\mathfrak{D}}, \boldsymbol{\xi}_{\mathfrak{D}})_{\mathfrak{D}}^{\frac{1}{2}}, \quad \forall \boldsymbol{\xi} \in (\mathbb{R}^2)^{\mathfrak{D}}.$$

The DDFV method allows to construct two-dimensional discrete gradient and divergence operators which are in duality in a discrete sense. This is precisely the duality property that gives the name to the method. We refer to [15] and [2] for a complete description of the duality framework. We first define the discrete gradient operator. It is a mapping from $\mathbb{R}^{\mathcal{T}}$ to $(\mathbb{R}^2)^{\mathfrak{D}}$ defined by $\nabla^{\mathfrak{D}} v_{\mathcal{T}} = \left(\nabla^{\mathcal{D}} v_{\mathcal{T}}\right)_{\mathcal{D} \in \mathfrak{D}}$ for all $v_{\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}$, where for all $\mathcal{D} \in \mathfrak{D}$,

$$\nabla^{\mathcal{D}} v_{\mathcal{T}} = \frac{1}{2m_{\mathcal{D}}} \left[m_\sigma (v_L - v_K) \mathbf{n}_{\sigma K} + m_{\sigma^*} (v_{L^*} - v_{K^*}) \mathbf{n}_{\sigma^* K^*} \right]. \tag{20}$$

We then define the discrete divergence operator. It is a mapping from $(\mathbb{R}^2)^{\mathfrak{D}}$ to $\mathbb{R}^{\mathcal{T}}$ given for all $\boldsymbol{\xi}_{\mathfrak{D}} \in (\mathbb{R}^2)^{\mathfrak{D}}$ by

$$\mathrm{div}^{\mathcal{T}} \boldsymbol{\xi}_{\mathfrak{D}} = \left(\mathrm{div}^{\mathfrak{M}} \boldsymbol{\xi}_{\mathfrak{D}}, \mathrm{div}^{\partial \mathfrak{M}} \boldsymbol{\xi}_{\mathfrak{D}}, \mathrm{div}^{\mathfrak{M}^*} \boldsymbol{\xi}_{\mathfrak{D}}, \mathrm{div}^{\partial \mathfrak{M}^*} \boldsymbol{\xi}_{\mathfrak{D}} \right), \tag{21}$$

where the divergence on the primal mesh is $\mathrm{div}^{\mathfrak{M}} \boldsymbol{\xi}_{\mathfrak{D}} = (\mathrm{div}^{K} \boldsymbol{\xi}_{\mathfrak{D}})_{K \in \mathfrak{M}}$ and $\mathrm{div}^{\partial \mathfrak{M}} \boldsymbol{\xi}_{\mathfrak{D}} = 0$, with

$$\mathrm{div}^{K} \boldsymbol{\xi}_{\mathfrak{D}} = \frac{1}{m_K} \sum_{\mathcal{D}_{\sigma,\sigma^*} \in \mathfrak{D}} m_\sigma \boldsymbol{\xi}_{\mathcal{D}} \cdot \mathbf{n}_{\sigma K}, \quad \forall K \in \mathfrak{M}.$$

Similarly the divergence on the dual mesh is $\mathrm{div}^{\mathfrak{M}^*} \boldsymbol{\xi}_{\mathfrak{D}} = (\mathrm{div}^{K^*} \boldsymbol{\xi}_{\mathfrak{D}})_{K^* \in \mathfrak{M}^*}$ and $\mathrm{div}^{\partial \mathfrak{M}^*} \boldsymbol{\xi}_{\mathfrak{D}} = (\mathrm{div}^{K^*} \boldsymbol{\xi}_{\mathfrak{D}})_{K^* \in \partial \mathfrak{M}^*}$ with an analogous definition for $\mathrm{div}^{K^*} \boldsymbol{\xi}_{\mathfrak{D}}$, $K^* \in \mathfrak{M}^*$, and for all $K^* \in \partial \mathfrak{M}^*$,

$$\mathrm{div}^{K^*} \boldsymbol{\xi}_{\mathfrak{D}} = \frac{1}{m_{K^*}} \left(\sum_{\mathcal{D}_{\sigma,\sigma^*} \in \mathfrak{D}_{K^*}} m_{\sigma^*} \boldsymbol{\xi}_{\mathcal{D}} \cdot \mathbf{n}_{\sigma^* K^*} + \sum_{\mathcal{D}_{\sigma,\sigma^*} \in \mathfrak{D}_{K^*} \cap \mathfrak{D}_{ext}} \frac{m_\sigma}{2} \boldsymbol{\xi}_{\mathcal{D}} \cdot \mathbf{n}_{\sigma K} \right).$$

This definition of the discrete divergence operator comes from a discrete version of the divergence theorem

$$\int_K \mathrm{div}\, (\boldsymbol{\xi}(x)) dx = \sum_{\sigma \in \partial K} \int_\sigma \boldsymbol{\xi}(s) \cdot \mathbf{n}_{\sigma K} ds, \quad \forall K \in \mathfrak{M},$$

and similarly for $K^* \in \mathfrak{M}^*$ or $\partial \mathfrak{M}^*$.

The discrete gradient and discrete divergence are linked by a discrete Green's formula. In this article, we only apply this formula to quantities with zero Dirichlet boundary conditions prescribed on Γ_D. Hence the Green's formula simplifies in this particular case. For all $\boldsymbol{\xi}_{\mathfrak{D}} \in (\mathbb{R}^2)^{\mathfrak{D}}$, for all $v_{\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}$ such that $v_K = 0$ for all $K \in \partial \mathfrak{M}_D$ and $v_{K^*} = 0$ for all $K^* \in \partial \mathfrak{M}_D^*$, it reads

$$[[\mathrm{div}^{\mathcal{T}} \boldsymbol{\xi}_{\mathfrak{D}}, v_{\mathcal{T}}]]_{\mathcal{T}} = -(\boldsymbol{\xi}_{\mathfrak{D}}, \nabla^{\mathfrak{D}} v_{\mathcal{T}})_{\mathfrak{D}}. \tag{22}$$

We end this section with the introduction of $\varphi^{\mathfrak{D}}$, a reconstruction operator on diamonds. It is a mapping from $\mathbb{R}^{\mathcal{T}}$ to $\mathbb{R}^{\mathfrak{D}}$, the linear space of scalar fields constant

on the diamonds, defined for all $v_{\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}$ by $\varphi^{\mathfrak{D}}(u_{\mathcal{T}}) = (\varphi^{\mathcal{D}}(u_{\mathcal{T}}))_{\mathcal{D}\in\mathfrak{D}}$. For a diamond cell $\mathcal{D} \in \mathfrak{D}_{int}$, whose vertices are x_K, x_L, x_{K^*}, x_{L^*}, see Figures 1-right and 2, we set

$$\varphi^{\mathcal{D}}(v_{\mathcal{T}}) = \frac{m_{\mathcal{D}\cap K} v_K + m_{\mathcal{D}\cap L} v_L + m_{\mathcal{D}\cap K^*} v_{K^*} + m_{\mathcal{D}\cap L^*} v_{L^*}}{2m_{\mathcal{D}}} \tag{23}$$

and for a diamond cell $\mathcal{D} \in \mathfrak{D}_{ext}$ on the boundary, whose vertices are x_K, x_{K^*}, x_{L^*}, we set

$$\varphi^{\mathcal{D}}(v_{\mathcal{T}}) = \frac{m_{\mathcal{D}\cap K} v_K + m_{\mathcal{D}\cap K^*} v_{K^*} + m_{\mathcal{D}\cap L^*} v_{L^*}}{2m_{\mathcal{D}}}. \tag{24}$$

To lighten the notations, we will simply denote by $v_{\mathfrak{D}}$ the reconstruction $\varphi^{\mathfrak{D}}(v_{\mathcal{T}})$.

2.2. **DDFV scheme for the energy-transport system.** We now introduce the numerical scheme for the energy-transport system (5)–(6). The discrete initial condition $\mathbf{u}^0_{\mathcal{T}}$ is defined as the mean value of $\mathbf{u}_{init}$ on the primal and dual meshes:

$$\mathbf{u}^0_K = \frac{1}{m_K}\int_K \mathbf{u}_{init} \quad \forall K \in \overline{\mathfrak{M}}, \qquad \mathbf{u}^0_{K^*} = \frac{1}{m_{K^*}}\int_{K^*} \mathbf{u}_{init} \quad \forall K^* \in \overline{\mathfrak{M}^*}. \tag{25}$$

Similarly, since the Dirichlet boundary conditions $\mathbf{g} = (g_1, g_2)$ and h are assumed to be defined on the whole domain Ω, we define their discretizations $\mathbf{g}_{\mathcal{T}}$, $h_{\mathcal{T}}$ by:

$$\mathbf{g}_K = \frac{1}{m_K}\int_K \mathbf{g} \quad \forall K \in \overline{\mathfrak{M}}, \qquad \mathbf{g}_{K^*} = \frac{1}{m_{K^*}}\int_{K^*} \mathbf{g} \quad \forall K^* \in \overline{\mathfrak{M}^*}, \tag{26}$$

$$h_K = \frac{1}{m_K}\int_K h \quad \forall K \in \overline{\mathfrak{M}}, \qquad h_{K^*} = \frac{1}{m_{K^*}}\int_{K^*} h \quad \forall K^* \in \overline{\mathfrak{M}^*}. \tag{27}$$

The approximate doping profile $C_{\mathcal{T}}$ is given on the primal and dual meshes by

$$C_K = \frac{1}{m_K}\int_K C \quad \forall K \in \overline{\mathfrak{M}}, \qquad C_{K^*} = \frac{1}{m_{K^*}}\int_{K^*} C \quad \forall K^* \in \overline{\mathfrak{M}^*}.$$

Let $N \in \mathbb{N}^*$. We note $\delta t = t_{\max}/N$ and $t_n = n\delta t$ for $n \in \{0, \dots N\}$. We choose to use an implicit Euler time discretization to avoid a constraining parabolic stability condition. In practice, this will lead to solve a nonlinear system of coupled equations at each time step. Some details about the practical implementation will be provided in Section 4.2, focusing on the Newton method which is employed to solve the nonlinear system at each time iteration.

The DDFV scheme is obtained by integrating equations (5) over all cells of $\mathfrak{M}$ and $\mathfrak{M}^* \cup \partial\mathfrak{M}^*_N$. Since it will simplify future computations, we proceed as in [6] and rewrite the Joule heating term as follows:

$$\nabla V \cdot J_1 = \text{div}\,(VJ_1) - V\,\text{div}\,J_1. \tag{28}$$

We also reformulate the term $u_2\nabla V$ as

$$u_2\nabla V = \text{div}\,(u_2 V) - V\nabla u_2. \tag{29}$$

This reformulation will allow to prove the equivalence between the schemes for the energy-transport system (5) and its symmetrized form (11). Indeed, contrary to the continuous framework, the discrete version of (29) is generally not true. In [6, 5], the authors constructed interfacial approximations of u_2 and V in such a way that a discrete version of (29) is obtained. Here, we choose to have a unique definition of the diamond reconstruction operator, given by (23)–(24). Doing so,

we do not need to verify a discrete counterpart of (29) but directly discretize the required formulation, that is the right hand side of (29).

Using equality (29), the current densities (6) thus become

$$J_i(\mathbf{u}) = -L_{i1}(\nabla u_1 + \nabla(u_2 V) - V\nabla u_2) - L_{i2}\nabla u_2, \quad i = 1, 2. \tag{30}$$

For all $n \geq 0$, knowing $\mathbf{u}_\mathcal{T}^n$ and $V_\mathcal{T}^n$, we are looking for $\mathbf{u}_\mathcal{T}^{n+1}$ and $V_\mathcal{T}^{n+1}$ solutions of the following nonlinear problem:

$$\frac{\rho_{1,K}^{n+1} - \rho_{1,K}^n}{\delta t} + \mathrm{div}^K(J_{1,\mathfrak{D}}^{n+1}) = 0, \qquad \forall K \in \mathfrak{M}, \tag{31}$$

$$\frac{\rho_{1,K^*}^{n+1} - \rho_{1,K^*}^n}{\delta t} + \mathrm{div}^{K^*}(J_{1,\mathfrak{D}}^{n+1}) = 0, \qquad \forall K^* \in \mathfrak{M}^* \cup \partial\mathfrak{M}_N^*, \tag{32}$$

$$\begin{aligned}\frac{\rho_{2,K}^{n+1} - \rho_{2,K}^n}{\delta t} + \mathrm{div}^K(J_{2,\mathfrak{D}}^{n+1}) &= \mathrm{div}^K(V_\mathfrak{D}^{n+1} J_{1,\mathfrak{D}}^{n+1}) \\ &\quad - V_K^{n+1}\mathrm{div}^K(J_{1,\mathfrak{D}}^{n+1}) + \mathrm{W}_K^{n+1}, \qquad \forall K \in \mathfrak{M},\end{aligned} \tag{33}$$

$$\begin{aligned}\frac{\rho_{2,K^*}^{n+1} - \rho_{2,K^*}^n}{\delta t} + \mathrm{div}^{K^*}(J_{2,\mathfrak{D}}^{n+1}) &= \mathrm{div}^{K^*}(V_\mathfrak{D}^{n+1} J_{1,\mathfrak{D}}^{n+1}) \\ &\quad - V_{K^*}^{n+1}\mathrm{div}^{K^*}(J_{1,\mathfrak{D}}^{n+1}) + \mathrm{W}_{K^*}^{n+1}, \qquad \forall K^* \in \mathfrak{M}^* \cup \partial\mathfrak{M}_N^*,\end{aligned} \tag{34}$$

$$-\lambda^2\mathrm{div}^K(\nabla^\mathfrak{D} V_\mathcal{T}^{n+1}) = C_K - \rho_{1,K}^{n+1}, \qquad \forall K \in \mathfrak{M}, \tag{35}$$

$$-\lambda^2\mathrm{div}^{K^*}(\nabla^\mathfrak{D} V_\mathcal{T}^{n+1}) = C_{K^*} - \rho_{1,K^*}^{n+1}, \qquad \forall K^* \in \mathfrak{M}^* \cup \partial\mathfrak{M}_N^*. \tag{36}$$

We adopt the following notations in (31)–(36): $\rho_{i,\mathcal{T}}^{n+1} = \rho_i(\mathbf{u}_\mathcal{T}^{n+1})$, $i = 1, 2$, and $W_\mathcal{T}^{n+1} = W(\mathbf{u}_\mathcal{T}^{n+1})$. The approximated current densities are defined by

$$J_{i,\mathfrak{D}}^{n+1} = -L_{i1}^{\mathfrak{D},n}\left(\nabla^\mathfrak{D} u_{1,\mathcal{T}}^{n+1} + \nabla^\mathfrak{D}(u_{2,\mathcal{T}}^{n+1} V_\mathcal{T}^{n+1}) - V_\mathfrak{D}^{n+1}\nabla^\mathfrak{D} u_{2,\mathcal{T}}^{n+1}\right) - L_{i2}^{\mathfrak{D},n}\nabla^\mathfrak{D} u_{2,\mathcal{T}}^{n+1}, \tag{37}$$

where the coefficients $L_{ij}^{\mathfrak{D},n}$ are defined as $L_{ij}^{\mathfrak{D},n} = L_{ij}(\mathbf{u}_\mathfrak{D}^n)$, $i, j = 1, 2$. Let us recall here that since the unknowns $(\mathbf{u}_\mathcal{T}, V_\mathcal{T})$ are associated to the mesh $\mathcal{T}$, we need to reconstruct their approximate values on the diamond mesh $\mathfrak{D}$. This is done by the reconstruction operator $\varphi^\mathfrak{D}$ defined by (23)–(24), and we set $\mathbf{u}_\mathfrak{D} := \varphi^\mathfrak{D}(\mathbf{u}_\mathcal{T})$, $V_\mathfrak{D} = \varphi^\mathfrak{D}(V_\mathcal{T})$ to lighten the notations.

The prescribed mixed boundary conditions are given by

$$\mathbf{u}_K^{n+1} = \mathbf{g}_K, \quad V_K^{n+1} = h_K, \qquad \forall K \in \partial\mathfrak{M}_D, \tag{38}$$

$$\mathbf{u}_{K^*}^{n+1} = \mathbf{g}_{K^*}, \quad V_{K^*}^{n+1} = h_{K^*}, \qquad \forall K^* \in \partial\mathfrak{M}_D^*, \tag{39}$$

and

$$J_{1,\mathcal{D}}^{n+1} \cdot \mathbf{n}_{\sigma K} = J_{2,\mathcal{D}}^{n+1} \cdot \mathbf{n}_{\sigma K} = \nabla^\mathcal{D} V_\mathcal{T}^{n+1} \cdot \mathbf{n}_{\sigma K} = 0, \quad \forall \mathcal{D} = \mathcal{D}_{\sigma,\sigma^*} \in \mathfrak{D}_{ext,N}, \tag{40}$$

where $\mathfrak{D}_{ext,N}$ denotes the subset of diamond cells $\mathcal{D}_{\sigma,\sigma^*}$ such that the edge σ belongs to the Neumann boundary domain Γ_N.

2.3. **DDFV scheme for the symmetrized system.** We now turn to the approximation of the symmetrized system (11). To do so, for all $n \geq 0$, we apply the change of variables (10) to define the discrete entropic variables:

$$w_{1,\mathcal{T}}^n := u_{1,\mathcal{T}}^n + u_{2,\mathcal{T}}^n V_\mathcal{T}^n, \qquad w_{2,\mathcal{T}}^n := u_{2,\mathcal{T}}^n, \tag{41}$$

which are associated to the definition (12) of the functions $b_i(\mathbf{w}, V)$, $i = 1, 2$:

$$b_{1,\mathcal{T}}^{n+1} = \rho_{1,\mathcal{T}}^{n+1}, \qquad b_{2,\mathcal{T}}^{n+1} = \rho_{2,\mathcal{T}}^{n+1} - V_{\mathcal{T}}^{n+1}\rho_{1,\mathcal{T}}^{n+1}. \tag{42}$$

The initial condition $\mathbf{w}_{\mathcal{T}}^0$ is deduced from $\mathbf{u}_{\mathcal{T}}^0$ defined by (25), by means of the change of variables (41), and $V_{\mathcal{T}}^0$ is obtained by solving (35)–(36). Similarly the Dirichlet boundary conditions $\overline{\mathbf{g}}_K$ for $K \in \partial\mathfrak{M}_D$ and $\overline{\mathbf{g}}_{K^*}$ for $K^* \in \partial\mathfrak{M}_D^*$ are deduced from the boundary conditions (26)–(27).

Plugging the change of variables (41)–(42) in the scheme (31)–(37) yields

$$\frac{b_{1,K}^{n+1} - b_{1,K}^{n}}{\delta t} + \mathrm{div}^K(I_{1,\mathfrak{D}}^{n+1}) = 0, \qquad \forall K \in \mathfrak{M}, \tag{43}$$

$$\frac{b_{1,K^*}^{n+1} - b_{1,K^*}^{n}}{\delta t} + \mathrm{div}^{K^*}(I_{1,\mathfrak{D}}^{n+1}) = 0, \qquad \forall K^* \in \mathfrak{M}^* \cup \partial\mathfrak{M}_N^*, \tag{44}$$

$$\frac{b_{2,K}^{n+1} - b_{2,K}^{n}}{\delta t} + \mathrm{div}^K(I_{2,\mathfrak{D}}^{n+1}) = -b_{1,K}^{n}\frac{V_K^{n+1} - V_K^n}{\delta t} + \bar{\mathrm{W}}_K^{n+1}, \qquad \forall K \in \mathfrak{M}, \tag{45}$$

$$\frac{b_{2,K^*}^{n+1} - b_{2,K^*}^{n}}{\delta t} + \mathrm{div}^{K^*}(I_{2,\mathfrak{D}}^{n+1}) = -b_{1,K^*}^{n}\frac{V_{K^*}^{n+1} - V_{K^*}^n}{\delta t} + \bar{\mathrm{W}}_{K^*}^{n+1}, \qquad \forall K^* \in \mathfrak{M}^* \cup \partial\mathfrak{M}_N^*, \tag{46}$$

$$-\lambda^2 \mathrm{div}^K(\nabla^{\mathfrak{D}} V_{\mathcal{T}}^{n+1}) = C_K - b_{1,K}^{n+1}, \qquad \forall K \in \mathfrak{M}, \tag{47}$$

$$-\lambda^2 \mathrm{div}^{K^*}(\nabla^{\mathfrak{D}} V_{\mathcal{T}}^{n+1}) = C_{K^*} - b_{1,K^*}^{n+1}, \qquad \forall K^* \in \mathfrak{M}^* \cup \partial\mathfrak{M}_N^*, \tag{48}$$

where the numerical fluxes $I_{i,\mathfrak{D}}$ are given by

$$I_{1,\mathfrak{D}}^{n+1} = J_{1,\mathfrak{D}}^{n+1}, \qquad I_{2,\mathfrak{D}}^{n+1} = J_{2,\mathfrak{D}}^{n+1} - V_{\mathfrak{D}}^{n+1} J_{1,\mathfrak{D}}^{n+1}. \tag{49}$$

The relaxation term $\bar{W}_{\mathcal{T}}$ is defined by $\bar{W}_{\mathcal{T}} = \bar{W}(\mathbf{w}_{\mathcal{T}}, V_{\mathcal{T}})$. After the change of variables, the obtained scheme is supplemented with mixed boundary conditions:

$$\mathbf{w}_K^{n+1} = \overline{\mathbf{g}}_K, \quad V_K^{n+1} = h_K, \qquad \forall K \in \partial\mathfrak{M}_D, \tag{50}$$

$$\mathbf{w}_{K^*}^{n+1} = \overline{\mathbf{g}}_{K^*}, \quad V_{K^*}^{n+1} = h_{K^*}, \qquad \forall K^* \in \partial\mathfrak{M}_D^*, \tag{51}$$

and

$$I_{1,\mathcal{D}}^{n+1} \cdot \mathbf{n}_{\sigma K} = I_{2,\mathcal{D}}^{n+1} \cdot \mathbf{n}_{\sigma K} = \nabla^{\mathcal{D}} V_{\mathcal{T}}^{n+1} \cdot \mathbf{n}_{\sigma K} = 0, \quad \forall \mathcal{D} = \mathcal{D}_{\sigma,\sigma^*} \in \mathfrak{D}_{ext,N}. \tag{52}$$

The point now is to determine if the obtained numerical fluxes $I_{i,\mathfrak{D}}$ are actually approximations of the symmetrized current densities I_i given in (13), *i.e.* if they can be written in the following form:

$$\begin{aligned} I_{1,\mathfrak{D}}^{n+1} &= -D_{11}^{\mathfrak{D}} \nabla^{\mathfrak{D}} w_{1,\mathcal{T}}^{n+1} - D_{12}^{\mathfrak{D}} \nabla^{\mathfrak{D}} w_{2,\mathcal{T}}^{n+1}, \\ I_{2,\mathfrak{D}}^{n+1} &= -D_{21}^{\mathfrak{D}} \nabla^{\mathfrak{D}} w_{1,\mathcal{T}}^{n+1} - D_{22}^{\mathfrak{D}} \nabla^{\mathfrak{D}} w_{2,\mathcal{T}}^{n+1}, \end{aligned} \tag{53}$$

up to a proper definition of the coefficients $D_{ij}^{\mathfrak{D}}$, i, $j = 1$, 2, which are approximations of the coefficients of the matrix D defined in (15). If so, the schemes (31)-(37) and (43)-(53) will be equivalent. It is actually the case, the result being established in the following section.

3. Numerical analysis and properties of the schemes

This section is devoted to the analysis of the two obtained numerical schemes. First of all, we establish the equivalence between the scheme (31)–(37) for the energy-transport model and the scheme (43)–(53) for the symmetrized system. This equivalence property allows to establish an entropy inequality from which we deduce discrete *a priori* estimates. Based on these results, we finally prove the existence of a solution to the implicit schemes, by means of a Leray-Schauder theorem.

3.1. **Equivalence of the schemes.**

Proposition 1. The scheme (31)–(37) written in variables $(\mathbf{u}_\mathcal{T}, V_\mathcal{T})$ is equivalent to the scheme (43)–(53) written in variables $(\mathbf{w}_\mathcal{T}, V_\mathcal{T})$, provided that the matrix $D^\mathfrak{D} = (D^\mathfrak{D}_{ij})_{1\leq i,j\leq 2}$ reads

$$D^\mathfrak{D} =^t P^{n+1}_\mathfrak{D} L^{\mathfrak{D},n} P^{n+1}_\mathfrak{D} \text{ with } P^{n+1}_\mathfrak{D} = \begin{pmatrix} 1 & -V^{n+1}_\mathfrak{D} \\ 0 & 1 \end{pmatrix}. \tag{54}$$

Proof. Starting from the scheme (31)–(37) for the scaled system, it remains to prove that fluxes (49) can be written under the form (53) with the definition (54) of the $D^\mathfrak{D}$. First observe that the formulation (54) is equivalent to

$$\begin{cases} D^\mathfrak{D}_{11} := L^{\mathfrak{D},n}_{11}, \\ D^\mathfrak{D}_{12} := L^{\mathfrak{D},n}_{12} - L^{\mathfrak{D},n}_{11} V^{n+1}_\mathfrak{D}, \\ D^\mathfrak{D}_{22} := L^{\mathfrak{D},n}_{22} - 2V^{n+1}_\mathfrak{D} L^{\mathfrak{D},n}_{12} + (V^{n+1}_\mathfrak{D})^2 L^{\mathfrak{D},n}_{11}. \end{cases} \tag{55}$$

Next, using the change of variables (41), we compute the discrete gradients of $\mathbf{w}^{n+1}_\mathcal{T}$:

$$\begin{cases} \nabla^\mathfrak{D} w^{n+1}_{1,\mathcal{T}} = \nabla^\mathfrak{D} u^{n+1}_{1,\mathcal{T}} + \nabla^\mathfrak{D}(u^{n+1}_{2,\mathcal{T}} V^{n+1}_\mathcal{T}), \\ \nabla^\mathfrak{D} w^{n+1}_{2,\mathcal{T}} = \nabla^\mathfrak{D} u^{n+1}_{2,\mathcal{T}}. \end{cases} \tag{56}$$

Plugging these gradients (56) into (49) together with the flux definition (37) of $J^{n+1}_{1,\mathfrak{D}}$ gives

$$\begin{aligned} I^{n+1}_{1,\mathfrak{D}} &= -L^{\mathfrak{D},n}_{11}(\nabla^\mathfrak{D} w^{n+1}_{1,\mathcal{T}} - V^{n+1}_\mathfrak{D} \nabla^\mathfrak{D} w^{n+1}_{2,\mathcal{T}}) - L^{\mathfrak{D},n}_{12} \nabla^\mathfrak{D} w^{n+1}_{2,\mathcal{T}} \\ &= -D^\mathfrak{D}_{11} \nabla^\mathfrak{D} w^{n+1}_{1,\mathcal{T}} - D^\mathfrak{D}_{12} \nabla^\mathfrak{D} w^{n+1}_{2,\mathcal{T}}, \end{aligned}$$

where the formulation (55) of the matrix $D^\mathfrak{D}$ has been used. Similarly, the definition (49) of the flux $I^{n+1}_{2,\mathfrak{D}}$ combined with the definition (37) of the fluxes $J^{n+1}_{i,\mathfrak{D}}$, $i = 1, 2$, leads to

$$\begin{aligned} I^{n+1}_{2,\mathfrak{D}} &= \nabla^\mathfrak{D} w^{n+1}_{1,\mathcal{T}} \left(-L^{\mathfrak{D},n}_{21} + V^{n+1}_\mathfrak{D} L^{\mathfrak{D},n}_{11} \right) \\ &\quad - \nabla^\mathfrak{D} w^{n+1}_{2,\mathcal{T}} \left(L^{\mathfrak{D},n}_{22} + 2V^{n+1}_\mathfrak{D} L^{\mathfrak{D},n}_{21} + \left(V^{n+1}_\mathfrak{D}\right)^2 L^{\mathfrak{D},n}_{11} \right) \\ &= -D^\mathfrak{D}_{21} \nabla^\mathfrak{D} w^{n+1}_{1,\mathcal{T}} - D^\mathfrak{D}_{22} \nabla^\mathfrak{D} w^{n+1}_{2,\mathcal{T}}. \end{aligned}$$

Hence the relations (53) are established with the definition (55) of the coefficients of the matrix $D^\mathfrak{D}$, which is the discrete counterpart of (15). Conversely applying the inverse change of variables in the scheme (43)-(49) for the symmetrized system with the definition (15) of the matrix $D^\mathfrak{D}$ gives the scheme (31)-(37). □

3.2. **Discrete entropy inequality and *a priori* estimates.** The scaled energy-transport model (5) is endowed with an entropy function which is proved in [12] to be decreasing in time as soon as the boundary conditions are at thermal equilibrium. The proof relies on Assumption 1-ii. of uniform positivity of the matrix L and Assumption 1-iv. of dissipation of the relaxation term W, the key ingredient being the passage from the scaled energy-transport model (5) to the symmetrized one (11) and *vice versa*.

We adapt hereafter the same kind of computations at the discrete level, taking advantage of the two equivalent DDFV formulations established in previous section.

For all $n \geq 0$, we define the discrete entropy function

$$\begin{aligned} S^n = &\frac{1}{2} \sum_{K \in \mathfrak{M}} m_K \left(\boldsymbol{\rho}_K^n \cdot (\mathbf{u}_K^n - \mathbf{g}_K) - (\chi(\mathbf{u}_K^n) - \chi(\mathbf{g}_K)) \right) \\ &+ \frac{1}{2} \sum_{K^* \in \mathfrak{M}^* \cup \partial\mathfrak{M}_N^*} m_{K^*} \left(\boldsymbol{\rho}_{K^*}^n \cdot (\mathbf{u}_{K^*}^n - \mathbf{g}_{K^*}) - (\chi(\mathbf{u}_{K^*}^n) - \chi(\mathbf{g}_{K^*})) \right) \\ &- \frac{\lambda^2}{2} g_2 ||\nabla^{\mathfrak{D}}(V_{\mathcal{T}}^n - h_{\mathcal{T}})||_{\mathfrak{D}}^2. \end{aligned} \tag{57}$$

It satisfies the following discrete dissipation property.

Proposition 2 (Discrete entropy dissipation)**.** Under Assumptions 1, if the Dirichlet boundary conditions are at thermal equilibrium

$$\nabla^{\mathfrak{D}} \overline{g}_{1,\mathcal{T}} = \nabla^{\mathfrak{D}} \overline{g}_{2,\mathcal{T}} = 0, \tag{58}$$

then the discrete entropy is a decreasing function in time in the sense that

$$\begin{aligned} \frac{S^{n+1} - S^n}{\delta t} \leq &- \sum_{\mathcal{D} \in \mathfrak{D}} {}^t\nabla^{\mathfrak{D}} \mathbf{w}_{\mathcal{T}}^{n+1} \, D^{\mathfrak{D}} \, \nabla^{\mathfrak{D}} \mathbf{w}_{\mathcal{T}}^{n+1} \\ &+ \left[\left[W_{\mathcal{T}}^{n+1} \, , \, u_{2,\mathcal{T}}^{n+1} - g_{2,\mathcal{T}} \right] \right]_{\mathcal{T}} \leq 0. \end{aligned} \tag{59}$$

Proof. By definition (57) of the discrete entropy functional, it holds

$$S^{n+1} - S^n = A + B,$$

with

$$\begin{aligned} A = &\frac{1}{2} \sum_{K \in \mathfrak{M}} m_K \left(\boldsymbol{\rho}_K^{n+1} \cdot (\mathbf{u}_K^{n+1} - \mathbf{g}_K) \right. \\ &\qquad \left. - \boldsymbol{\rho}_K^n \cdot (\mathbf{u}_K^n - \mathbf{g}_K) - \left(\chi(\mathbf{u}_K^{n+1}) - \chi(\mathbf{u}_K^n) \right) \right) \\ &+ \frac{1}{2} \sum_{K^* \in \mathfrak{M}^* \cup \partial\mathfrak{M}_N^*} m_{K^*} \left(\boldsymbol{\rho}_{K^*}^{n+1} \cdot (\mathbf{u}_{K^*}^{n+1} - \mathbf{g}_{K^*}) \right. \\ &\qquad \left. - \boldsymbol{\rho}_{K^*}^n \cdot (\mathbf{u}_{K^*}^n - \mathbf{g}_{K^*}) - \left(\chi(\mathbf{u}_{K^*}^{n+1}) - \chi(\mathbf{u}_{K^*}^n) \right) \right), \\ B = &- \frac{\lambda^2}{2} g_2 \left(||\nabla^{\mathfrak{D}}(V_{\mathcal{T}}^{n+1} - h_{\mathcal{T}})||_{\mathfrak{D}}^2 - ||\nabla^{\mathfrak{D}}(V_{\mathcal{T}}^n - h_{\mathcal{T}})||_{\mathfrak{D}}^2 \right). \end{aligned} \tag{60}$$

We first focus on term A. The convexity of χ, stated in Assumption 1-i., implies that $\chi(a) - \chi(b) \leq \nabla\chi(a) \cdot (a-b)$, for all $a, b \in \mathbb{R}^2$. It yields

$$\begin{aligned} A \leq & \frac{1}{2} \sum_{K \in \mathfrak{M}} m_K (\boldsymbol{\rho}_K^{n+1} - \boldsymbol{\rho}_K^n) \cdot (\mathbf{u}_K^{n+1} - \mathbf{g}_K) \\ & + \frac{1}{2} \sum_{K^* \in \mathfrak{M}^* \cup \partial\mathfrak{M}_N^*} m_{K^*} (\boldsymbol{\rho}_{K^*}^{n+1} - \boldsymbol{\rho}_{K^*}^n) \cdot (\mathbf{u}_{K^*}^{n+1} - \mathbf{g}_{K^*}). \end{aligned} \tag{61}$$

Developping the scalar products on the right-hand side gives

$$A \leq A_1 + A_2,$$

where

$$A_i = \left[\left[\rho_{i,\mathcal{T}}^{n+1} - \rho_{i,\mathcal{T}}^n, u_{i,T}^{n+1} - g_{i,\mathcal{T}} \right] \right]_{\mathcal{T}}, \quad i = 1, 2, \tag{62}$$

the notation of the scalar products on the mesh $\mathcal{T}$ being introduced in Section 2.1.3. We can now use the numerical scheme (31)-(36) to reformulate the term A_i, $i = 1, 2$. It holds

$$\begin{aligned} A_1 = & -\delta t \left[\left[\text{div}^{\mathcal{T}} J_{1,\mathfrak{D}}^{n+1}, u_{1,\mathcal{T}}^{n+1} - g_{1,\mathcal{T}} \right] \right]_{\mathcal{T}}, \\ A_2 = & -\delta t \left[\left[\text{div}^{\mathcal{T}} J_{2,\mathfrak{D}}^{n+1} - \text{div}^{\mathcal{T}} (V_{\mathfrak{D}}^{n+1} J_{1,\mathfrak{D}}^{n+1}), u_{2,\mathcal{T}}^{n+1} - g_{2,\mathcal{T}} \right] \right]_{\mathcal{T}} \\ & - \delta t \left[\left[V_{\mathcal{T}}^{n+1} \text{div}^{\mathcal{T}} (J_{1,\mathfrak{D}}^{n+1}), u_{2,\mathcal{T}}^{n+1} - g_{2,\mathcal{T}} \right] \right]_{\mathcal{T}} \\ & + \delta t \left[\left[W_{\mathcal{T}}^{n+1}, u_{2,\mathcal{T}}^{n+1} - g_{2,\mathcal{T}} \right] \right]_{\mathcal{T}}. \end{aligned} \tag{63}$$

Since by (38)–(39) we remark that $u_{i,\mathcal{T}}^{n+1} - g_{i,\mathcal{T}}$ satisfies zero Dirichlet boundary conditions, one can apply the simplified Green's formula (22) to A_1 and to the first term of A_2, that is

$$\begin{aligned} A_1 = & \delta t (J_{1,\mathfrak{D}}^{n+1}, \nabla^{\mathfrak{D}} (u_{1,\mathcal{T}}^{n+1} - g_{1,\mathcal{T}}))_{\mathfrak{D}}, \\ A_2 = & \delta t (J_{2,\mathfrak{D}}^{n+1} - V_{\mathfrak{D}}^{n+1} J_{1,\mathfrak{D}}^{n+1}, \nabla^{\mathfrak{D}} (u_{2,\mathcal{T}}^{n+1} - g_{2,\mathcal{T}}))_{\mathfrak{D}} \\ & - \delta t \left[\left[V_{\mathcal{T}}^{n+1} \text{div}^{\mathcal{T}} (J_{1,\mathfrak{D}}^{n+1}), u_{2,T}^{n+1} - g_{2,\mathcal{T}} \right] \right]_{\mathcal{T}} \\ & + \delta t \left[\left[W_{\mathcal{T}}^{n+1}, u_{2,\mathcal{T}}^{n+1} - g_{2,\mathcal{T}} \right] \right]_{\mathcal{T}}. \end{aligned} \tag{64}$$

We now turn to the term B given in (60). Since $a^2 - b^2 \leq 2a(a-b)$, it holds

$$B \leq -\lambda^2 g_2 \left(\nabla^{\mathfrak{D}} (V_{\mathcal{T}}^{n+1} - h_{\mathcal{T}}), \nabla^{\mathfrak{D}} (V_{\mathcal{T}}^{n+1} - V_{\mathcal{T}}^n) \right)_{\mathfrak{D}}. \tag{65}$$

Using the zero Dirichlet boundary conditions satisfied by $V_{\mathcal{T}}^{n+1} - h_{\mathcal{T}}$, we apply the Green's formula (22), leading to

$$B \leq \lambda^2 g_2 \left[\left[V_{\mathcal{T}}^{n+1} - h_{\mathcal{T}}, \text{div}^{\mathcal{T}} (\nabla^{\mathfrak{D}} (V_{\mathcal{T}}^{n+1} - V_{\mathcal{T}}^n)) \right] \right]_{\mathcal{T}}. \tag{66}$$

We now can use the numerical scheme and in particular the discrete Poisson equations (35)–(36) to get

$$B \leq g_2 \left[\left[V_{\mathcal{T}}^{n+1} - h_{\mathcal{T}}, \rho_{1,\mathcal{T}}^{n+1} - \rho_{1,\mathcal{T}}^n \right] \right]_{\mathcal{T}}. \tag{67}$$

Using the numerical schemes (31)–(32) on $\rho_{1,\mathcal{T}}$, we obtain

$$B \leq -\delta t g_2 \left[\left[V_{\mathcal{T}}^{n+1} - h_{\mathcal{T}}, \text{div}^{\mathcal{T}} (J_{1,\mathfrak{D}}^{n+1}) \right] \right]_{\mathcal{T}}. \tag{68}$$

Combining expressions (64) of terms A_i, $i = 1, 2$, and the bound (68) of the term B, it holds

$$\begin{aligned}\frac{S^{n+1} - S^n}{\delta t} &\leq (J_{1,\mathfrak{D}}^{n+1}, \nabla^{\mathfrak{D}}(u_{1,\mathcal{T}} - g_{1,\mathcal{T}}))_{\mathfrak{D}} + (J_{2,\mathfrak{D}}^{n+1} - V_{\mathfrak{D}}^{n+1} J_{1,\mathfrak{D}}^{n+1}, \nabla^{\mathfrak{D}}(u_{2,\mathcal{T}}^{n+1} - g_{2,\mathcal{T}}))_{\mathfrak{D}} \\ &\quad - \left[\left[V_{\mathcal{T}}^{n+1} \mathrm{div}^{\mathcal{T}}(J_{1,\mathfrak{D}}^{n+1}), u_{2,\mathcal{T}}^{n+1} - g_{2,\mathcal{T}}\right]\right]_{\mathcal{T}} \\ &\quad + \left[\left[W_{\mathcal{T}}^{n+1}, u_{2,\mathcal{T}}^{n+1} - g_{2,\mathcal{T}}\right]\right]_{\mathcal{T}} \\ &\quad - g_2 \left[\left[V_{\mathcal{T}}^{n+1} - h_{\mathcal{T}}, \mathrm{div}^{\mathcal{T}}(J_{1,\mathfrak{D}}^{n+1})\right]\right]_{\mathcal{T}}.\end{aligned} \tag{69}$$

Using the change of variables (41) and the property (49) on the fluxes, as well as the assumption (58) on the boundary conditions, the inequality (69) becomes, up to a reassortment of the terms,

$$\begin{aligned}\frac{S^{n+1} - S^n}{\delta t} &\leq (I_{1,\mathfrak{D}}^{n+1}, \nabla^{\mathfrak{D}} w_{1,\mathcal{T}})_{\mathfrak{D}} + (I_{2,\mathfrak{D}}^{n+1}, \nabla^{\mathfrak{D}} w_{2,\mathcal{T}}^{n+1})_{\mathfrak{D}} \\ &\quad + \left[\left[W_{\mathcal{T}}^{n+1}, u_{2,\mathcal{T}}^{n+1} - g_{2,\mathcal{T}}\right]\right]_{\mathcal{T}} \\ &\quad - (I_{1,\mathfrak{D}}^{n+1}, \nabla^{\mathfrak{D}}(w_{2,\mathcal{T}}^{n+1} V_{\mathcal{T}}^{n+1} - \bar{g}_{2,\mathcal{T}} h_{\mathcal{T}}))_{\mathfrak{D}} \\ &\quad - \left[\left[V_{\mathcal{T}}^{n+1} \mathrm{div}^{\mathcal{T}}(I_{1,\mathfrak{D}}^{n+1}), w_{2,\mathcal{T}}^{n+1} - \bar{g}_{2,\mathcal{T}}\right]\right]_{\mathcal{T}} \\ &\quad - g_2 \left[\left[V_{\mathcal{T}}^{n+1} - h_{\mathcal{T}}, \mathrm{div}^{\mathcal{T}}(I_{1,\mathfrak{D}}^{n+1})\right]\right]_{\mathcal{T}}.\end{aligned} \tag{70}$$

It remains now to prove that the three last terms cancel each others. Using the adapted Green formula (22), we have

$$\begin{aligned}&(I_{1,\mathfrak{D}}^{n+1}, \nabla^{\mathfrak{D}}(\bar{g}_{2,\mathcal{T}} h_{\mathcal{T}} - w_{2,\mathcal{T}}^{n+1} V_{\mathcal{T}}^{n+1}))_{\mathfrak{D}} \\ &= - \left[\left[\mathrm{div}^{\mathcal{T}}(I_{1,\mathfrak{D}}^{n+1}), \bar{g}_{2,\mathcal{T}} h_{\mathcal{T}} - w_{2,\mathcal{T}}^{n+1} V_{\mathcal{T}}^{n+1}\right]\right]_{\mathcal{T}} \\ &= \left[\left[V_{\mathcal{T}}^{n+1} \mathrm{div}^{\mathcal{T}}(I_{1,\mathfrak{D}}^{n+1}), w_{2,\mathcal{T}}^{n+1}\right]\right]_{\mathcal{T}} - \bar{g}_2 \left[\left[\mathrm{div}^{\mathcal{T}}(I_{1,\mathfrak{D}}^{n+1}), h_{\mathcal{T}}\right]\right]_{\mathcal{T}} \\ &= \left[\left[V_{\mathcal{T}}^{n+1} \mathrm{div}^{\mathcal{T}}(I_{1,\mathfrak{D}}^{n+1}), w_{2,\mathcal{T}}^{n+1} - \bar{g}_{2,\mathcal{T}}\right]\right]_{\mathcal{T}} + \left[\left[V_{\mathcal{T}}^{n+1} \mathrm{div}^{\mathcal{T}}(I_{1,\mathfrak{D}}^{n+1}), \bar{g}_{2,\mathcal{T}}\right]\right]_{\mathcal{T}} \\ &\quad - \bar{g}_2 \left[\left[\mathrm{div}^{\mathcal{T}}(I_{1,\mathfrak{D}}^{n+1}), h_{\mathcal{T}}\right]\right]_{\mathcal{T}} \\ &= \left[\left[V_{\mathcal{T}}^{n+1} \mathrm{div}^{\mathcal{T}}(I_{1,\mathfrak{D}}^{n+1}), w_{2,\mathcal{T}}^{n+1} - \bar{g}_{2,\mathcal{T}}\right]\right]_{\mathcal{T}} + \bar{g}_2 \left[\left[\mathrm{div}^{\mathcal{T}}(I_{1,\mathfrak{D}}^{n+1}), V_{\mathcal{T}}^{n+1} - h_{\mathcal{T}}\right]\right]_{\mathcal{T}}.\end{aligned} \tag{71}$$

Finally the inequality (70) sums up to

$$\begin{aligned}\frac{S^{n+1} - S^n}{\delta t} &\leq (I_{1,\mathfrak{D}}^{n+1}, \nabla^{\mathfrak{D}} w_{1,\mathcal{T}})_{\mathfrak{D}} + (I_{2,\mathfrak{D}}^{n+1}, \nabla^{\mathfrak{D}} w_{2,\mathcal{T}}^{n+1})_{\mathfrak{D}}, \\ &\quad + \left[\left[W_{\mathcal{T}}^{n+1}, u_{2,\mathcal{T}}^{n+1} - g_{2,\mathcal{T}}\right]\right]_{\mathcal{T}},\end{aligned} \tag{72}$$

which is equivalent to the desired inequality (59), using the definitions (53) of the fluxes $I_{i,\mathfrak{D}}^{n+1}$, $i = 1, 2$. The dissipation in time of the entropy function results from the positive definiteness of the matrix $D^{\mathfrak{D}}$ and from the dissipation Assumption 1-iv. of the relaxation term W. □

The dissipation of the discrete entropy allows to state some *a priori* estimate on the discrete solution to the energy-transport model.

Proposition 3 (*A priori* estimate)**.** Under Assumptions 1, let us assume moreover that the boundary conditions are at thermal equilibrium that is (58) is satisfied. Let $(\mathbf{u}_\mathcal{T}^n, V_\mathcal{T}^n)_{n\geq 0}$ be a solution to the energy-transport scheme (31)-(37). Then there exists a constant C_1, depending only on the initial and boundary conditions and the Debye length λ^2 such that

$$\sup_{n=0,\dots,N} \left(\|\mathbf{u}_\mathcal{T}^n - \mathbf{g}_\mathcal{T}\|_\mathcal{T}^2 + \|\nabla^\mathfrak{D}(V_\mathcal{T}^n - h_\mathcal{T})\|_\mathfrak{D}^2 \right) \leq C_1. \tag{73}$$

Proof. The proof relies on a reformulation of the discrete entropy fonction using the function χ. Since $\chi \in C^1(\mathbb{R}^2, \mathbb{R})$ such that $\nabla_\mathbf{u}\chi = \boldsymbol{\rho}$, according to Assumption 1-i., a Taylor expansion gives

$$\chi(\mathbf{u}_K^n) - \chi(\mathbf{g}_K) = \int_0^1 \boldsymbol{\rho}(\mathbf{g}_K + s(\mathbf{u}_K^n - \mathbf{g}_K)) \cdot (\mathbf{u}_K^n - \mathbf{g}_K) ds, \tag{74}$$

which can be rewritten as follows

$$\begin{aligned} &\int_0^1 \boldsymbol{\rho}(\mathbf{g}_K + s(\mathbf{u}_K^n - \mathbf{g}_K)) \cdot (\mathbf{u}_K^n - \mathbf{g}_K) ds \\ &= \int_0^1 \boldsymbol{\rho}(\mathbf{g}_K + s(\mathbf{u}_K^n - \mathbf{g}_K)) \cdot (\mathbf{u}_K^n - (\mathbf{g}_K + s(\mathbf{u}_K^n - \mathbf{g}_K))) \frac{ds}{1-s}. \end{aligned} \tag{75}$$

Plugging these expressions in the entropy definition (57) and using the monotony Assumption 1-i. on $\boldsymbol{\rho}$ yields

$$\begin{aligned} S^0 \geq S^n \geq \frac{C_0}{2} \left(\sum_{K\in\mathfrak{M}} m_K |\mathbf{u}_K^n - \mathbf{g}_K|^2 + \sum_{K^*\in\overline{\mathfrak{M}^*}} m_{K^*} |\mathbf{u}_{K^*}^n - \mathbf{g}_{K^*}|^2 \right) \\ - \frac{\lambda^2}{2} g_2 \|\nabla^\mathfrak{D}(V_T^n - h_\mathcal{T})\|_\mathfrak{D}^2, \end{aligned} \tag{76}$$

which is the desired estimate. □

3.3. **Existence of a solution to the schemes.** Using the estimates obtained in the previous section, we are now able to prove the existence of a solution to the scheme (31)–(37) in primal entropy variables $(\mathbf{u}, V)$ using a Leray-Schauder fixed point theorem. As a consequence of the equivalence of the schemes stated in Proposition 1, we also get the existence of a solution to the scheme (43)–(53) in entropic variables $(\mathbf{w}, V)$.

Theorem 1 (Existence of a solution to the schemes)**.** Under Assumptions 1, assume that the mesh $\mathcal{T}$ satisfies the regularity conditions (18)–(19) and that (58) holds such that the Dirichlet boundary conditions are at thermal equilibrium. Then, the scheme (31)–(37) for the energy-transport system admits a solution $(\mathbf{u}_\mathcal{T}^n, V_\mathcal{T}^n)$ for all $n \geq 0$. As a consequence of the equivalence property, the scheme (43)–(53) for the symmetrized system has also a solution $(\mathbf{w}_\mathcal{T}^n, V_\mathcal{T}^n)$ for all $n \geq 0$.

Proof. The proof is done by induction on $n \geq 0$, using the Leray-Schauder fixed point theorem. Let $n \geq 0$, and assume that $\mathbf{w}_\mathcal{T}^n$ and $V_\mathcal{T}^n$ are known. We build an application

$$\begin{aligned} \mathcal{L}^n : (\mathbb{R}^2)^\mathcal{T} \times [0,1] &\to (\mathbb{R}^2)^\mathcal{T} \\ (\tilde{\mathbf{u}}_\mathcal{T}, \kappa) &\mapsto \mathbf{u}_\mathcal{T}, \end{aligned}$$

to which we will apply the fixed point theorem. More precisely, starting from $(\tilde{\mathbf{u}}_\mathcal{T}, \kappa) \in (\mathbb{R}^2)^\mathcal{T} \times [0,1]$, $\mathbf{u}_\mathcal{T} := \mathcal{L}^n(\tilde{\mathbf{u}}_\mathcal{T})$ is defined in four steps as follows.

Step 1: We define $\mathrm{V}_{\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}$ as the solution of the following linear problem:

$$\begin{cases} -\lambda^2 \mathrm{div}^K(\nabla^{\mathfrak{D}} \mathrm{V}_{\mathcal{T}}) = C_K - \rho_1(\tilde{\mathbf{u}}_K), & \forall K \in \mathfrak{M}, \\ -\lambda^2 \mathrm{div}^{K^*}(\nabla^{\mathfrak{D}} \mathrm{V}_{\mathcal{T}}) = C_{K^*} - \rho_1(\tilde{\mathbf{u}}_{K^*}), & \forall K^* \in \mathfrak{M}^* \cup \partial\mathfrak{M}^*_N, \end{cases} \tag{77}$$

with

$$V_K = h_K, \quad \forall K \in \partial\mathfrak{M}_D, \quad V_{K^*} = h_{K^*}, \quad \forall K^* \in \partial\mathfrak{M}^*_D, \tag{78}$$

$$\nabla^{\mathcal{D}} V_{\mathcal{T}} \cdot \mathbf{n}_{\sigma K} = 0, \quad \forall \mathcal{D} \in \mathcal{D}_{\sigma,\sigma^*} \in \mathfrak{D}_{ext,N}. \tag{79}$$

Step 2: We use the change of variables (41) to define $\tilde{\mathbf{w}}_{\mathcal{T}} \in (\mathbb{R}^2)^{\mathcal{T}}$ as

$$\tilde{w}_{1,\mathcal{T}} = \tilde{u}_{1,\mathcal{T}} + \tilde{u}_{2,\mathcal{T}} V_{\mathcal{T}}, \qquad \tilde{w}_{2,\mathcal{T}} = \tilde{u}_{2,\mathcal{T}}. \tag{80}$$

Step 3: We define $\mathbf{w}_{\mathcal{T}} \in (\mathbb{R}^2)^{\mathcal{T}}$ as the solution of the following linear problem:

$$\mathrm{div}^K\left(-\tilde{D}^{\mathfrak{D}}_{11}\nabla^{\mathfrak{D}} w_{1,\mathcal{T}} - \tilde{D}^{\mathfrak{D}}_{12}\nabla^{\mathfrak{D}} w_{2,\mathcal{T}}\right) = - \kappa \frac{b_1(\tilde{\mathbf{w}}_K, \mathrm{V}_K) - b_1(\mathbf{w}^n_K, \mathrm{V}^n_K)}{\delta t}, \qquad \forall K \in \mathfrak{M}, \tag{81}$$

$$\mathrm{div}^{K^*}\left(-\tilde{D}^{\mathfrak{D}}_{11}\nabla^{\mathfrak{D}} w_{1,\mathcal{T}} - \tilde{D}^{\mathfrak{D}}_{12}\nabla^{\mathfrak{D}} w_{2,\mathcal{T}}\right) = - \kappa \frac{b_1(\tilde{\mathbf{w}}_{K^*}, \mathrm{V}_{K^*}) - b_1(\mathbf{w}^n_{K^*}, \mathrm{V}^n_{K^*})}{\delta t}, \qquad \forall K^* \in \mathfrak{M}^* \cup \partial\mathfrak{M}^*_N, \tag{82}$$

$$\mathrm{div}^K\left(-\tilde{D}^{\mathfrak{D}}_{21}\nabla^{\mathfrak{D}} w_{1,\mathcal{T}} - \tilde{D}^{\mathfrak{D}}_{22}\nabla^{\mathfrak{D}} w_{2,\mathcal{T}}\right) = \kappa \left(- \frac{b_2(\tilde{\mathbf{w}}_K, \mathrm{V}_K) - b_2(\mathbf{w}^n_K, \mathrm{V}^n_K)}{\delta t} + \bar{\mathrm{W}}(\tilde{\mathbf{w}}_K, \mathrm{V}_K) - b_1(\mathbf{w}^n_K, \mathrm{V}^n_K) \frac{\mathrm{V}_K - \mathrm{V}^n_K}{\delta t} \right), \qquad \forall K \in \mathfrak{M}, \tag{83}$$

$$\mathrm{div}^{K^*}\left(-\tilde{D}^{\mathfrak{D}}_{21}\nabla^{\mathfrak{D}} w_{1,\mathcal{T}} - \tilde{D}^{\mathfrak{D}}_{22}\nabla^{\mathfrak{D}} w_{2,\mathcal{T}}\right) = \kappa \left(- \frac{b_2(\tilde{\mathbf{w}}_{K^*}, \mathrm{V}_{K^*}) - b_2(\mathbf{w}^n_{K^*}, \mathrm{V}^n_{K^*})}{\delta t} + \bar{\mathrm{W}}(\tilde{\mathbf{w}}_{K^*}, \mathrm{V}_{K^*}) - b_1(\mathbf{w}^n_{K^*}, \mathrm{V}^n_{K^*}) \frac{\mathrm{V}_{K^*} - \mathrm{V}^n_{K^*}}{\delta t} \right), \quad \forall K^* \in \mathfrak{M}^* \cup \partial\mathfrak{M}^*_N, \tag{84}$$

where $\tilde{D}^{\mathfrak{D}} = (\tilde{D}^{\mathfrak{D}}_{ij})_{1\le i,j\le 2}$ is defined by $\tilde{D}^{\mathfrak{D}} = {}^t P_{\mathfrak{D}} \tilde{L}^{\mathfrak{D}} P_{\mathfrak{D}}$, with $P_{\mathfrak{D}} = \begin{pmatrix} 1 & -V_{\mathfrak{D}} \\ 0 & 1 \end{pmatrix}$ and $\tilde{L}^{\mathfrak{D}} = L(\tilde{\mathbf{u}}_{\mathcal{T}})$. Problem (81)–(84) is supplemented with the following boundary conditions:

$$\mathbf{w}_K = \kappa\, \bar{\mathbf{g}}_K \quad \forall K \in \partial\mathfrak{M}_{\mathcal{D}}, \qquad \mathbf{w}_{K^*} = \kappa\, \bar{\mathbf{g}}_{K^*} \quad \forall K^* \in \partial\mathfrak{M}^*_D, \tag{85}$$

$$\left(-\tilde{D}^{\mathfrak{D}}_{11}\nabla^{\mathfrak{D}} w_{1,\mathcal{T}} - \tilde{D}^{\mathfrak{D}}_{12}\nabla^{\mathfrak{D}} w_{2,\mathcal{T}}\right) \cdot \mathbf{n}_{\sigma K} = 0, \quad \forall \mathcal{D} = \mathcal{D}_{\sigma,\sigma^*} \in \mathfrak{D}_{ext,N}, \tag{86}$$

$$\left(-\tilde{D}^{\mathfrak{D}}_{21}\nabla^{\mathfrak{D}} w_{1,\mathcal{T}} - \tilde{D}^{\mathfrak{D}}_{22}\nabla^{\mathfrak{D}} w_{2,\mathcal{T}}\right) \cdot \mathbf{n}_{\sigma K} = 0, \quad \forall \mathcal{D} = \mathcal{D}_{\sigma,\sigma^*} \in \mathfrak{D}_{ext,N}. \tag{87}$$

Step 4: Finally, using again the change of variables (41), we define $\mathbf{u}_{\mathcal{T}} \in (\mathbb{R}^2)^{\mathcal{T}}$ as

$$u_{1,\mathcal{T}} = w_{1,\mathcal{T}} - w_{2,\mathcal{T}} V_{\mathcal{T}}, \qquad u_{2,\mathcal{T}} = w_{2,\mathcal{T}}. \tag{88}$$

We emphasize that a fixed point $\mathbf{u}_{\mathcal{T}}$ of $\mathcal{L}^n(\cdot, 1)$, together with the corresponding $V_{\mathcal{T}}$ defined at Step 1 is a solution to the DDFV scheme (31)–(37).

Let us now verify that $\mathcal{L}^n$ is well-defined. Existence and uniqueness of the solution $V_\mathcal{T}$ to (77)–(79) is guaranteed by the positive definiteness of the matrix corresponding to the linear system. This result is standard and we refer to [21] for details. Let us now prove that there exists a unique $\mathbf{w}_\mathcal{T}$ solution to (81)–(87). Since it is a linear system of finite dimension, it is sufficient to prove that if $\kappa = 0$, then $\mathbf{w}_\mathcal{T} = 0$. Hence assume that $\kappa = 0$. By multiplying (81) by $w_{1,K}$, (82) by w_{1,K^*}, (83) by $w_{2,K}$ and (84) by w_{2,K^*}, and then summing over $K \in \mathfrak{M}$ and $K^* \in \mathfrak{M}^* \cup \partial\mathfrak{M}_N^*$, and finally summing the four obtained expressions, it yields

$$\begin{aligned}&\left[\left[\text{div}^{\,\mathcal{T}}\left(-\tilde{D}_{11}^{\mathfrak{D}}\nabla^{\mathfrak{D}}w_{1,\mathcal{T}} - \tilde{D}_{12}^{\mathfrak{D}}\nabla^{\mathfrak{D}}w_{2,\mathcal{T}}\right),\, w_{1,\mathcal{T}}\right]\right]_\mathcal{T} + \\ &\left[\left[\text{div}^{\,\mathcal{T}}\left(-\tilde{D}_{21}^{\mathfrak{D}}\nabla^{\mathfrak{D}}w_{1,\mathcal{T}} - \tilde{D}_{22}^{\mathfrak{D}}\nabla^{\mathfrak{D}}w_{2,\mathcal{T}}\right),\, w_{2,\mathcal{T}}\right]\right]_\mathcal{T} = 0.\end{aligned}$$

Since $\kappa = 0$, the Dirichlet boundary conditions vanish, which allows to apply the discrete Green formula (22), leading to

$$\begin{aligned}&\left(-\tilde{D}_{11}^{\mathfrak{D}}\nabla^{\mathfrak{D}}w_{1,\mathcal{T}} - \tilde{D}_{12}^{\mathfrak{D}}\nabla^{\mathfrak{D}}w_{2,\mathcal{T}}, \nabla^{\mathfrak{D}}w_{1,\mathcal{T}}\right)_{\mathfrak{D}} + \\ &\qquad\left(-\tilde{D}_{21}^{\mathfrak{D}}\nabla^{\mathfrak{D}}w_{1,\mathcal{T}} - \tilde{D}_{22}^{\mathfrak{D}}\nabla^{\mathfrak{D}}w_{2,\mathcal{T}}, \nabla^{\mathfrak{D}}w_{2,\mathcal{T}}\right)_{\mathfrak{D}} = 0,\end{aligned}$$

which can be rewritten under the following condensed form

$$\sum_{\mathcal{D}\in\mathfrak{D}} {}^t\nabla^{\mathcal{D}}\mathbf{w}_\mathcal{T}\, \tilde{D}^{\mathfrak{D}}\, \nabla^{\mathcal{D}}\mathbf{w}_\mathcal{T} = 0.$$

Since $\tilde{D}^{\mathfrak{D}}$ is a symmetric positive definite matrix, it follows that $\nabla^{\mathfrak{D}}\mathbf{w}_\mathcal{T} = 0$. Using the discrete Poincaré inequality, see for example [4, Theorem 5.4], we conclude that $\mathbf{w}_\mathcal{T} = 0$. Hence it exists a unique solution $\mathbf{w}_\mathcal{T}$ to (81)-(87) and the application $\mathcal{L}^n$ is well-defined.

Now we have to verify that assumptions of the Leray-Schauder fixed point theorem are fulfilled. We already proved that $\mathcal{L}^n(\tilde{\mathbf{u}}_\mathcal{T}, \kappa = 0) = 0$. It remains to show that there exists a constant $M > 0$ such that, for all $\kappa \in [0,1]$, for all $\mathbf{u}_\mathcal{T} \in (\mathbb{R}^2)^\mathcal{T}$ satisfying $\mathcal{L}^n(\mathbf{u}_\mathcal{T}, \kappa) = \mathbf{u}_\mathcal{T}$, it holds $\|\mathbf{u}_\mathcal{T}\|_2 \leq M$. Let us first emphasize that if $\mathcal{L}^n(\mathbf{u}_\mathcal{T}, \kappa) = \mathbf{u}_\mathcal{T}$, then $\tilde{\mathbf{u}}_\mathcal{T} = \mathbf{u}_\mathcal{T}$, and, as a consequence, $\tilde{\mathbf{w}}_\mathcal{T} = \mathbf{w}_\mathcal{T}$. Thus, we can adapt the proofs of Propositions 2 and 3, obtained with $\kappa = 1$, to the more general case $\kappa \in [0,1]$. It gives

$$\|\mathbf{u}_\mathcal{T}\|_2^2 \leq 2\|\mathbf{u}_\mathcal{T} - \kappa\mathbf{g}_\mathcal{T}\|_2^2 + 2\kappa\|\mathbf{g}_\mathcal{T}\|_2^2 \leq \frac{4S^n}{C_0} + 2\|\mathbf{g}_\mathcal{T}\|_2^2,$$

where S^n is the discrete entropy (57) defined using the known quantities $\mathbf{w}_\mathcal{T}^n$ and $\mathrm{V}_\mathcal{T}^n$, $\mathbf{g}_\mathcal{T}$ is the known discrete Dirichlet boundary condition and C_0 is the constant arising in Assumptions 1-i.. Finally, since $\mathcal{L}^n$ is continuous, the Leray-Schauder fixed point theorem applies and leads to the desired conclusion. □

In order to prove the convergence of the schemes, one would need H^1 estimates on $\mathbf{w}_\mathcal{T}$, which require L^∞ bounds on the potential $V_\mathcal{T}$. Such bounds are available for the two-point flux finite volume approximation [6, 5] and are deduced from the particular structure of the matrix involved in the numerical scheme which ensures the maximum principle on the potential $V_\mathcal{T}$ [7]. In the present DDFV formalism, the maximum principle is not guaranteed. As a consequence it is not possible to adapt the proof proposed in [12] to exhibit H^1 estimates on $\mathbf{w}_\mathcal{T}$ and to prove compactness results leading to the convergence of the schemes.

4. Numerical experiments

4.1. **Physical models and parameters.** The physical models we consider enter the framework introduced in [22, 25]. The physical assumptions are the following:

- the electron density ρ_1 is given by nondegenerate Boltzmann statistics,
- the energy-band diagram is parabolic,
- the energy relaxation term W is given by a Fokker-Planck approximation.

The densities ρ_1 and ρ_2 are then defined as the following nonlinear functions of $\mathbf{u}$:

$$\begin{cases} \rho_1(\mathbf{u}) = \left(-\dfrac{1}{u_2}\right)^{\frac{3}{2}} \exp(u_1), \\ \rho_2(\mathbf{u}) = \dfrac{3}{2}\left(-\dfrac{1}{u_2}\right)^{\frac{5}{2}} \exp(u_1) = \dfrac{3}{2}\rho_1 T, \end{cases}$$

where the temperature is $T = -1/u_2$. The matrix L and the relaxation term W read:

$$L = c_0\rho_1 T^{1/2-\beta} \begin{pmatrix} 1 & (2-\beta)T \\ (2-\beta)T & (3-\beta)(2-\beta)T^2 \end{pmatrix}, \tag{89}$$

$$W = c_1\rho_1 \frac{\ell^2}{\tau_0\mu_0 U_T} T^{\beta-1/2}(1-T), \tag{90}$$

where c_0, c_1 and β are constants. The usual values are $c_0 = 1$, $c_1 = 3/2$, $\beta = 1/2$, for the Chen model, and $c_0 = 2/\sqrt{\pi}$, $c_1 = 2/\sqrt{\pi}$, $\beta = 0$ for the Lyumkis model. We can then rewrite:

- for the Chen model:

$$L_{chen} = \rho_1 \begin{pmatrix} 1 & \dfrac{3}{2}T \\ \dfrac{3}{2}T & \dfrac{15}{4}T^2 \end{pmatrix},$$

$$W_{chen} = \left(\frac{3}{2}\frac{\ell^2}{\tau_0\mu_0 U_T}\right)\rho_1 - \left(\frac{\ell^2}{\tau_0\mu_0 U_T}\right)\rho_2,$$

- for the Lyumkis model:

$$L_{lyum} = \frac{2}{\sqrt{\pi}}\sqrt{T}\ \rho_1 \begin{pmatrix} 1 & 2T \\ 2T & 6T^2 \end{pmatrix},$$

$$W_{lyum} = \left(\frac{2}{\sqrt{\pi}}\frac{1}{\sqrt{T}}\frac{\ell^2}{\tau_0\mu_0 U_T}\right)\rho_1 - \left(\frac{4}{3\sqrt{\pi}}\frac{1}{\sqrt{T}}\frac{\ell^2}{\tau_0\mu_0 U_T}\right)\rho_2.$$

We focus on the test case of a n^+nn^+ silicon diode which is uniform in one space dimension and compare to the numerical experiments obtained in [22, 8].

The semiconductor domain is $\Omega = (0, l_x) \times (0, l_y)$ where $l_x = 0.6\mu m$ and $l_y = 0.2\mu m$ and the length of the channel is $0.4\mu m$. The device geometry is illustrated in Figure 4-top. All the numerical results are presented using the corresponding scaled geometry, see 4-bottom.

The physical parameters appearing in the scaled energy-tranport model (5)-(6), correspond to those of [22] and are listed in Table 1.

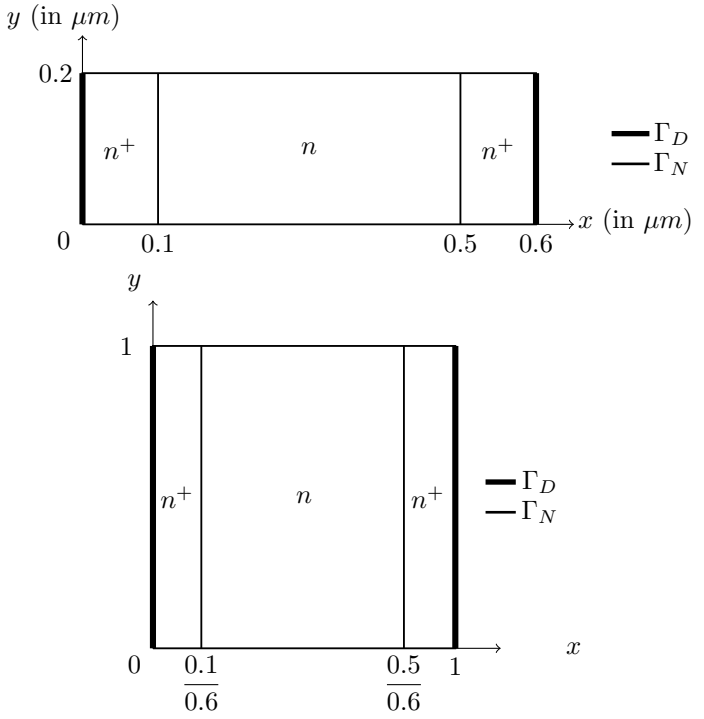

FIGURE 4. Geometry of the n^+nn^+ ballistic diode: physical geometry (top) and scaled geometry (bottom).

Parameter	**Physical meaning**	**Numerical value**
q	elementary charge	1.6d-19 As
ϵ_s	permittivity constant	1.d-12 As $V^{-1}s^{-1}$
μ_0	low field mobility	1.5d3 $cm^2V^{-1}s^{-1}$
T_0	ambient temperature	300 K
U_T	thermal voltage at T_0	0.0259 V
τ_0	energy relaxation time	0.4d-12 s

TABLE 1. Physical parameters for n^+nn^+ silicon diode.

The doping profile is

$$C(x,y) = \begin{cases} C_n/C_m, & \text{in the } n \text{ region,} \\ 1, & \text{in the } n^+ \text{ region,} \end{cases}$$

where $C_m = 5 \times 10^{17} cm^{-3}$ and $C_n = 2 \times 10^{15} cm^{-3}$.

The Dirichlet boundary conditions for the scaled system are defined as follows:

- $u_1 = 0$, $u_2 = -1$
- $V = 1.5V/U_T$ if $x = 0$, $V = 0$ otherwise.

4.2. **Details on the implementation of the method.** The discretization is fully implicit in time, which allows us to avoid too restrictive time steps; we implement a Newton method. It implies to solve a large nonlinear system at each time iteration; the inversion of the associated Jacobian matrix is done by a direct method, appealing to *Umfpack* libraries.

The Newton method stops when one of these two conditions is fullfilled: a) the difference between two successive solutions is smaller than 10^{-9}, b) the number of iterations is larger than 30.

In order to allow the convergence of the method, an adaptative time stepping technique is implemented. The minimal time step is $\Delta t_{\min} = 10^{-8}$ and the maximal time step is $\Delta t_{\max} = 10^{-4}$. The adaptative algorithm starts with $\Delta t = \Delta t_{\max}$. At each time iteration, we consider that the Newton method has converged if the stopping criterion a) is fulfilled and the quantity $u_{2,\mathcal{T}}$ is negative, that is if the temperature is positive. If not, the time step is modified and set to $\max(\Delta t/2, \Delta t_{\min})$ and the Newton algorithm is relaunched until convergence is achieved. Note that, at the next time iteration, the current time step is set to $\min(2\Delta t, \Delta t_{\max})$.

We present all the simulations on a refinement with 2560 cells of the non-conformal locally refined mesh of Figure 5, since equivalent results have been obtained on all the meshes presented in Figure 5.

4.3. **Comparison of the two schemes.** We propose a comparison of the schemes for the scaled energy transport model (31)–(40) and the symmetrized model (43)–(52), focusing on their performances regarding the number of Newton iterations.

	Newton Iterations	**Number of time steps**	**Adaptative time steps**
Scaled system (Chen)	13405	10010	19
Symmetrized system (Chen)	27155	10409	1023
Scaled system (Lyumkis)	20704	10013	25
Symmetrized system (Lyumkis)	37417	10326	663

TABLE 2. Comparison of the schemes for the scaled energy-tranport model and for the symmetrized system for both the Chen and Lyumkis models.

Table 2 records for each scheme and model: a) the total number of Newton iterations, including iterations for which the Newton method did not converge, b) the total number of time steps, c) the number of adapted time steps. One observes that the number of iterations is much more important with the scheme for the symmetrized system. Such an observation has already been done for the two-point flux approximation in the one dimensional setting [5]. Figure 6 illustrates the repartition in time of the Newton iterations. It emphasizes that the scaled energy-transport scheme requires less Newton iterations and one recovers the erratic amount of Newton iterations for the symmetrized model.

The Figures 7, 8 and 9 present the electron mean velocity, defined as $v^{el} = \|J_1\|_1/(q\rho_1)$, the electrostatic potential V and the temperature $-1/u_2$ respectively. The results correspond to a final time of computation $T_{\max} = 1$, which ensures to reach the equilibrium, for both the Chen and the Lyumkis models. Note that

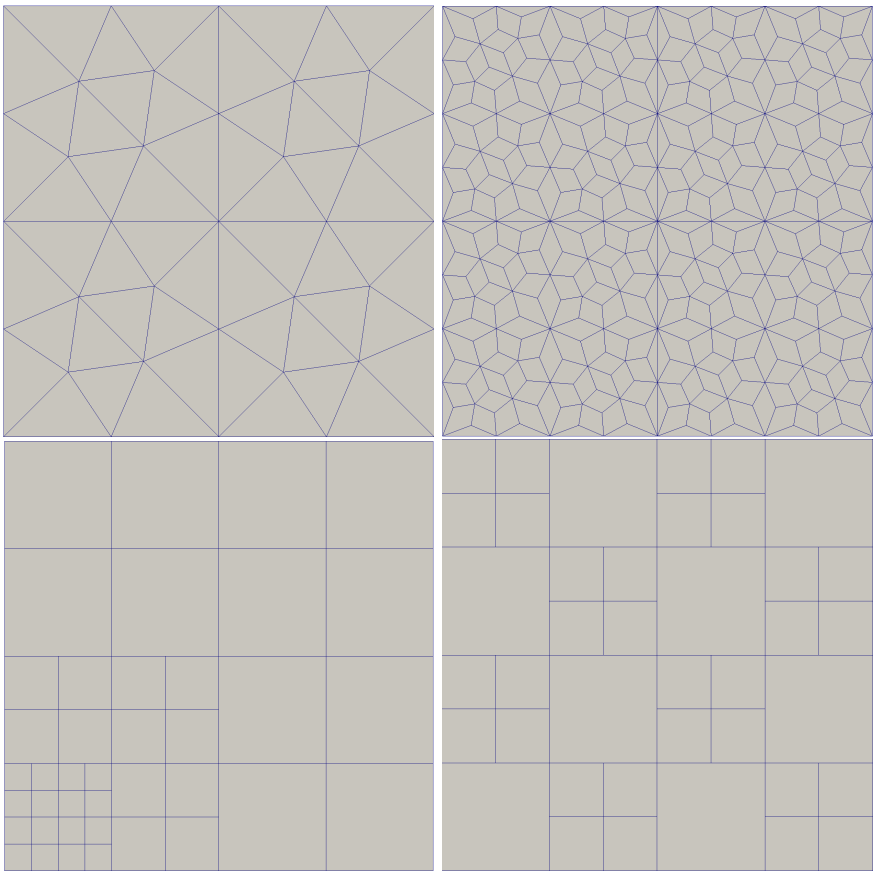

FIGURE 5. Considered meshes for the 2D ballistic diode.

the same profiles are obtained whatever the considered scheme and mesh are. The maximal time step is $\Delta t_{\max} = 10^{-4}$ but one could consider $\Delta t_{\max} = 10^{-3}$ for the scheme associated to the scaled energy-transport model. As expected, one observe that the computed quantities are uniform in one space direction. We compare now the results obtained with the DDFV schemes with the ones in [13, 22]; in particular, we consider as references values the ones in [13]. The temperature is high in the n-channel due to the hot electron effect. For the Chen model, its maximum is attained at $T = 7.86$, which corresponds (before scaling) to $T = 2358K$: the reference is $T_{ref} = 2330K$. For the Lyumkis model, the maximum is at $T = 12.9$, which corresponds (before scaling) to $T = 3871K$: the reference is $T_{ref} = 3970K$. In Fig. 7 are presented the values of the electron mean velocity. For the Chen model, the maximum is attained at $v^{el} = 23.2$, which corresponds (before scaling) to $v^{el} = 1.49 \cdot 10^7$ cm/s: the reference is $v^{el} = 1.44 \cdot 10^7$ cm/s. For the Lyumkis model, the maximum is at $v^{el} = 42.48$, which corresponds (before scaling) to $v^{el} = 2.75 \cdot 10^7$ cm/s: the reference is $v^{el} = 2.92 \cdot 10^7$ cm/s. All the results are in agreement with the numerical simulations presented in [13, 22].

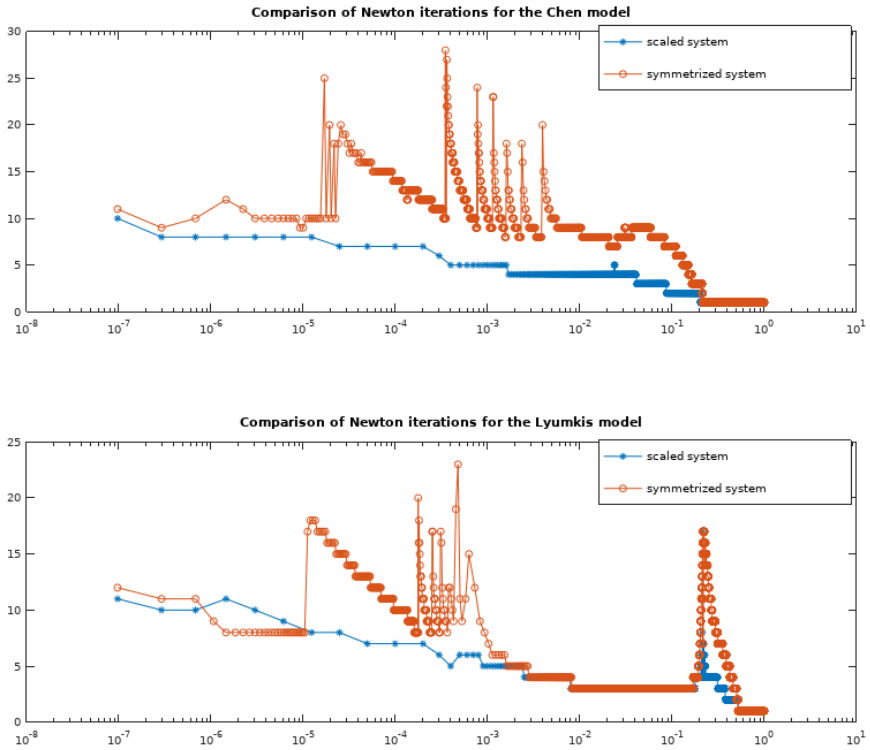

FIGURE 6. Repartition of the Newton iterations for the scaled energy-transport model and the symmetrized model.

5. Conclusion

We have presented in this paper DDFV schemes for the energy-transport systems. By adapting at the discrete level the techniques of [12], we manage to prove the equivalence of the schemes for the scaled system and the symmetrized one. It allows to establish a discrete entropy inequality from which we deduce *a priori* estimates and as a by-product, existence of solutions to the schemes. The two schemes are compared on a two-dimensional test case. It turns out that the scheme for the symmetrized system requires much more Newton iterations than the scheme for the scaled model. However both schemes provide qualitatively the same results, whatever the mesh is. In particular the numerical results obtained with a nonconformal refined mesh are in good agreement with the reference solutions provided in the literature.

Acknowledgements: The authors were partially funded by the Centre Henri Lebesgue (ANR-11-LABX-0020-01) and ANR Project MoHyCon (ANR-17-CE40-0027-01).

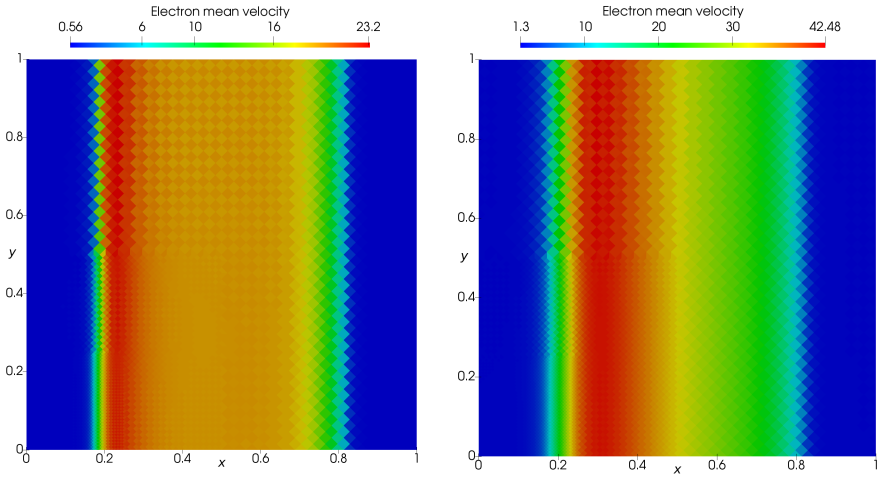

FIGURE 7. Electron mean velocity for the Chen (left) and the Lyumkis (right) models.

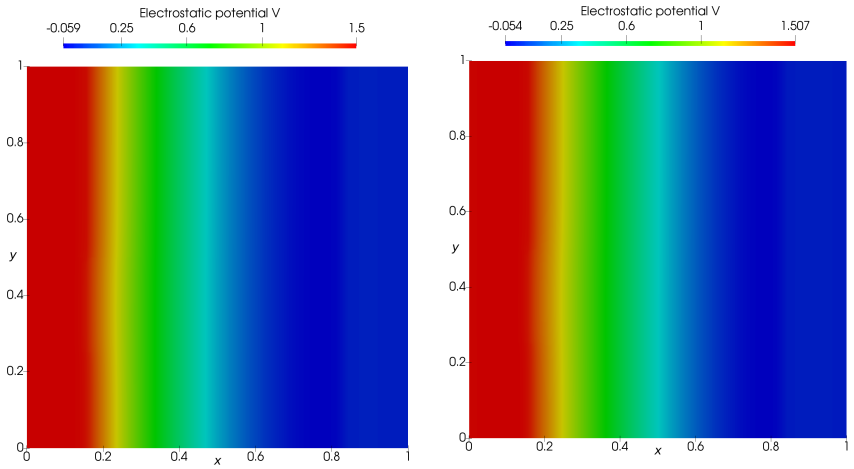

FIGURE 8. Electrostatic potential for the Chen (left) and the Lyumkis (right) models.

REFERENCES

[1] B. Andreianov, M. Bendahmane, F. Hubert, and S. Krell. On 3D DDFV discretization of gradient and divergence operators. I. Meshing, operators and discrete duality. *IMA J. Numer. Anal.*, 32(4):1574–1603, 2012.

[2] B. Andreianov, F. Boyer, and F. Hubert. Discrete duality finite volume schemes for Leray-Lions type elliptic problems on general 2D-meshes. *Num. Meth. for PDEs*, 23(1):145–195, 2007.

[3] N. Ben Abdallah and P. Degond. On a hierarchy of macroscopic models for semiconductors. *Journal of Mathematical Physics*, 37(7):3306–3333, 1996.

[4] M. Bessemoulin-Chatard, C. Chainais-Hillairet, and F. Filbet. On discrete functional inequalities for some finite volume schemes. *IMA J. Numer. Anal.*, 35(3):1125–1149, 2015.

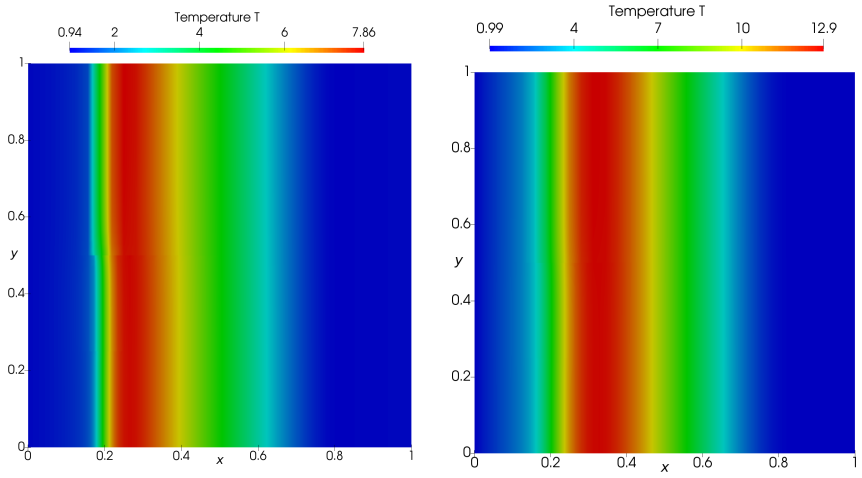

FIGURE 9. Temperature field for the Chen (left) and the Lyumkis (right) models.

[5] M. Bessemoulin-Chatard, C. Chainais-Hillairet, and H. Mathis. Analysis of numerical schemes for semiconductors energy-transport models. submitted, hal-02940224, 2020.

[6] M. Bessemoulin-Chatard, C. Chainais-Hillairet, and H. Mathis. Numerical schemes for semiconductors energy-transport models. *to appear in FVCA IX, Springer Proceedings in Mathematics and Statistics*, 2020.

[7] C. Cancès, C. Chainais-Hillairet, J. Fuhrmann, and B. Gaudel. A numerical analysis focused comparison of several finite volume schemes for a unipolar degenerate drift-diffusion model. Preprint, https://hal.archives-ouvertes.fr/hal-02194604v3/document, 2020.

[8] C. Chainais-Hillairet. Discrete duality finite volume schemes for two-dimensional drift-diffusion and energy-transport models. *Internat. J. Numer. Methods Fluids*, 59(3):239–257, 2009.

[9] C. Chainais-Hillairet and Y.-J. Peng. Finite volume scheme for semiconductor energy-transport model. In *Elliptic and parabolic problems*, volume 63 of *Progr. Nonlinear Differential Equations Appl.*, pages 139–146. Birkhäuser, Basel, 2005.

[10] L. Chen and L. Hsiao. The solution of Lyumkis energy transport model in semiconductor science. *Math. Methods Appl. Sci.*, 26(16):1421–1433, 2003.

[11] L. Chen, L. Hsiao, and Y. Li. Large time behavior and energy relaxation time limit of the solutions to an energy transport model in semiconductors. *J. Math. Anal. Appl.*, 312(2):596–619, 2005.

[12] P. Degond, S. Génieys, and A. Jüngel. A system of parabolic equations in nonequilibrium thermodynamics including thermal and electrical effects. *J. Math. Pures Appl. (9)*, 76(10):991–1015, 1997.

[13] P. Degond, A. Jüngel, and P. Pietra. Numerical discretization of energy-transport models for semiconductors with nonparabolic band structure. *SIAM J. Sci. Comput.*, 22(3):986–1007, 2000.

[14] P. Degond, C. D. Levermore, and C. Schmeiser. A note on the energy-transport limit of the semiconductor Boltzmann equation. In N. Ben Abdallah, I. M. Gamba, C. Ringhofer, A. Arnold, R. T. Glassey, P. Degond, and C. D. Levermore, editors, *Transport in Transition Regimes*, pages 137–153, New York, NY, 2004. Springer New York.

[15] K. Domelevo and P. Omnes. A finite volume method for the Laplace equation on almost arbitrary two-dimensional grids. *M2AN Math. Model. Numer. Anal.*, 39(6):1203–1249, 2005.

[16] R. Eymard, T. Gallouët, and R. Herbin. Finite volume methods. In *Handbook of numerical analysis*, volume VII, pages 713–1020. North-Holland, Amsterdam, 2000.

[17] W. Fang and K. Ito. Existence of stationary solutions to an energy drift-diffusion model for semiconductor devices. *Math. Models Methods Appl. Sci.*, 11(5):827–840, 2001.

[18] M. Fournié. Numerical discretization of energy-transport model for semiconductors using high-order compact schemes. *Appl. Math. Lett.*, 15(6):721–726, 2002.
[19] S. Gadau and A. Jüngel. A three-dimensional mixed finite-element approximation of the semiconductor energy-transport equations. *SIAM J. Sci. Comput.*, 31(2):1120–1140, 2008/09.
[20] J. A. Griepentrog. An application of the implicit function theorem to an energy model of the semiconductor theory. *ZAMM Z. Angew. Math. Mech.*, 79(1):43–51, 1999.
[21] F. Hermeline. A finite volume method for the approximation of diffusion operators on distorted meshes. *J. Comput. Phys.*, 160(2):481–499, 2000.
[22] S. Holst, A. Jüngel, and P. Pietra. A mixed finite-element discretization of the energy-transport model for semiconductors. *SIAM J. Sci. Comput.*, 24(6):2058–2075, 2003.
[23] S. Holst, A. Jüngel, and P. Pietra. An adaptive mixed scheme for energy-transport simulations of field-effect transistors. *SIAM J. Sci. Comput.*, 25(5):1698–1716, 2004.
[24] A. Jüngel. Regularity and uniqueness of solutions to a parabolic system in nonequilibrium thermodynamics. *Nonlinear Anal.*, 41(5-6, Ser. A: Theory Methods):669–688, 2000.
[25] A. Jüngel. *Transport equations for semiconductors*, volume 773 of *Lecture Notes in Physics*. Springer-Verlag, Berlin, 2009.
[26] A. Jüngel, R. Pinnau, and E. Röhrig. Existence analysis for a simplified transient energy-transport model for semiconductors. *Math. Methods Appl. Sci.*, 36(13):1701–1712, 2013.
[27] Ansgar Jüngel. *Quasi-hydrodynamic semiconductor equations*, volume 41 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser Verlag, Basel, 2001.
[28] S. Krell. *Schémas Volumes Finis en mécanique des fluides complexes*. PhD thesis, Univ. de Provence, 2010.
[29] P. A. Markowich, C. A. Ringhofer, and C. Schmeiser. *Semiconductor equations*. Springer-Verlag, Vienna, 1990.
[30] V. Romano. 2D numerical simulation of the MEP energy-transport model with a finite difference scheme. *J. Comput. Phys.*, 221(2):439–468, 2007.
[31] N. Zamponi and A. Jüngel. Global existence analysis for degenerate energy-transport models for semiconductors. *J. Differential Equations*, 258(7):2339–2363, 2015.

Laboratoire de Mathématiques Jean Leray, Université de Nantes & CNRS UMR 6629, BP 92208, F-44322 Nantes Cedex 3, France
Email address: marianne.bessemoulin@univ-nantes.fr

CEMEF - Mines ParisTech - CNRS UMR 7635, CS 10207, F-06904 Sophia Antipolis, France
Email address: giulia.lissoni@mines-paristech.fr

Laboratoire de Mathématiques Jean Leray, Université de Nantes & CNRS UMR 6629, BP 92208, F-44322 Nantes Cedex 3, France
Email address: helene.mathis@univ-nantes.fr