Journal of Computational Physics 520 (2025) 113477

Contents lists available at ScienceDirect

Journal of Computational Physics

journal homepage: www.elsevier.com/locate/jcp

DG-IMEX method for a two-moment model for radiation transport in the $O(v/c)$ limit ☆

M. Paul Laiu [a], Eirik Endeve [a,b,*], J. Austin Harris [c], Zachary Elledge [b], Anthony Mezzacappa [b]

[a] Multiscale Methods and Dynamics Group, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA
[b] Department of Physics and Astronomy, University of Tennessee Knoxville, TN 37996-1200, USA
[c] Advanced Computing for Nuclear, Particles, and Astrophysics Group, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

ARTICLE INFO

Keywords:
Boltzmann equation
Radiation transport
Hyperbolic conservation laws
Discontinuous Galerkin
Implicit-explicit
Moment realizability

ABSTRACT

We consider neutral particle systems described by moments of a phase-space density and propose a realizability-preserving numerical method to evolve a spectral two-moment model for particles interacting with a background fluid moving with nonrelativistic velocities. The system of nonlinear moment equations, with special relativistic corrections to $O(v/c)$, expresses a balance between phase-space advection and collisions and includes velocity-dependent terms that account for spatial advection, Doppler shift, and angular aberration. The model is conservative for the correct $O(v/c)$ Eulerian-frame number density and is consistent, to $O(v/c)$, with Eulerian-frame energy and momentum conservation. This model is closely related to the one promoted by Lowrie et al. [1] and similar to models currently used to study transport phenomena in large-scale simulations of astrophysical environments. The proposed numerical method is designed to preserve moment realizability, which guarantees that the moments correspond to a nonnegative phase-space density. The realizability-preserving scheme consists of the following key components: (i) a strong stability-preserving implicit-explicit (IMEX) time-integration method; (ii) a discontinuous Galerkin (DG) phase-space discretization with carefully constructed numerical fluxes; (iii) a realizability-preserving implicit collision update; and (iv) a realizability-enforcing limiter. In time integration, nonlinearity of the moment model necessitates solution of nonlinear equations, which we formulate as fixed-point problems and solve with tailored iterative solvers that preserve moment realizability with guaranteed global convergence. We also analyze the simultaneous Eulerian-frame number and energy conservation properties of the semi-discrete DG scheme and propose a "spectral redistribution" scheme that promotes Eulerian-frame energy conservation. Through numerical experiments, we demonstrate the accuracy and robustness of this DG-IMEX method and investigate its Eulerian-frame energy conservation properties.

☆ This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

* Corresponding author at: Multiscale Methods and Dynamics Group, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA.
E-mail addresses: laiump@ornl.gov (M. Paul Laiu), endevee@ornl.gov (E. Endeve), harrisja@ornl.gov (J. Austin Harris), zelledge@vols.utk.edu (Z. Elledge), mezz@utk.edu (A. Mezzacappa).

https://doi.org/10.1016/j.jcp.2024.113477
Received 20 September 2023; Received in revised form 12 August 2024; Accepted 30 September 2024

Available online 3 October 2024

1. Introduction

In this paper, we design and analyze a numerical method for solving a system of moment equations that model transport of neutral particles (e.g., photons, neutrons, or neutrinos) interacting with a background fluid moving with nonrelativistic velocities — i.e., flows in which the ratio of the background flow velocity to the speed of light, v/c, is sufficiently small such that special relativistic corrections of order $(v/c)^2$ and higher can be neglected. Similar $O(v/c)$ models have been used to study transport phenomena in astrophysical environments [2], including neutrino transport in core-collapse supernovae (e.g., [3–7]) and binary neutron star mergers (e.g., [8,9]). The numerical method is based on the discontinuous Galerkin (DG) phase-space discretization and an implicit-explicit (IMEX) method for time integration, and we pay particular attention to the preservation of certain physical bounds by the fully discrete scheme. The bound-preserving property is achieved by carefully considering the phase-space and temporal discretizations, as well as the formulation of associated iterative nonlinear solvers.

Neutral particle transport in physical systems where the particle mean-free path may be similar to, or exceed, other characteristic length scales demands a kinetic description based on the distribution function $f(\boldsymbol{p},\boldsymbol{x},t)$, which is a phase-space density providing, at time t, the number of particles in an infinitesimal phase-space volume $d\boldsymbol{x}d\boldsymbol{p}$ centered around phase-space coordinates $\{\boldsymbol{p},\boldsymbol{x}\}$. Here, $\boldsymbol{p}$ and $\boldsymbol{x}$ are momentum- and position-space coordinates, respectively. The evolution of f is governed by a kinetic equation that expresses a balance between phase-space advection and collisions (e.g., interparticle collisions and/or collisions with a background); see, e.g., [10,2] for detailed expositions. In this paper, as a simplification, we consider the situation where particles described by a kinetic distribution function interact with an external background whose properties are prescribed and unaffected by f.

The design of numerical methods to model transport of particles interacting with a moving fluid is complicated, in part, by the necessity to choose coordinates for discretization of momentum space. While relativistic kinetic theory provides the framework to freely specify momentum-space coordinates, the two most obvious reference frame choices, the Eulerian and comoving frames, come with distinct computational challenges (e.g., [11–13,2]). On the one hand, choosing momentum-space coordinates associated with an Eulerian observer eases the discretization of the phase-space advection problem at the expense of complicating the particle-fluid interaction kinematics and, for moment models, the closure procedure. On the other hand, choosing momentum-space coordinates associated with the comoving frame (or comoving observer) — defined as the sequence of inertial frames whose velocity instantaneously coincides with the fluid velocity [13,2] — simplifies the description of particle-fluid interaction kinematics but at the expense of increased complexity in solving the phase-space advection problem numerically. Moreover, when particles equilibrate with the fluid, the distribution function becomes isotropic in the comoving frame, which simplifies the closure procedure for moment-based methods [13]. We also mention the mixed-frame approach (e.g., [14]), where the distribution function depends on Eulerian-frame momentum coordinates. Then, to evaluate comoving-frame emissivities and opacities at Eulerian-frame momentum coordinates, appropriate transformation laws and expansions to $O(v/c)$ are applied (see Section 7.2 in [2]). The mixed-frame approach attempts to combine the best of both coordinate choices but has difficulties with certain collision operators and does not generalize to the relativistic case. Nagakura et al. [15] combine both coordinate choices in a relativistic framework, using a discrete ordinates method, which requires mapping of numerical data between momentum space coordinate systems. (Recently, this approach has been adopted also to thermal radiation transport [16].) This approach has yet to be applied to moment models.

Our primary goal is to model neutrino transport in large-scale core-collapse supernova simulations, which require the inclusion of a wide range of neutrino–matter interactions — with various kinematic forms (e.g., [17–19,7,20]) — which tend to dominate the overall computational cost. Therefore, we opt for relative simplicity in the collision term, adopt momentum-space coordinates associated with the comoving frame, and focus our effort here on the discretization of the phase-space advection problem.

Because of the high computational cost associated with solving kinetic equations numerically in full dimensionality with sufficient phase-space resolution, dimension-reduction techniques are frequently employed. One commonly used method is to define and solve for a sequence of moments, instead of f directly. Specifically, we employ spherical-polar momentum-space coordinates $(\varepsilon,\vartheta,\varphi)$ and integrate the distribution function against angular basis functions (depending on momentum-space angles $\omega=(\vartheta,\varphi)$) to obtain spectral, angular moments (depending on particle energy ε, and $\boldsymbol{x}$ and t) representing number densities, number fluxes, etc. The hierarchy of moment equations is obtained by taking corresponding moments of the kinetic equation. In this study, we consider a so-called two-moment model, where we solve for the zeroth (scalar) and first (vector) moments. The resulting system of moment equations, accurate to $O(v/c)$, describes the evolution of the moments due to advection in phase-space (the left-hand side) and collisions with the background fluid (the right-hand side). Due to the choice of comoving-frame momentum coordinates, the left-hand side contains velocity-dependent terms that account for spatial advection, Doppler shift, and angular aberration. Moreover, the moment equations contain higher-order moments (rank-two and rank-three tensors) that must be expressed in terms of the lower-order moments to close the system of equations. Specifically, we consider an approximate, algebraic moment closure originating from the maximum-entropy closure proposed by Minerbo [21] (see also [22,4]). Related two-moment models have recently been used to model neutrino transport in core-collapse supernova simulations (e.g., [4,5]).

In this paper, we consider a number-conservative two-moment model obtained by taking the flat spacetime, $O(v/c)$ limit of general-relativistic moment models, e.g., from [23,24,7]. We refer to the model as number-conservative because, in the absence of collisions, the zeroth moment equation is conservative for the correct $O(v/c)$ Eulerian-frame number density. The model is closely related to the two-moment model promoted by Lowrie et al. [1]: With the assumption of one-dimensional, planar geometry, we obtain their equations by multiplying our equations with the particle energy ε. This two-moment model supports wave speeds that are bounded by the speed of light. It is also consistent, to $O(v/c)$, with conservation laws for Eulerian-frame energy and momentum. *Key to this consistency is retention of certain $O(v/c)$ terms in the time derivative of the moment equations*, which are often omitted (e.g., [25,4,5]). However, retention of these terms increases the computational complexity of the algorithm because the evolved moments

become nonlinear functions of the primitive (comoving-frame) moments needed to evaluate closure relations, which then introduces nonlinear, iterative solves that contribute to increased computational costs.

We use the DG method [26] to discretize the moment equations. The choice of comoving-frame momentum coordinates results in advection-type terms along the energy dimension and four-dimensional divergence operators in the left-hand side of the moment equations. We use the DG method to discretize all four phase-space dimensions. DG methods have advantages for modeling particle transport because of their ability to capture the asymptotic diffusion limit with coarse meshes [27–29] without modification of numerical fluxes (as in, e.g., [30]), and we leverage this property here. Moreover, their variational formulation and flexibility with respect to test functions make them suitable for designing methods that conserve particle number and total energy *simultaneously* (e.g., [31,32]), which can be more difficult to achieve with, e.g., finite-difference or finite-volume methods. We use IMEX time stepping [33,34] to integrate the ordinary differential equations resulting from the semi-discretization of the moment equations by the DG method. Following our prior works [35,36], we integrate the phase-space advection problem explicitly and the collision term implicitly. However, different from our prior works, due to the additional $O(v/c)$ terms in the time derivatives of the moment equations, the implicit part is nonlinear, even for the simplified collision term we consider here, and requires an iterative solution procedure, which we formulate in this paper.

Given appropriate initial and boundary conditions, the solution to moment models with maximum-entropy closure is known to be *realizable*; i.e., the moment solution is consistent with a kinetic distribution f that satisfies required physical bounds [37,38]. For particle systems obeying Bose–Einstein or Maxwell–Boltzmann statistics, f is nonnegative, whereas for particle systems obeying Fermi–Dirac statistics, $f \in [0, 1]$. These bounds translate into constraints on the associated moments, and moments satisfying these constraints are referred to as "realizable" moments. Although moment realizability is preserved by continuous moment models, solving moment models numerically can result in unrealizable moments, which leads to ill-posedness of the closure procedure and can give unphysical results when coupling moment models to other physical models, such as fluid models. Therefore, maintaining moment realizability has been a key challenge in the design of numerical schemes for solving moment equations and has been explored in existing work from different perspectives, including development of realizability-preserving spatio-temporal discretizations [39,40, 35], design of realizability-enforcing limiters [35], and relaxation of the realizability constraints via regularization [38]. While these existing approaches provide some essential components to construct a realizability-preserving scheme for the $O(v/c)$ two-moment model considered in this work, they focus on models without any relativistic corrections and do not fully address the challenges of preserving moment realizability when relativistic corrections are included.

The realizability-preserving numerical scheme proposed in this paper consists of the following key components. First, for time integration, we adopt a strong stability-preserving (SSP) IMEX method, which treats the advection terms explicitly and the collision term implicitly. This choice avoids excessive time-step restrictions in the highly collisional regime and gives explicit stage updates that can be expressed as a convex combination of multiple forward Euler steps, which is necessary for preserving realizability. Second, the DG method is equipped with tailored numerical fluxes, which, together with the SSP IMEX time integration method, maintains nonnegative cell-averaged number densities in the explicit update under a time-step restriction that takes the form of a hyperbolic-type Courant–Friedrichs–Lewy (CFL) condition. Third, the realizability-enforcing limiter proposed in [35] is used to recover pointwise realizable moments after each stage of the IMEX method. As discussed above, the moment closure procedure requires an iterative solver for nonlinear equations that convert evolved (conserved) moments to the primitive moments. To preserve realizability in this conversion process, we formulate the nonlinear equation as a fixed-point problem and apply an iterative solver analogous to the modified Richardson iteration (e.g., [41,42]) to ensure realizability in each iteration. We prove the global convergence property of this iterative solver in the $O(v/c)$ regime. The convergence analysis is applicable to the maximum-entropy closure as well as its algebraic approximation. Finally, the nonlinear systems arising from the implicit step of the IMEX method can also be formulated as a fixed-point problem and solved in a similar fashion. The realizability-preserving and convergence analyses both carry through with minor modifications. With these components in hand, we prove that the proposed DG-IMEX scheme for solving the $O(v/c)$ two-moment model indeed preserves moment realizability.

The two-moment model we consider is number conservative and, in the continuum limit, consistent to $O(v/c)$ with phase-space conservation laws for Eulerian-frame energy and momentum. Because the Eulerian-frame energy is not a primary evolved quantity of the model, but is instead obtained from a nontrivial combination of the evolved quantities, similar consistency with this conservation law is not guaranteed at the discrete level. In the context of finite-difference methods, Liebendörfer et al. [43] proposed a consistent discretization by carefully matching specific numerical flux terms in the finite-difference representation of the general-relativistic Boltzmann equation (see also [44] for an approach in the case of moment models). For the semi-discrete DG scheme proposed here, the numerical fluxes are tailored to maintain moment realizability, which limits the flexibility of following this procedure. However, the flexibility provided by the approximation spaces of the DG method can be helpful in this respect. For example, by testing with the particle energy ε, which is represented exactly by the DG approximation space with linear functions in the energy dimension, we obtain the two-moment model promoted in [1]. We further analyze the *simultaneous* Eulerian-frame number and energy conservation properties of the semi-discrete DG scheme, and point out that our DG approximation of the background velocity, which is allowed to be discontinuous, can impact the ability to achieve consistency with Eulerian-frame energy conservation to $O(v/c)$. Moreover, we design a "spectral redistribution" that corrects for Eulerian-frame energy conservation violations introduced by the realizability-enforcing limiter mentioned above. Through numerical experiments, we observe that Eulerian-frame energy conservation violations grow as $(v/c)^2$, indicating the desired consistency for an $O(v/c)$ method.

The paper is organized as follows. The mathematical formulation of the two-moment model is presented in Section 2, while the closure procedure and wave propagation speeds supported by the resulting moment model are presented and discussed in Section 3. Section 4 provides an overview of the numerical method, including the DG phase-space discretization, IMEX time discretiza-

tion, and iterative solvers for the nonlinear systems arising from the conserved-to-primitive conversion problem and time-implicit evaluation of the collision term. Section 5, where the realizability-preserving property of the method is established, contains the main technical results of the paper. The simultaneous conservation of Eulerian-frame number and energy of the DG method is discussed in Section 6, where the spectral redistribution that corrects for Eulerian-frame energy conservation violations introduced by the realizability-enforcing limiter is also presented. The algorithms have been implemented in the toolkit for high-order neutrino radiation-hydrodynamics (THORNADO[1]) and have been ported to utilize graphics processing units (GPUs). Our GPU programming model and implementation strategy is briefly discussed in Section 7. Results from numerical experiments demonstrating the robustness and accuracy of our method are presented in Section 8, where we also present GPU and multi-core performance results and highlight the relative computational cost of algorithmic components. The kinetic equation from which our two-moment model can be easily derived is provided in Appendix A. Some technical proofs are given in Appendix B.

For the remainder of this paper we employ units in which the speed of light is unity ($c = 1$).

2. Mathematical model

We consider a kinetic model where we solve for angular moments of the distribution function $f : (\omega, \varepsilon, \boldsymbol{x}, t) \in \mathbb{S}^2 \times \mathbb{R}^+ \times \mathbb{R}^3 \times \mathbb{R}^+ \to \mathbb{R}^+$, which gives the number of particles propagating in the direction $\omega \in \mathbb{S}^2 := \{ \omega = (\vartheta, \varphi) \mid \vartheta \in [0, \pi], \varphi \in [0, 2\pi) \}$, with energy $\varepsilon \in \mathbb{R}^+$, at position $\boldsymbol{x} \in \mathbb{R}^3$ and time $t \in \mathbb{R}^+$. We define angular moments of f as

$$\left\{ \mathcal{D}, \mathcal{I}^i, \mathcal{K}^{ij}, \mathcal{Q}^{ijk} \right\}(\varepsilon, \boldsymbol{x}, t) = \frac{1}{4\pi} \int_{\mathbb{S}^2} f(\omega, \varepsilon, \boldsymbol{x}, t) \left\{ 1, \ell^i, \ell^i \ell^j, \ell^i \ell^j \ell^k \right\} d\omega, \tag{1}$$

where $\ell^i(\omega)$ is the ith component of a unit vector parallel to the particle three-momentum $\boldsymbol{p} = \varepsilon \boldsymbol{\ell}$, and $d\omega = \sin\vartheta \, d\vartheta \, d\varphi$. We take $\boldsymbol{p} = \left(p^1, p^2, p^3\right)^{\mathsf{T}}$ to be the particle three-momentum, and ε and ω the particle energy and direction in a spherical-polar momentum-space coordinate system associated with an observer instantaneously moving with the fluid three-velocity $\boldsymbol{v}$ (the comoving observer). This choice of momentum-space coordinates is commonly used to model particles interacting with a moving material, as it simplifies the particle–material interaction (collision) terms (see, e.g., [12,2]). For simplicity, we will assume that the components of the three-velocity v^i are given functions of position $\boldsymbol{x}$, *independent* of time t. In Eq. (1), $\mathcal{D}$ and $\mathcal{I}^i$ are the comoving-frame, spectral particle density and flux density components, respectively.

Moment models that incorporate moving fluid effects are derived in the framework of relativistic kinetic theory [45], and the moment model considered here is obtained from taking angular moments of the $O(v)$ kinetic equation derived by Munier & Weaver [46, 47], which we provide for convenience in Appendix A. In this limit, the zeroth-moment equation is given by

$$\partial_t \left(\mathcal{D} + v^i \mathcal{I}_i \right) + \partial_i \left(\mathcal{I}^i + v^i \mathcal{D} \right) - \frac{1}{\varepsilon^2} \partial_\varepsilon \left(\varepsilon^3 \mathcal{K}^i{}_k \, \partial_i v^k \right) = \chi \left(\mathcal{D}_0 - \mathcal{D} \right), \tag{2}$$

where $\partial_t = \partial/\partial t$, $\partial_i = \partial/\partial x^i$, and $\partial_\varepsilon = \partial/\partial\varepsilon$. We use the Einstein summation convention, where repeated Latin indices run from 1 to 3. In flat spacetime, assuming Cartesian spatial coordinates, we can raise and lower indices on vectors and tensors with the Kronecker tensor; e.g., $\mathcal{I}_i = \delta_{ij} \mathcal{I}^j$. On the right-hand side of Eq. (2), $\chi \geq 0$ is the absorption opacity, and $\mathcal{D}_0$ is the zeroth moment of an equilibrium distribution f_0. The corresponding first-moment equation is given by

$$\begin{aligned} &\partial_t \left(\mathcal{I}_j + v^i \mathcal{K}_{ij} \right) + \partial_i \left(\mathcal{K}^i{}_j + v^i \mathcal{I}_j \right) - \frac{1}{\varepsilon^2} \partial_\varepsilon \left(\varepsilon^3 \mathcal{Q}^i{}_{kj} \, \partial_i v^k \right) \\ &\quad + \mathcal{I}^i \, \partial_i v_j - \mathcal{Q}^i{}_{kj} \, \partial_i v^k = -\kappa \mathcal{I}_j, \end{aligned} \tag{3}$$

where $\kappa = \chi + \sigma$ is the sum of the absorption opacity and the opacity due to elastic and isotropic scattering ($\sigma \geq 0$).

The two-moment model given by Eqs. (2) and (3) correspond to the moment equations for number transport given by Just et al. [4]; their Equations (9a) and (9b). (See also Eq. (125) in [48] for the number-density equation.) The velocity-dependent terms in the spatial and energy derivatives in Eqs. (2) and (3) account for spatial advection and Doppler shift between adjacent comoving observers, respectively, while the fourth and fifth terms on the left-hand side of Eq. (3) account for angular aberration between adjacent comoving observers (e.g., [43]). We point out that the velocity-dependent terms inside the time derivatives in Eqs. (2) and (3) were dropped in [4]. By retaining these terms, Eq. (2) evolves the $O(v)$ Eulerian-frame number density, and, as emphasized by Lowrie et al. [1], wave speeds remain bounded by the speed of light and the model is consistent with the correct $O(v)$ Eulerian-frame energy and momentum equations. To elaborate on the latter, we define the "conserved" moments that are evolved in Eqs. (2) and (3) as

$$\mathcal{N} := \mathcal{D} + v^i \mathcal{I}_i \quad \text{and} \quad \mathcal{G}_j := \mathcal{I}_j + v^i \mathcal{K}_{ij}, \tag{4}$$

respectively. Here, $\mathcal{N}$ is the correct $O(v)$ Eulerian-frame number density, and, in the absence of sources on the right-hand side, Eq. (2) is a phase-space conservation law. The Eulerian-frame energy and momentum densities are related to $\mathcal{N}$ and $\mathcal{G}_j$ by

$$\mathcal{E} = \varepsilon \left(\mathcal{N} + v^i \mathcal{G}_i \right) = \varepsilon \left(\mathcal{D} + 2 v^i \mathcal{I}_i \right) + O(v^2) \tag{5}$$

[1] www.github.com/endeve/thornado.

and

$$\mathcal{P}_j = \varepsilon\,(\mathcal{G}_j + v_j\,\mathcal{N}) = \varepsilon\,(\mathcal{I}_j + v^i\,\mathcal{K}_{ij} + v_j\,\mathcal{D}) + O(v^2), \tag{6}$$

respectively. The following proposition gives the energy and momentum conservation properties of the two-moment model in Eqs. (2)–(3).

Proposition 1. *The two-moment model given by Eqs. (2)–(3) is, up to $O(v)$, consistent with phase-space conservation laws for the energy density $\mathcal{E}$ and momentum density $\mathcal{P}_j$.*

Proof. By multiplying Eqs. (2) and (3) with appropriate factors and summing up the resulting equations, the evolution equations for the energy and momentum densities can be derived, respectively, as

$$\partial_t \mathcal{E} + \partial_i \mathcal{P}^i - \frac{1}{\varepsilon^2}\partial_\varepsilon\left(\varepsilon^4\,\mathcal{K}^i{}_k\,\partial_i v^k\right) = \varepsilon\,\chi\left(\mathcal{D}_0 - \mathcal{D}\right) - \varepsilon\,\kappa\,v^j\,\mathcal{I}_j \tag{7}$$

and

$$\partial_t \mathcal{P}_j + \partial_i \mathcal{S}^i{}_j - \frac{1}{\varepsilon^2}\partial_\varepsilon\left(\varepsilon^4\,\mathcal{Q}^i{}_{kj}\,\partial_i v^k\right) = -\varepsilon\,\kappa\,\mathcal{I}_j + \varepsilon\,v_j\,\chi\left(\mathcal{D}_0 - \mathcal{D}\right). \tag{8}$$

Here, all $O(v^2)$ terms are dropped, and the momentum flux density is denoted as $\mathcal{S}^{ij} := \varepsilon\,(\mathcal{K}^{ij} + \mathcal{I}^i\,v^j + v^i\,\mathcal{I}^j)$. In the absence of sources on the right-hand side, Eqs. (7) and (8) become phase-space conservation laws for $\mathcal{E}$ and $\mathcal{P}_j$, respectively. □

To close the two-moment model (2)–(3), the higher-order moments $\mathcal{K}^{ij}$ and $\mathcal{Q}^{ijk}$ must be specified. We will use an algebraic closure, which we discuss in more detail in Section 3. To this end, we write the second-order moments as

$$\mathcal{K}^{ij} = \mathsf{k}^{ij}\,\mathcal{D}, \tag{9}$$

where the symmetric variable Eddington tensor components are given by (e.g., [49])

$$\mathsf{k}^{ij} = \frac{1}{2}\left[(1-\psi)\,\delta^{ij} + (3\psi - 1)\,\hat{\mathsf{n}}^i\,\hat{\mathsf{n}}^j\right], \tag{10}$$

where $\hat{\mathsf{n}}^i = \mathcal{I}^i/\mathcal{I}$ and $\mathcal{I} = \sqrt{\mathcal{I}_i\mathcal{I}^i}$. The expression given by Eq. (10) satisfies the trace condition $\mathsf{k}^i{}_i = \delta_{ij}\mathsf{k}^{ij} = 1$ (cf. Eq. (1)), and the Eddington factor can be obtained from

$$\psi = \hat{\mathsf{n}}_i\,\hat{\mathsf{n}}_j\,\mathsf{k}^{ij} = \frac{\int_{\mathbb{S}^2} f\,(\hat{\mathsf{n}}_i\ell^i)^2\,d\omega}{\int_{\mathbb{S}^2} f\,d\omega}. \tag{11}$$

Similarly, the third-order moments can be written as

$$\mathcal{Q}^{ijk} = \mathsf{q}^{ijk}\,\mathcal{D}, \tag{12}$$

where we define the symmetric "heat-flux" tensor (e.g., [4]),

$$\mathsf{q}^{ijk} = \frac{1}{2}\left[(h-\zeta)\left(\hat{\mathsf{n}}^i\,\delta^{jk} + \hat{\mathsf{n}}^j\,\delta^{ik} + \hat{\mathsf{n}}^k\,\delta^{ij}\right) + (5\zeta - 3h)\,\hat{\mathsf{n}}^i\,\hat{\mathsf{n}}^j\,\hat{\mathsf{n}}^k\right], \tag{13}$$

where $h = \mathcal{I}/\mathcal{D}$ is the flux factor. The expression in Eq. (13) satisfies the trace condition $\delta_{jk}\,\mathsf{q}^{ijk} = \mathsf{q}^{ij}{}_j = \mathcal{I}^i/\mathcal{D}$, and the "heat-flux" factor can be obtained from

$$\zeta = \hat{\mathsf{n}}_i\,\hat{\mathsf{n}}_j\,\hat{\mathsf{n}}_k\,\mathsf{q}^{ijk} = \frac{\int_{\mathbb{S}^2} f\,(\hat{\mathsf{n}}_i\ell^i)^3\,d\omega}{\int_{\mathbb{S}^2} f\,d\omega}. \tag{14}$$

Eqs (2) and (3) are closed by specifying the Eddington and heat-flux factors in terms of the "primitive" moments $\boldsymbol{\mathcal{M}} = \left(\mathcal{D}, \boldsymbol{\mathcal{I}}\right)^{\mathsf{T}}$; i.e., $\psi = \psi(\boldsymbol{\mathcal{M}})$ and $\zeta = \zeta(\boldsymbol{\mathcal{M}})$.

Assuming a closure for the higher-order tensors, we define the vector of evolved moments,

$$\boldsymbol{\mathcal{U}}(\boldsymbol{\mathcal{M}}, \boldsymbol{v}) = \begin{bmatrix} \mathcal{N} \\ \mathcal{G}_j \end{bmatrix} = \begin{bmatrix} \mathcal{D} + v^i\,\mathcal{I}_i \\ \mathcal{I}_j + v^i\,\mathcal{K}_{ij} \end{bmatrix}, \tag{15}$$

the phase-space fluxes,

$$\boldsymbol{\mathcal{F}}^i(\boldsymbol{\mathcal{U}}, \boldsymbol{v}) = \begin{bmatrix} \mathcal{I}^i + v^i\,\mathcal{D} \\ \mathcal{K}^i{}_j + v^i\,\mathcal{I}_j \end{bmatrix} \quad \text{and} \quad \boldsymbol{\mathcal{F}}^\epsilon(\boldsymbol{\mathcal{U}}, \boldsymbol{v}) = -\begin{bmatrix} \mathcal{K}^i{}_k \\ \mathcal{Q}^i{}_{kj} \end{bmatrix}\partial_i v^k, \tag{16}$$

and the sources,

$$\boldsymbol{\mathcal{S}}(\boldsymbol{\mathcal{U}}, \boldsymbol{v}) = \begin{bmatrix} 0 \\ \mathcal{Q}^i{}_{kj}\,\partial_i v^k - \mathcal{I}^i\,\partial_i v_j \end{bmatrix} \quad \text{and} \quad \boldsymbol{\mathcal{C}}(\boldsymbol{\mathcal{U}}) = \begin{bmatrix} \chi\left(\mathcal{D}_0 - \mathcal{D}\right) \\ -\kappa\,\mathcal{I}_j \end{bmatrix}, \tag{17}$$

so we can write the two-moment model in the compact form,

$$\partial_t \boldsymbol{\mathcal{U}} + \frac{\partial}{\partial x^i}\left(\boldsymbol{\mathcal{F}}^i(\boldsymbol{\mathcal{U}},\boldsymbol{v})\right) + \frac{1}{\varepsilon^2}\frac{\partial}{\partial \varepsilon}\left(\varepsilon^3 \boldsymbol{\mathcal{F}}^\varepsilon(\boldsymbol{\mathcal{U}},\boldsymbol{v})\right) = \boldsymbol{\mathcal{S}}(\boldsymbol{\mathcal{U}},\boldsymbol{v}) + \boldsymbol{\mathcal{C}}(\boldsymbol{\mathcal{U}}). \tag{18}$$

Note that the collision term $\boldsymbol{\mathcal{C}}$ does not depend explicitly on the three-velocity $\boldsymbol{v}$. This is a consequence of choosing comoving-frame, momentum-space coordinates.

The moment closure is defined in terms of the primitive moments $\boldsymbol{\mathcal{M}}$, while we will evolve the "conserved" moments $\boldsymbol{\mathcal{U}} = \left(\mathcal{N}, \mathcal{G}_j\right)^\intercal$. The relation between the conserved and primitive moments can be written as

$$\boldsymbol{\mathcal{U}} = \boldsymbol{\mathcal{L}}(\boldsymbol{\mathcal{M}},\boldsymbol{v})\,\boldsymbol{\mathcal{M}}, \tag{19}$$

where

$$\boldsymbol{\mathcal{L}}(\boldsymbol{\mathcal{M}},\boldsymbol{v}) = \begin{bmatrix} 1 & v^1 & v^2 & v^3 \\ v^k\,\mathsf{k}_{k1}(\boldsymbol{\mathcal{M}}) & 1 & 0 & 0 \\ v^k\,\mathsf{k}_{k2}(\boldsymbol{\mathcal{M}}) & 0 & 1 & 0 \\ v^k\,\mathsf{k}_{k3}(\boldsymbol{\mathcal{M}}) & 0 & 0 & 1 \end{bmatrix}. \tag{20}$$

When solving Eq. (18) numerically, it is necessary to convert between primitive and conserved moments. Computing the conserved moments from the primitive moments is straightforward, but obtaining the primitive moments from the conserved moments is non-trivial because, for a given nontrivial velocity $\boldsymbol{v}$, there is no closed-form expression for $\boldsymbol{\mathcal{M}}$ in terms of $\boldsymbol{\mathcal{U}}$, due to the nonlinear dependence $\mathsf{k}_{ij}(\boldsymbol{\mathcal{M}})$. Thus, the primitive moments must be obtained through an iterative procedure, which we discuss in more detail later, where we will pay particular attention to maintaining physically-realizable moments throughout the iteration process. One is faced with a similar problem, e.g., when solving the relativistic Euler and magnetohydrodynamics equations (e.g., [50]).

3. Moment closure

We use the maximum-entropy closure of Minerbo [21] to close the two-moment model. We let the admissible set of kinetic distribution functions be

$$\mathfrak{R} := \left\{ f \mid f \geq 0 \quad \text{and} \quad \frac{1}{4\pi}\int_{\mathbb{S}^2} f\, d\omega > 0 \right\}, \tag{21}$$

which is then used to define moment realizability as below.

Definition 1. The moments $\boldsymbol{\mathcal{M}} = (\mathcal{D}, \mathcal{I})^\intercal$ are realizable if they can be obtained from a distribution function $f(\omega) \in \mathfrak{R}$. The set of all realizable moments $\mathcal{R}$ is

$$\mathcal{R} := \left\{ \boldsymbol{\mathcal{M}} = (\mathcal{D},\mathcal{I})^\intercal \mid \mathcal{D} > 0 \text{ and } \gamma(\boldsymbol{\mathcal{M}}) = \mathcal{D} - \mathcal{I} \geq 0 \right\}, \tag{22}$$

where the function $\gamma(\boldsymbol{\mathcal{M}})$ is concave.

The Minerbo closure is based on the maximum-entropy principle, assuming an entropy functional of the form $s[f] = f\,\ln f - f$. The functional form of the distribution maximizing this entropy functional is, in this case, the Maxwell–Boltzmann distribution,

$$f_{\mathrm{ME}}(\omega) = \exp\left(\alpha + \beta\,(\hat{\mathsf{n}}_i \ell^i)\right), \tag{23}$$

where α and β are determined from the constraints,

$$\mathcal{D} = \frac{1}{4\pi}\int_{\mathbb{S}^2} f_{\mathrm{ME}}(\omega)\, d\omega \quad \text{and} \quad \hat{\mathsf{n}}_i\, \mathcal{I}^i = \mathcal{I} = \frac{1}{4\pi}\int_{\mathbb{S}^2} f_{\mathrm{ME}}(\omega)\,(\hat{\mathsf{n}}_i \ell^i)\, d\omega. \tag{24}$$

(Note that $f_{\mathrm{ME}} \in \mathfrak{R}$.) Letting $\hat{\mathsf{n}}_i \ell^i = \mu$, we can write f_{ME} as a function of μ and perform a change of variable to write the integrals in Eq. (24) in terms of μ, which allows us to evaluate the constraints in Eq. (24) analytically (cf. [21]) and leads to

$$\mathcal{D} = e^\alpha\, \sinh(\beta)/\beta \quad \text{and} \quad \mathcal{I} = e^\alpha\left(\beta\cosh(\beta) - \sinh(\beta)\right)/\beta^2. \tag{25}$$

The flux factor can then be written solely as a function of β; i.e., $h = \coth(\beta) - 1/\beta =: L(\beta)$, where $L(\beta)$ is the Langevin function. Thus, for a given h, we can obtain $\beta(h) = L^{-1}(h)$. Note that $L(\beta) \in (-1,1)$, so that solutions for β only exist for $h < 1$ (i.e., for $\boldsymbol{\mathcal{M}}$ in the interior of $\mathcal{R}$). Using the maximum-entropy distribution in Eq. (23), direct calculations give, for $h \in [0,1)$,

$$\psi(h) = 1 - \frac{2\,h}{\beta(h)} \quad \text{and} \quad \zeta(h) = \coth(\beta(h)) - 3\psi(h)/\beta(h). \tag{26}$$

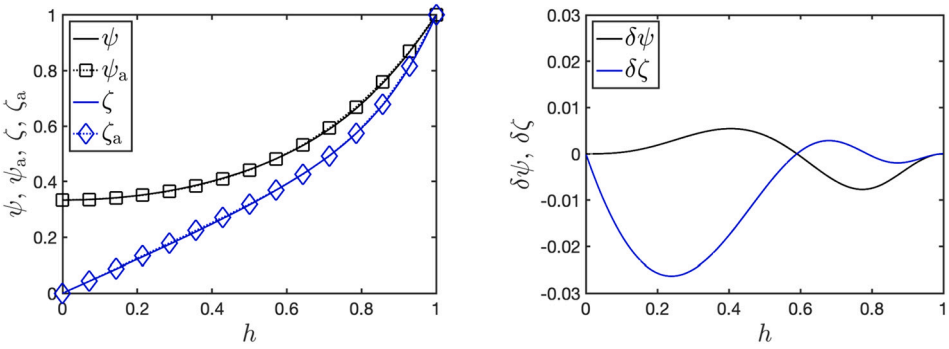


Fig. 1. The left plot shows the values of the Eddington factor, ψ, the heat-flux factor, ζ, and their polynomial approximations, ψ_a and ζ_a, versus the flux factor, h. The right plot illustrates the relative errors, $\delta\psi = (\psi - \psi_a)/\psi$ and $\delta\zeta = (\zeta - \zeta_a)/\zeta$, versus h.

When $h = 1$ (i.e., when $\mathcal{M}$ is on the boundary of $\mathcal{R}$), it is known [51] that, for the two-moment case considered here, the underlying kinetic distribution is a weighted Dirac delta function. In this case, the moment closure is given by the associated Eddington and heat-flux factors $\psi(1) = \zeta(1) = 1$. Instead of inverting the Langevin function for β, the Eddington and heat-flux factors, ψ and ζ, can be accurately approximated by polynomials in h. For ψ, the following polynomial approximation leads to a relative approximation error, $\delta\psi := (\psi - \psi_a)/\psi$, within 1% [22]:

$$\psi_a(h) = \frac{1}{3} + \frac{2}{15}\left(3h^2 - h^3 + 3h^4\right). \tag{27}$$

For ζ, the following approximation, given by [4],

$$\zeta_a(h) = h\left(45 + 10h - 12h^2 - 12h^3 + 38h^4 - 12h^5 + 18h^6\right)/75, \tag{28}$$

has a relative approximation error, $\delta\zeta := (\zeta - \zeta_a)/\zeta$, lower than 3%. In Fig. 1, we plot the Eddington factor, ψ, the heat-flux factor, ζ, and their polynomial approximations, ψ_a and ζ_a, and report the relative approximation error versus the flux factor, h. It can be seen from Fig. 1 that ψ_a and ζ_a are quite accurate polynomial approximations to the Eddington and heat-flux factors. Thus, the approximate closure is used in the numerical tests for the two-moment model reported in Section 8, in which the two-moment model is closed by plugging the algebraic expressions given in Eqs. (27) and (28) into the Eddington and heat-flux tensors in Eqs. (10) and (13), respectively.

Remark 1. In this work, we focus on the Minerbo closure, which is based on the Maxwell–Boltzmann entropy $s[f] = f\ln f - f$. The results can be extended to two-moment models with maximum-entropy closures based on the Bose–Einstein entropy $s[f] = f\ln f - (1+f)\ln(1+f)$ with maximum entropy distribution $f_{\mathrm{ME}} = [\exp(-\alpha - \beta(\hat{\mathsf{n}}_i \ell^i)) - 1]^{-1}$. For particle systems obeying Bose–Einstein or Maxwell–Boltzmann statistics, the admissible set $\mathfrak{R}$ and the realizable set $\mathcal{R}$ in this work are appropriate. For Fermi–Dirac particles, e.g., neutrinos, the entropy functional is $s[f] = f\ln f + (1-f)\ln(1-f)$, with bounds $0 \le f \le 1$ and maximum entropy distribution $f_{\mathrm{ME}} = [\exp(-\alpha - \beta(\hat{\mathsf{n}}_i \ell^i)) + 1]^{-1}$ [22]. In the low-occupancy limit, $f \ll 1$, this simplifies to the Maxwell–Boltzmann case considered here. However, the extension of this work to systems obeying Fermi–Dirac statistics, where f is also bounded from above, is non-trivial and deferred to future work.

Next we explore the wave propagation speeds of the moment system in Eq. (18) with the approximate Minerbo closure introduced above. To calculate the wave speed, we compute the maximum magnitude of the eigenvalues of the spatial-flux Jacobians with respect to the conserved moments, $(\partial_{\mathcal{U}}\mathcal{F}^i)$, $i = 1,2,3$. Specifically, we compute the spatial-flux Jacobian by

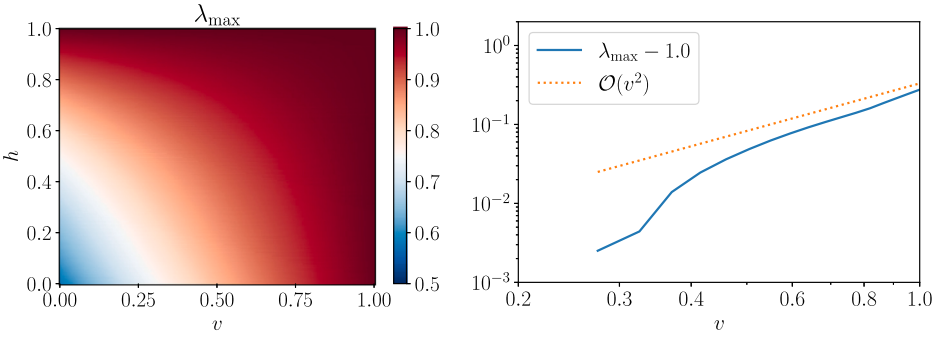
$$\left(\frac{\partial \mathcal{F}^i}{\partial \mathcal{U}}\right) = \left(\frac{\partial \mathcal{F}^i}{\partial \mathcal{M}}\right)\left(\frac{\partial \mathcal{U}}{\partial \mathcal{M}}\right)^{-1}, \tag{29}$$

where

$$\left(\frac{\partial \mathcal{U}}{\partial \mathcal{M}}\right)_{ij} = \begin{bmatrix} 1 & v^j \\ v^k\left[\left(\frac{\partial \mathsf{k}_{ik}}{\partial \mathcal{D}}\right)\mathcal{D} + \mathsf{k}_{ik}\right] & \delta_{ij} + v^k\left(\frac{\partial \mathsf{k}_{ik}}{\partial I^j}\right)\mathcal{D} \end{bmatrix} \tag{30}$$

and

$$\left(\frac{\partial \mathcal{F}^i}{\partial \mathcal{M}}\right) = \begin{bmatrix} v^i & \delta^{i1} & \delta^{i2} & \delta^{i3} \\ \left(\frac{\partial \mathsf{k}^i{}_1}{\partial \mathcal{D}}\right)\mathcal{D} + \mathsf{k}^i{}_1 & \left(\frac{\partial \mathsf{k}^i{}_1}{\partial I^1}\right)\mathcal{D} + v^i & \left(\frac{\partial \mathsf{k}^i{}_1}{\partial I^2}\right)\mathcal{D} & \left(\frac{\partial \mathsf{k}^i{}_1}{\partial I^3}\right)\mathcal{D} \\ \left(\frac{\partial \mathsf{k}^i{}_2}{\partial \mathcal{D}}\right)\mathcal{D} + \mathsf{k}^i{}_2 & \left(\frac{\partial \mathsf{k}^i{}_2}{\partial I^1}\right)\mathcal{D} & \left(\frac{\partial \mathsf{k}^i{}_2}{\partial I^2}\right)\mathcal{D} + v^i & \left(\frac{\partial \mathsf{k}^i{}_2}{\partial I^3}\right)\mathcal{D} \\ \left(\frac{\partial \mathsf{k}^i{}_3}{\partial \mathcal{D}}\right)\mathcal{D} + \mathsf{k}^i{}_3 & \left(\frac{\partial \mathsf{k}^i{}_3}{\partial I^1}\right)\mathcal{D} & \left(\frac{\partial \mathsf{k}^i{}_3}{\partial I^2}\right)\mathcal{D} & \left(\frac{\partial \mathsf{k}^i{}_3}{\partial I^3}\right)\mathcal{D} + v^i \end{bmatrix} \tag{31}$$



(a) One-dimensional case (b) Three-dimensional case

Fig. 2. Figures on both panels show the value of $\lambda_{\max}$, the maximum magnitude of the spatial-flux Jacobian eigenvalues in various configurations. Fig. 2a plots the computed values of $\lambda_{\max} = \max(\lambda(\partial_{\boldsymbol{u}}\mathcal{F}^1))$ at $v \in [0,1]$ and $h \in [0,1]$ in a simplified one-dimensional case considered in Proposition 2. The result verifies the claim $\lambda_{\max} \leq 1$ in Proposition 2. Fig. 2b shows that, in the three-dimensional case, the maximum wave speed of the two-moment model Eq. (18) scales as $1 + O(v^2)$. Here, the maximum wave speed is given by $\lambda_{\max} := \max_{i=1,2,3} \max_{\boldsymbol{u}\in\mathcal{R}}(\lambda^i_{\max})$, where $\lambda^i_{\max} = \max(\lambda(\partial_{\boldsymbol{u}}\mathcal{F}^i))$. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

follow from the definitions given in Eqs. (15) and (16), respectively. With this expression, we are able to demonstrate the following proposition, which states that the maximum wave speed is bounded above by the speed of light in a one-dimensional setting.

Proposition 2. *Suppose* $\boldsymbol{v} = (v,0,0)$ *and* $\mathcal{I} = (\mathcal{I},0,0)$*, with* $|v| \leq 1$*,* $|\mathcal{I}| \leq \mathcal{D}$*, and* $\mathcal{D} > 0$*. Let* $\lambda_{\max} := \max\left(|\lambda(\partial_{\boldsymbol{u}}\mathcal{F}^1)|\right)$ *denote the maximum magnitude of the spatial-flux Jacobian eigenvalues. Then* $\lambda_{\max} \leq 1$*.*

Proof. In this setting, the spatial-flux Jacobian reduces to a 2-by-2 matrix, because the entries associated with the x^2 and x^3 axes are all zeros. In addition, the only nonzero component of the Eddington tensor is k_{11}, which takes the values of the (approximate) Eddington factor ψ_{a}. Thus, the partial derivatives $\frac{\partial \mathsf{k}_{11}}{\partial \mathcal{D}}$ and $\frac{\partial \mathsf{k}^1_{\ 1}}{\partial \mathcal{I}^1}$ become $\frac{\partial \psi_{\mathsf{a}}}{\partial \mathcal{D}}$ and $\frac{\partial \psi_{\mathsf{a}}}{\partial \mathcal{I}^1}$, respectively. Evaluating these partial derivatives using the chain rule then leads to

$$\left(\frac{\partial \mathcal{F}^1}{\partial \boldsymbol{u}}\right) = \frac{1}{1 - v^2\psi_{\mathsf{a}} + v(1+vh)\psi'_{\mathsf{a}}} \begin{bmatrix} v - v\psi_{\mathsf{a}} + v(v+h)\psi'_{\mathsf{a}} & 1 - v^2 \\ (1-v^2)(\psi_{\mathsf{a}} - h\psi'_{\mathsf{a}}) & v - v\psi_{\mathsf{a}} + (1+vh)\psi'_{\mathsf{a}} \end{bmatrix}, \tag{32}$$

where ψ'_{a} denotes the derivative of the approximate Eddington factor, ψ_{a}, in Eq. (27) with respect to the flux factor, h. To prove the claim, we need to show that the eigenvalues of $(\partial_{\boldsymbol{u}}\mathcal{F}^1)$ are in $[-1,1]$. Since ψ_{a} and ψ'_{a} are both one-dimensional polynomials in h, the proof of the claim is straightforward but tedious. Here we omit the detailed analysis and show in Fig. 2a the computed values of $\lambda_{\max}$ for $v \in [0,1]$ and $h \in [0,1]$, which illustrates that $\lambda_{\max}$ is bounded from above by one. □

This result is an extension of the wave speed analysis in [1, Section 6.2], in which it is assumed that the Eddington factor is independent of the flux factor; i.e., that $\psi'_{\mathsf{a}} = 0$.

Remark 2. In the three-dimensional case, the magnitude of the eigenvalues of the spatial-flux Jacobian are bounded above by $1 + O(v^2)$, which we provide verification of in Fig. 2b. Although the upper bound can exceed unity, which implies that the wave speed of the two-moment model in Eq. (18) can become unphysical, it shows that including the velocity-dependent term in the time derivatives in Eqs. (2) and (3) improves the maximum wave speed estimation from $1 + O(v)$ (see, e.g., discussions in [1]) to $1 + O(v^2)$. Note that in the design of the numerical flux discussed in Section 4.1, we use unity as the estimate for the maximum wave speed, which appears to be valid in the regimes for which the $O(v)$ model is applicable. In particular, unphysical wave speeds are not observed for $v \leq 0.25$, as shown in Fig. 2b, for which we do not yet have a theoretical explanation.

4. Numerical scheme

4.1. Discontinuous Galerkin phase-space discretization

We use the DG method to discretize Eq. (18) in phase-space. To this end we divide the phase-space domain $D = D_\varepsilon \times D_{\boldsymbol{x}}$ into a disjoint union $\mathcal{T}$ of open elements $\boldsymbol{K} = K_\varepsilon \times \boldsymbol{K}_{\boldsymbol{x}}$, so that $D = \cup_{\boldsymbol{K}\in\mathcal{T}} \boldsymbol{K}$. Here, D_ε is the energy domain and $D_{\boldsymbol{x}}$ is the $d_{\boldsymbol{x}}$-dimensional spatial domain, and

$$\boldsymbol{K}_{\boldsymbol{x}} = \left\{ \boldsymbol{x} : \ x^i \in K^i_{\boldsymbol{x}} := (x^i_{\mathrm{L}}, x^i_{\mathrm{H}}) \mid i = 1, \ldots, d_{\boldsymbol{x}} \right\} \quad \text{and} \quad K_\varepsilon := (\varepsilon_{\mathrm{L}}, \varepsilon_{\mathrm{H}}), \tag{33}$$

where x^i_{L} (x^i_{H}) is the low (high) boundary of the spatial element in the ith spatial dimension, and ε_{L} (ε_{H}) is the low (high) boundary of the energy element. We also define $\tau(\varepsilon) = \varepsilon^2$ and denote the volume of a phase-space element by

$$|\boldsymbol{K}| = \int_{\boldsymbol{K}} \tau \, d\varepsilon \, d\boldsymbol{x}, \quad \text{where} \quad d\boldsymbol{x} = \prod_{i=1}^{d_{\boldsymbol{x}}} dx^i. \tag{34}$$

The length of an element in the ith dimension is $|K^i_{\boldsymbol{x}}| = x^i_{\mathrm{H}} - x^i_{\mathrm{L}}$, and $|K_\varepsilon| = \varepsilon_{\mathrm{H}} - \varepsilon_{\mathrm{L}}$. We also define the phase-space surface element $\tilde{\boldsymbol{K}}^i = (\times_{j\neq i} K^j_{\boldsymbol{x}}) \times K_\varepsilon$ and the spatial coordinates orthogonal to the ith spatial dimension $\tilde{\boldsymbol{x}}^i$, so that as a set $\boldsymbol{x} = \{ x^i, \tilde{\boldsymbol{x}}^i \}$. Finally, we let $\boldsymbol{z} = (\varepsilon, \boldsymbol{x})$ denote the phase-space coordinate, and define $d\boldsymbol{z} = d\varepsilon d\boldsymbol{x}$, $d\tilde{\boldsymbol{z}}^i = d\varepsilon d\tilde{\boldsymbol{x}}^i$, and let, again as a set, $\tilde{\boldsymbol{z}}^i = \{\varepsilon, \tilde{\boldsymbol{x}}^i\}$.

On each element $\boldsymbol{K}$, we let the approximation space for the DG method be

$$\mathbb{V}^k_h(\boldsymbol{K}) = \left\{ \varphi_h : \varphi_h|_{\boldsymbol{K}} \in \mathbb{Q}^k(\boldsymbol{K}), \forall \boldsymbol{K} \in \mathcal{T} \right\}, \tag{35}$$

where $\mathbb{Q}^k(\boldsymbol{K})$ is the phase-space tensor product of one-dimensional polynomials of maximal degree k. We will denote the approximation space on spatial elements as $\mathbb{V}^k_h(\boldsymbol{K}_{\boldsymbol{x}})$, which is defined as in Eq. (35), where $\mathbb{Q}^k(\boldsymbol{K}_{\boldsymbol{x}})$ is the spatial tensor product of one-dimensional polynomials of maximal degree k. We will use $\mathbb{V}^k_h(\boldsymbol{K}_{\boldsymbol{x}})$ to approximate the fluid three-velocity $\boldsymbol{v} = (v^1, v^2, v^3)$, which will be assumed to be a given function of $\boldsymbol{x}$.

The semi-discrete DG problem is then to find $\mathcal{U}_h \in \mathbb{V}^k_h(\boldsymbol{K})$, which approximates $\mathcal{U}$ in Eq. (18), such that

$$\left(\partial_t \mathcal{U}_h, \varphi_h \right)_{\boldsymbol{K}} = \mathcal{B}_h\left(\mathcal{U}_h, \boldsymbol{v}_h, \varphi_h \right)_{\boldsymbol{K}} + \left(C(\mathcal{U}_h), \varphi_h \right)_{\boldsymbol{K}}, \tag{36}$$

for all test functions $\varphi_h \in \mathbb{V}^k_h(\boldsymbol{K})$, $\boldsymbol{v}_h \in \mathbb{V}^k_h(\boldsymbol{K}_{\boldsymbol{x}})$, and all $\boldsymbol{K} \in \mathcal{T}$. In Eq. (36), we have defined the inner product

$$\left(a_h, b_h \right)_{\boldsymbol{K}} = \int_{\boldsymbol{K}} a_h \, b_h \, \tau \, d\boldsymbol{z}, \quad a_h, b_h \in \mathbb{V}^k_h(\boldsymbol{K}) \tag{37}$$

and the phase-space advection operator

$$\mathcal{B}_h\left(\mathcal{U}_h, \boldsymbol{v}_h, \varphi_h \right)_{\boldsymbol{K}} = \mathcal{B}^{\boldsymbol{x}}_h\left(\mathcal{U}_h, \boldsymbol{v}_h, \varphi_h \right)_{\boldsymbol{K}} + \mathcal{B}^{\varepsilon}_h\left(\mathcal{U}_h, \boldsymbol{v}_h, \varphi_h \right)_{\boldsymbol{K}} + \left(\mathcal{S}(\mathcal{U}_h, \boldsymbol{v}_h), \varphi_h \right)_{\boldsymbol{K}}, \tag{38}$$

where the contribution from position space fluxes is

$$\begin{aligned} \mathcal{B}^{\boldsymbol{x}}_h\left(\mathcal{U}_h, \boldsymbol{v}_h, \varphi_h \right)_{\boldsymbol{K}} = &- \sum_{i=1}^{d_{\boldsymbol{x}}} \int_{\tilde{\boldsymbol{K}}^i} \left[\widehat{\mathcal{F}^i}\left(\mathcal{U}_h, \boldsymbol{v}_h \right) \varphi_h|_{x^i_{\mathrm{H}}} - \widehat{\mathcal{F}^i}\left(\mathcal{U}_h, \boldsymbol{v}_h \right) \varphi_h|_{x^i_{\mathrm{L}}} \right] \tau \, d\tilde{\boldsymbol{z}}^i \\ &+ \sum_{i=1}^{d_{\boldsymbol{x}}} \left(\mathcal{F}^i(\mathcal{U}_h, \boldsymbol{v}_h), \partial_i \varphi_h \right)_{\boldsymbol{K}} \end{aligned} \tag{39}$$

and the contribution from energy space fluxes is

$$\begin{aligned} \mathcal{B}^{\varepsilon}_h\left(\mathcal{U}_h, \boldsymbol{v}_h, \varphi_h \right)_{\boldsymbol{K}} = &- \int_{\boldsymbol{K}_{\boldsymbol{x}}} \left[\varepsilon^3 \widehat{\mathcal{F}^\varepsilon}\left(\mathcal{U}_h, \boldsymbol{v}_h \right) \varphi_h|_{\varepsilon_{\mathrm{H}}} - \varepsilon^3 \widehat{\mathcal{F}^\varepsilon}\left(\mathcal{U}_h, \boldsymbol{v}_h \right) \varphi_h|_{\varepsilon_{\mathrm{L}}} \right] d\boldsymbol{x} \\ &+ \left(\varepsilon \mathcal{F}^\varepsilon(\mathcal{U}_h, \boldsymbol{v}_h), \partial_\varepsilon \varphi_h \right)_{\boldsymbol{K}}. \end{aligned} \tag{40}$$

In Eq. (39), $\widehat{\mathcal{F}^i}\left(\mathcal{U}_h, \boldsymbol{v}_h \right)$ is a numerical flux approximating the flux on the surface $\tilde{\boldsymbol{K}}^i$, which is evaluated using the global Lax–Friedrichs (LF) flux

$$\widehat{\mathcal{F}^i}\left(\mathcal{U}_h, \boldsymbol{v}_h \right)|_{x^i} = \mathcal{F}^i_{\mathrm{LF}}\left(\mathcal{U}_h(x^{i,-}, \tilde{\boldsymbol{z}}^i), \mathcal{U}_h(x^{i,+}, \tilde{\boldsymbol{z}}^i), \hat{\boldsymbol{v}}(x^i, \tilde{\boldsymbol{x}}^i) \right), \tag{41}$$

where $x^{i,\mp} = \lim_{\delta \to 0^+} x^i \mp \delta$ and where we write the global LF flux function as

$$\mathcal{F}^i_{\mathrm{LF}}\left(\mathcal{U}_a, \mathcal{U}_b, \hat{\boldsymbol{v}} \right) = \frac{1}{2} \left(\mathcal{F}^i(\mathcal{U}_a, \hat{\boldsymbol{v}}) + \mathcal{F}^i(\mathcal{U}_b, \hat{\boldsymbol{v}}) - \alpha^i \left(\mathcal{U}_b[\hat{\boldsymbol{v}}^i] - \mathcal{U}_a[\hat{\boldsymbol{v}}^i] \right) \right), \tag{42}$$

where α^i is the largest (absolute) eigenvalue of the flux Jacobian $\partial \mathcal{F}^i / \partial \mathcal{U}$ over the entire domain, for which we simply set $\alpha^i = 1$.[2] The components of the fluid three-velocity at the element interface is computed as the average

$$\hat{\boldsymbol{v}}(x^i, \tilde{\boldsymbol{x}}^i) = \frac{1}{2} \left(\boldsymbol{v}_h(x^{i,-}, \tilde{\boldsymbol{x}}^i) + \boldsymbol{v}_h(x^{i,+}, \tilde{\boldsymbol{x}}^i) \right). \tag{43}$$

Note that the three-velocity components can be discontinuous across element interfaces.

Remark 3. In the flux function in Eq. (42), we have defined the dissipative term to be proportional to $(\mathcal{U}_b[\hat{\boldsymbol{v}}^i] - \mathcal{U}_a[\hat{\boldsymbol{v}}^i])$, where $\hat{\boldsymbol{v}}^i = \left(\delta^{i1} \hat{v}^1, \delta^{i2} \hat{v}^2, \delta^{i3} \hat{v}^3 \right)^{\mathsf{T}}$, as opposed to the standard LF flux where the dissipative term is proportional to $(\mathcal{U}_b[\hat{\boldsymbol{v}}] - \mathcal{U}_a[\hat{\boldsymbol{v}}])$. We

[2] With this choice, at the expense of potentially increased numerical dissipation when the flux factor is small (see Fig. 2a), computation of flux Jacobian eigenvalues are avoided, and the realizability analysis is simplified.

have found this to be necessary in order to improve the realizability-preserving property of the scheme in the multi-dimensional setting (see Section 5).

In order to compute the energy space fluxes $\mathcal{F}^\epsilon(\boldsymbol{\mathcal{U}}_h, \boldsymbol{v}_h)$ and the sources $\mathcal{S}(\boldsymbol{\mathcal{U}}_h, \boldsymbol{v}_h)$, we need to approximate spatial derivatives of the three-velocity components within elements. We denote the derivative of the ith velocity component with respect to x^j by $(\partial_j v^i)_h \in \mathbb{V}_h^k(\boldsymbol{K}_{\boldsymbol{x}})$, and compute this by demanding that

$$\int_{\boldsymbol{K}_{\boldsymbol{x}}} (\partial_j v^i)_h \, \varphi_h \, d\boldsymbol{x} = \int_{\tilde{\boldsymbol{K}}_{\boldsymbol{x}}^j} \left[\hat{v}^i \, \varphi_h|_{x_{\mathrm{H}}^j} - \hat{v}^i \, \varphi_h|_{x_{\mathrm{L}}^j} \right] d\tilde{\boldsymbol{x}}^j - \int_{\boldsymbol{K}_{\boldsymbol{x}}} v_h^i \, \partial_j \varphi_h \, d\boldsymbol{x} \tag{44}$$

holds for all $\varphi_h \in \mathbb{V}_h^k(\boldsymbol{K}_{\boldsymbol{x}})$ and all $\boldsymbol{K}_{\boldsymbol{x}}$, and where $\hat{v}^i(x^j, \tilde{\boldsymbol{x}}^j)$ is computed as in Eq. (43).

The energy space flux $\widehat{\mathcal{F}^\epsilon}\left(\boldsymbol{\mathcal{U}}_h, \boldsymbol{v}_h\right)$ in Eq. (40) is also computed using an LF-type flux

$$\widehat{\mathcal{F}^\epsilon}\left(\boldsymbol{\mathcal{U}}_h, \boldsymbol{v}_h\right)|_\varepsilon = \mathcal{F}_{\mathrm{LF}}^\varepsilon\left(\boldsymbol{\mathcal{U}}_h(\varepsilon^-, \boldsymbol{x}), \boldsymbol{\mathcal{U}}_h(\varepsilon^+, \boldsymbol{x}), \boldsymbol{v}_h(\boldsymbol{x})\right), \tag{45}$$

where $\varepsilon^\mp = \lim_{\delta \to 0^+} \varepsilon \mp \delta$, and we take the LF flux function to be given by

$$\mathcal{F}_{\mathrm{LF}}^\varepsilon\left(\boldsymbol{\mathcal{U}}_a, \boldsymbol{\mathcal{U}}_b, \boldsymbol{v}_h\right) = \frac{1}{2}\left(\mathcal{F}^\epsilon(\boldsymbol{\mathcal{U}}_a, \boldsymbol{v}_h) + \mathcal{F}^\epsilon(\boldsymbol{\mathcal{U}}_b, \boldsymbol{v}_h) - \alpha^\epsilon \left(\boldsymbol{\mathcal{M}}_b - \boldsymbol{\mathcal{M}}_a \right) \right), \tag{46}$$

where α^ϵ is an estimate of the largest absolute eigenvalue of the flux Jacobian $\partial \mathcal{F}^\epsilon / \partial \boldsymbol{\mathcal{U}}$. To estimate α^ϵ we consider the quadratic form

$$Q(\boldsymbol{v}_h) = (-\partial_j v^i)_h \, \ell_i \, \ell^j = \boldsymbol{\ell}^\intercal A(\boldsymbol{v}_h) \boldsymbol{\ell}, \quad \text{where} \quad A_{ij}(\boldsymbol{v}_h) = -\frac{1}{2}\left((\partial_i v^j)_h + (\partial_j v^i)_h \right). \tag{47}$$

It can be shown that $|Q(\boldsymbol{v}_h)| \leq \lambda_A$, where λ_A is the largest absolute eigenvalue of the matrix A. (Since A is symmetric, the eigenvalues are real.) Hence, we set $\alpha^\epsilon = \lambda_A$.

Remark 4. In the energy space flux function in Eq. (46), the numerical dissipation term is given in terms of the primitive moments $\boldsymbol{\mathcal{M}}$ rather than the conserved moments $\boldsymbol{\mathcal{U}}$. This choice is motivated by the realizability analysis in Section 5.1.2.

Remark 5. For simplicity we assume that the absorption and scattering opacity (χ and σ, respectively), appearing in the second term on the right-hand side of Eq. (36), are constant within each phase-space element $\boldsymbol{K}$.

In this work, we consider the nodal DG scheme (see, e.g., [52] for an overview), which writes $\boldsymbol{\mathcal{U}}_h \in \mathbb{V}_h^k(\boldsymbol{K})$ as an expansion of tensor products of one-dimensional Lagrange polynomials of degrees up to k in each element. As in [36], we use the $(k+1)$-point Legendre–Gauss (LG) quadrature points (see, e.g., [53]) as the interpolation points for the Lagrange polynomials. Following the standard practice (i.e., for Ritz–Galerkin), we choose the test functions φ_h to be identical to the trial functions, which are the tensor products of Lagrange polynomials used in the expansion of $\boldsymbol{\mathcal{U}}_h$, and evaluate the inner products $(\cdot,\cdot)_{\boldsymbol{K}}$ using the $(k+1)$-point LG quadrature rule. In the remainder of this paper, we denote the sets of the $(k+1)$-point LG quadrature points in an element $\boldsymbol{K}$ on K_ε and $K_{\boldsymbol{x}}^i$ by $S_\varepsilon^{\boldsymbol{K}} := \{\varepsilon_1, \ldots, \varepsilon_{k+1}\}$ and $S_i^{\boldsymbol{K}} := \{x_1^i, \ldots, x_{k+1}^i\}$, respectively. Then the set of local DG nodes in element $\boldsymbol{K}$ is denoted as

$$S_\otimes^{\boldsymbol{K}} := S_\varepsilon^{\boldsymbol{K}} \otimes \left(\bigotimes_{i=1}^{d_{\boldsymbol{x}}} S_i^{\boldsymbol{K}} \right). \tag{48}$$

With this notation, the semidiscretized Eq. (36) can then be written as

$$\partial_t \boldsymbol{\mathcal{U}}_{\boldsymbol{k}} = \mathbf{B}\left(\boldsymbol{\mathcal{U}}_h, \boldsymbol{v}_h \right)_{\boldsymbol{k}} + \mathbf{C}(\boldsymbol{\mathcal{U}}_{\boldsymbol{k}}), \quad \forall \boldsymbol{K} \in \mathcal{T}, \tag{49}$$

where $\mathbf{B}$ and $\mathbf{C}$ denote the advection and collision operators acting on the collection of nodal values $\boldsymbol{\mathcal{U}}_{\boldsymbol{k}}(t) := \{\boldsymbol{\mathcal{U}}_h(\varepsilon, \boldsymbol{x}, t) : (\varepsilon, \boldsymbol{x}) \in S_\otimes^{\boldsymbol{K}}\}$. Here the subscript $\boldsymbol{k}$ implies evaluations at points in $S_\otimes^{\boldsymbol{K}}$. This nodal representation will become useful in the following sections. To simplify the notations therein, we will introduce a few auxiliary point sets in phase-space, which become useful in the realizability analysis in Sections 5.1.1 and 5.1.2. In element $\boldsymbol{K}$, let $\hat{S}_\varepsilon^{\boldsymbol{K}} := \{\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_{\hat{k}}\}$ and $\hat{S}_i^{\boldsymbol{K}} := \{\hat{x}_1^i, \ldots, \hat{x}_{\hat{k}}^i\}$ denote the sets of quadrature points given by the $\hat{k}$-point Legendre–Gauss–Lobatto (LGL) quadrature rule (see, e.g., [53]) on K_ε and $K_{\boldsymbol{x}}^i$, respectively. Here $\hat{k} \geq \frac{k+5}{2}$ is chosen so that the quadrature integrates polynomials up to degree $k+2$ exactly, which is required in the analysis. In element $\boldsymbol{K}$, we define the auxiliary sets $\hat{S}_{\varepsilon,\otimes}^{\boldsymbol{K}}$ and $\hat{S}_{i,\otimes}^{\boldsymbol{K}}$, $i = 1, \ldots, d_{\boldsymbol{x}}$, as

$$\hat{S}_{\varepsilon,\otimes}^{\boldsymbol{K}} := \hat{S}_\varepsilon^{\boldsymbol{K}} \otimes \left(\bigotimes_{i=1}^{d_{\boldsymbol{x}}} S_i^{\boldsymbol{K}} \right) \quad \text{and} \quad \hat{S}_{i,\otimes}^{\boldsymbol{K}} := S_\varepsilon^{\boldsymbol{K}} \otimes \left(\bigotimes_{j=1, j\neq i}^{d_{\boldsymbol{x}}} S_j^{\boldsymbol{K}} \right) \otimes \hat{S}_i^{\boldsymbol{K}}, \tag{50}$$

respectively. We denote the union of these auxiliary sets in element $\boldsymbol{K}$ as

$$\hat{S}_\otimes^{\boldsymbol{K}} := \hat{S}_{\varepsilon,\otimes}^{\boldsymbol{K}} \cup \left(\bigcup_{i=1}^{d_{\boldsymbol{x}}} \hat{S}_{i,\otimes}^{\boldsymbol{K}} \right) \tag{51}$$

and further denote the union of the auxiliary sets and the local DG nodes as

$$\tilde{S}_\otimes^{\boldsymbol{K}} := S_\otimes^{\boldsymbol{K}} \cup \hat{S}_\otimes^{\boldsymbol{K}}. \tag{52}$$

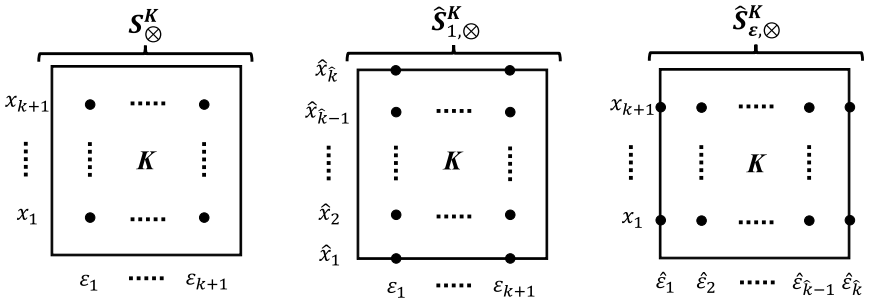


Fig. 3. Illustration of the collection of DG nodes S^K and the auxiliary phase-space point sets $\hat{S}^K_{1,\otimes}$ and $\hat{S}^K_{\varepsilon,\otimes}$ in an element $\boldsymbol{K}$ in a computational domain $\mathbb{R}\times\mathbb{R}^+$. These sets are defined in Eqs. (48) and (50), respectively. In this figure, $\hat{S}^K_{i,\otimes}$ reduces to $\hat{S}^K_{1,\otimes}$ since here $\boldsymbol{x}=x^1$ is considered.

An illustration of the local point sets S^K, $\hat{S}^K_{1,\otimes}$, and $\hat{S}^K_{\varepsilon,\otimes}$ is given in Fig. 3, in which the case $d_{\boldsymbol{x}}=1$ and $\boldsymbol{x}=x^1$ is considered. Therefore $\hat{S}^K_{i,\otimes}$ is simply $\hat{S}^K_{1,\otimes} := S^K_\varepsilon \otimes \hat{S}^K_1$, as defined in Eq. (50).

4.2. Time integration

We use IMEX methods to evolve the semi-discrete two-moment model in Eq. (36) forward in time, where the phase-space advection term is treated explicitly and the collision term is treated implicitly. The general s-stage IMEX scheme can then be written as [33,34]

$$\begin{aligned}\left(\mathcal{U}_h^{(i)},\varphi_h\right)_{\boldsymbol{K}} &= \left(\mathcal{U}_h^{n},\varphi_h\right)_{\boldsymbol{K}} \\ &\quad + \Delta t\sum_{j=1}^{i-1}\tilde{\alpha}_{ij}\,\mathcal{B}_h\left(\mathcal{U}_h^{(j)},\boldsymbol{v}_h,\varphi_h\right)_{\boldsymbol{K}} + \Delta t\sum_{j=1}^{i}\alpha_{ij}\left(C(\mathcal{U}_h^{(j)}),\varphi_h\right)_{\boldsymbol{K}},\end{aligned} \tag{53}$$

$$\begin{aligned}\left(\mathcal{U}_h^{n+1},\varphi_h\right)_{\boldsymbol{K}} &= \left(\mathcal{U}_h^{n},\varphi_h\right)_{\boldsymbol{K}} \\ &\quad + \Delta t\sum_{i=1}^{s}\tilde{w}_i\,\mathcal{B}_h\left(\mathcal{U}_h^{(i)},\boldsymbol{v}_h,\varphi_h\right)_{\boldsymbol{K}} + \Delta t\sum_{i=1}^{s}w_i\left(C(\mathcal{U}_h^{(i)}),\varphi_h\right)_{\boldsymbol{K}},\end{aligned} \tag{54}$$

for $i=1,\ldots,s$, all $\boldsymbol{K}\in\mathcal{T}$, and all $\varphi_h\in\mathbb{V}_h^k(\boldsymbol{K})$. Here the coefficients $\tilde{\alpha}_{ij}$, α_{ij}, $\tilde{w}_i$, and w_i are required to satisfy certain order conditions for achieving the desired accuracy of the IMEX scheme. In addition, to preserve realizability of the evolved moments, each stage in the IMEX update needs to be formulated as convex combinations of realizable terms, which results in additional restrictions on the choices of coefficients. We refer the readers to [35, Section 6] for details on the order and convex-invariant conditions on the coefficients in the IMEX scheme.

4.3. Iterative solvers for nonlinear systems

In this section, we introduce the iterative solvers for the nonlinear systems that occur in the evolution of the IMEX scheme in Eqs. (53)–(54). In Section 4.3.1, we present the iterative solver for the conversion of conserved moments $\mathcal{U}$ to primitive moments $\mathcal{M}$. This moment conversion is required to evaluate the closures for the higher-order moments $\mathcal{K}^{ij}$ and $\mathcal{Q}^{ijk}$ at each stage of the IMEX scheme, since these closures are defined in terms of the primitive moments as discussed in Section 3. In Section 4.3.2, we discuss the solver for the nonlinear equations arising from the implicit update in the IMEX scheme. Even though the simplified collision term $C(\mathcal{U})$ in Eq. (17) appears to be linear in terms of the primitive moments, the implicit system is still nonlinear because the IMEX scheme evolves the conserved moments. This nonlinear system formulation is also extendable to handle systems with collision terms that include more comprehensive physics; e.g., neutrino–electron scattering and thermal pair processes, as considered in [36].

Under the nodal DG framework (see Eq. (49)), each of these nonlinear systems can be formulated locally at each node in the phase-space element because there is no coupling between nodes in either the moment conversion or the collision solve. Therefore, the nonlinear systems considered in this section are written in terms of the nodal moments at a given phase-space node $\boldsymbol{z}\in S^K_\otimes$, $\forall\boldsymbol{K}\in\mathcal{T}$, where $S^K_\otimes$ is the set of DG nodes in element $\boldsymbol{K}$, as defined in Eq. (48). For convenience, we drop the subscript from the nodal representation in this section, and note that, such nonlinear systems must be solved at each $\boldsymbol{z}\in S^K_\otimes$ and in each $\boldsymbol{K}\in\mathcal{T}$ to perform the moment conversion from $\mathcal{U}$ to $\mathcal{M}$ or the implicit steps in the IMEX scheme.

4.3.1. Moment conversion solver

For a given conserved moment $\mathcal{U}\in\mathcal{R}$, finding a corresponding primitive moment $\mathcal{M}\in\mathcal{R}$ that satisfies Eq. (19) requires solving a nonlinear system. A naive approach is to formulate Eq. (19) as a fixed-point problem

$$\mathcal{M}=\begin{pmatrix}\mathcal{D}\\ \mathcal{I}_j\end{pmatrix}=\begin{pmatrix}-v^i\mathcal{I}_i+\mathcal{N}\\ -v^i\mathsf{k}_{ij}\mathcal{D}+\mathcal{G}_j\end{pmatrix}:=\tilde{\mathcal{H}}_{\mathcal{U}}(\mathcal{M}). \tag{55}$$

However, when standard fixed-point iteration, i.e., Picard iteration (see, e.g., [54, Section I.8]), is applied to solve Eq. (55), this formulation does not guarantee that the resulting moments are realizable at each iteration, which, in turn, may result in failures to convergence on problems in this form. To address these issues, we adopt the idea from Richardson iteration, see, e.g., [41] and [42, Section 13.2.1], for solving linear systems and reformulate the fixed-point problem in Eq. (55) as

$$\boldsymbol{\mathcal{M}} = \begin{pmatrix} \mathcal{D} \\ \mathcal{I}_j \end{pmatrix} = \begin{pmatrix} \mathcal{D} \\ \mathcal{I}_j \end{pmatrix} - \lambda \begin{pmatrix} \mathcal{D} + v^i \mathcal{I}_i - \mathcal{N} \\ \mathcal{I}_j + v^i \mathsf{k}_{ij} \mathcal{D} - \mathcal{G}_j \end{pmatrix} := \boldsymbol{\mathcal{H}}_{\boldsymbol{\mathcal{U}}}(\boldsymbol{\mathcal{M}}), \tag{56}$$

where $\lambda \in (0, 1]$ is a constant. Here we choose $\lambda = (1 + v)^{-1}$, where $v := |\boldsymbol{v}| = \sqrt{v_i v^i}$, to guarantee the realizability-preserving and convergence properties of the Picard iteration method; i.e.,

$$\boldsymbol{\mathcal{M}}^{[k+1]} = \boldsymbol{\mathcal{H}}_{\boldsymbol{\mathcal{U}}}(\boldsymbol{\mathcal{M}}^{[k]}). \tag{57}$$

The realizability-preserving and convergence properties of Eq. (57) are stated and proved in Section 5.3.

4.3.2. Collision solver

The implicit steps in Eq. (53) require solving nonlinear systems to find the updated conserved moments. Similar to the implicit systems considered in [36], these systems take the form

$$\boldsymbol{\mathcal{U}} = \boldsymbol{\mathcal{U}}^{(*)} + \Delta t_C \, \boldsymbol{C}(\boldsymbol{\mathcal{U}}) \,, \tag{58}$$

where $\boldsymbol{\mathcal{U}}^{(*)}$ denotes the known conserved moments from the explicit steps, $\boldsymbol{\mathcal{U}}$ denotes the unknown updated conserved moments driven by the implicit collision term $\boldsymbol{C}(\boldsymbol{\mathcal{U}})$ defined in Eq. (17), and Δt_C denotes the effective time step size for the implicit system. Since the sources are expressed in terms of primitive moments, we solve Eq. (58) as a nonlinear fixed-point problem on the unknown primitive moments and use the primitive moment solution to compute the collision term $\boldsymbol{C}(\boldsymbol{\mathcal{U}})$, which is then used to update the conserved moments $\boldsymbol{\mathcal{U}}$ in Eq. (58). As in the moment conversion case discussed in Section 4.3.1, we apply the idea from Richardson iteration to Eq. (58) and formulate it as a fixed-point problem in terms of the primitive moments; i.e.,

$$\boldsymbol{\mathcal{M}} = \begin{pmatrix} \mathcal{D} \\ \mathcal{I}_j \end{pmatrix} = \begin{pmatrix} \mathcal{D} \\ \mathcal{I}_j \end{pmatrix} - \lambda \begin{pmatrix} \mathcal{D} + v^i \mathcal{I}_i - \mathcal{N}^{(*)} - \Delta t_C \chi \left(\mathcal{D}_0 - \mathcal{D} \right) \\ \mathcal{I}_j + v^i \mathsf{k}_{ij} \mathcal{D} - \mathcal{G}_j^{(*)} + \Delta t_C \kappa \, \mathcal{I}_j \end{pmatrix} =: \tilde{\boldsymbol{Q}}(\boldsymbol{\mathcal{M}}), \tag{59}$$

where $\mathcal{N}^{(*)}$ and $\boldsymbol{\mathcal{G}}^{(*)}$ denote the number density and number flux components of the given conserved moment $\boldsymbol{\mathcal{U}}^{(*)}$, respectively, and the constant $\lambda \in (0, 1]$. Although, this formulation is consistent with the one considered in Section 4.3.1 when there are no collisions ($\chi = \kappa = 0$), it cannot guarantee that the realizability of moments is preserved when collisions are taken into account. To address this issue, we follow the approach taken in [36] and reformulate the fixed-point problem as

$$\boldsymbol{\mathcal{M}} = \begin{pmatrix} \mathcal{D} \\ \mathcal{I}_j \end{pmatrix} = \Lambda \begin{pmatrix} (1 - \lambda)\mathcal{D} + \lambda(-v^i \mathcal{I}_i + \mathcal{N}^{(*)} + \Delta t_C \chi \, \mathcal{D}_0) \\ (1 - \lambda)\mathcal{I}_j + \lambda(-v^i \mathsf{k}_{ij} \mathcal{D} + \mathcal{G}_j^{(*)}) \end{pmatrix} =: \boldsymbol{Q}(\boldsymbol{\mathcal{M}}) \,, \tag{60}$$

where $\Lambda := \mathrm{diag}(\mu_\chi, \mu_\kappa)$ with $\mu_\chi = (1 + \lambda \, \Delta t_C \, \chi)^{-1}$ and $\mu_\kappa = (1 + \lambda \, \Delta t_C \, \kappa)^{-1}$. Applying Picard iteration to this fixed-point problem then leads to the iterative scheme

$$\boldsymbol{\mathcal{M}}^{[k+1]} = \boldsymbol{Q}(\boldsymbol{\mathcal{M}}^{[k]}) \,. \tag{61}$$

In Section 5.4, we prove the realizability-preserving and convergence properties of this iterative solver with $\lambda = (1 + v)^{-1}$.

4.4. Flowcharts for the DG-IMEX method

The proposed DG-IMEX scheme is summarized in the flowchart in Fig. 4, in which the left chart shows the procedure of one forward-backward Euler step, and each of the three processes in one time step is described in one of the three charts on the right-hand side. This flowchart can be extended to more general IMEX time integration schemes. The advection and collision updates are given in Eqs. (62a) and (62c), respectively. In the advection update, conversion between conserved moment $\boldsymbol{\mathcal{U}}$ and primitive moment $\boldsymbol{\mathcal{M}}$ discussed in Section 4.3.1 is performed to allow for closure evaluation, which is needed for evaluating the flux of the moment system. The collision update invokes the moment conversion in Section 4.3.1 and the implicit collusion solver in Section 4.3.2. The steps for enforcing realizability and improving conservation properties are given later in Algorithms 1 and 2, respectively.

5. Realizability-preserving property

In this section, we show that, by imposing a proper time-step restriction and a realizability-enforcing limiter, the DG scheme with IMEX time integration given in Section 4 preserves the realizability of both conserved and primitive moments. To this end, we focus on the analysis of a forward-backward Euler method for its simplicity. The theoretical results can be extended to more general IMEX methods that are strong stability-preserving (SSP) with the size of time steps dependent only on the explicit part, such as the IMEX scheme implemented in the numerical tests reported in Section 8. Specifically, we analyze the realizability-preserving property of the following numerical scheme.

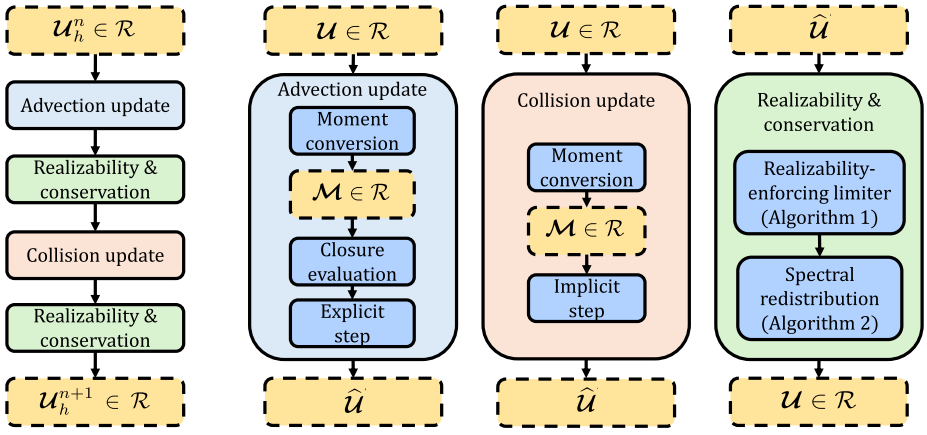


Fig. 4. DG-IMEX flowcharts. The left chart shows the procedure of one forward-backward Euler step in the proposed numerical scheme. Each of the three processes—advection update, collision update, and realizability enforcement and conservation improvement—is described in one of the three charts on the right-hand side.

$$\left(\widehat{\mathcal{U}}_h^{n+1/2}, \varphi_h\right)_{\boldsymbol{K}} = \left(\mathcal{U}_h^n, \varphi_h\right)_{\boldsymbol{K}} + \Delta t\, \mathcal{B}_h\left(\mathcal{U}_h^n, \boldsymbol{v}_h, \varphi_h\right)_{\boldsymbol{K}}, \tag{62a}$$

$$\mathcal{U}_h^{n+1/2} = \texttt{RealizabilityLimiter}(\widehat{\mathcal{U}}_h^{n+1/2}), \tag{62b}$$

$$\left(\widehat{\mathcal{U}}_h^{n+1}, \varphi_h\right)_{\boldsymbol{K}} = \left(\mathcal{U}_h^{n+1/2}, \varphi_h\right)_{\boldsymbol{K}} + \Delta t\left(C(\widehat{\mathcal{U}}_h^{n+1}), \varphi_h\right)_{\boldsymbol{K}}, \tag{62c}$$

$$\mathcal{U}_h^{n+1} = \texttt{RealizabilityLimiter}(\widehat{\mathcal{U}}_h^{n+1}). \tag{62d}$$

Here, `RealizabilityLimiter()` denotes the realizability-enforcing limiter proposed in [35], the details of which is given in Section 5.2 for completeness.

Loosely speaking, the realizability-preserving property of the scheme (62a)–(62d) requires that, if the current moments $\mathcal{U}_h^n$ are realizable, then the updated moments $\mathcal{U}_h^{n+1}$ remain realizable. In the following paragraphs, we summarize the realizability-preserving properties proved in this section, where more detailed realizability results and conditions are described using the sets of phase-space points defined in Section 4.1.

A key assumption in the realizability analysis is the exact closure assumption.

Assumption 1 *(Exact closures).* The moment closures for closing the higher order moments $\mathcal{K}^{ij}$ and $\mathcal{Q}^{ijk}$ in Eq. (62a) are exact, i.e., given lower order primitive moments $(\mathcal{D}, \mathcal{I}^i)$, the moments $(\mathcal{K}^{ij}, \mathcal{Q}^{ijk})$ are computed such that $(\mathcal{D}, \mathcal{I}^i, \mathcal{K}^{ij}, \mathcal{Q}^{ijk})$ satisfy Eq. (1) for some nonnegative distribution f.

We note that Assumption 1 holds when the exact Minerbo closure is used, i.e., when the Eddington and heat-flux factors are given in Eq. (26) (as opposed to the approximation given in Eqs. (27)–(28)). Evaluating (either the exact or approximate) Eddington factor and heat-flux factor uses the flux factor $h\,(=\mathcal{I}/\mathcal{D})$ of the primitive moments $\mathcal{M} = \left(\mathcal{D}, \mathcal{I}\right)^{\intercal}$. Since the numerical scheme Eqs. (62a)–(62d) evolves the conserved moments $\mathcal{U}$, evaluating moment closures requires the conversion between conserved and primitive moments. In other words, given $\mathcal{U}$ (or $\mathcal{M}$), the solver needs to compute the associated $\mathcal{M}$ (or $\mathcal{U}$) that satisfies Eq. (19).

Under Assumption 1, we state the main theoretical result of the realizability-preserving analysis for the scheme Eqs. (62a)–(62d) in Theorem 1, where this scheme is shown to preserve realizability of moments $\mathcal{U}_h$ on the point set $\widetilde{S}_{\otimes}^{\boldsymbol{K}}$ defined in Eq. (52), for all elements $\boldsymbol{K} \in \mathcal{T}$.

Theorem 1 *(Realizability preservation). Suppose (i) Assumption 1 holds, (ii)* $v_h := |\boldsymbol{v}_h| < 1$ *for all* $\boldsymbol{K} \in \mathcal{T}$*, and (iii) the time step* Δt *in Eq. (62a) satisfies the hyperbolic-type time-step restriction*

$$\begin{aligned} &\Delta t \le \min\left\{\Delta t_{\boldsymbol{x}}^{\min}, \Delta t_{\varepsilon}^{\min}\right\}, \text{ with} \\ &\Delta t_{\boldsymbol{x}}^{\min} := \min_{\boldsymbol{K}\in\mathcal{T}} \min_i (1 - v_h) C^i |K_{\boldsymbol{x}}^i| \quad \text{and} \quad \Delta t_{\varepsilon}^{\min} := \min_{\boldsymbol{K}\in\mathcal{T}} (1 - v_h) C^{\epsilon} |K_{\varepsilon}| / \varepsilon_H \end{aligned} \tag{63}$$

where C^i *and* C^{ϵ}*, which are independent of the size of elements in the discretization, are given in Eqs. (80) and (93), respectively. Then the scheme (62a)–(62d) is realizability-preserving, i.e.,* $\mathcal{U}_h^{n+1} \in \mathcal{R}$ *on* $\widetilde{S}_{\otimes}^{\boldsymbol{K}}$, $\forall \boldsymbol{K} \in \mathcal{T}$*, provided that* $\mathcal{U}_h^n \in \mathcal{R}$ *on* $\widetilde{S}_{\otimes}^{\boldsymbol{K}}$, $\forall \boldsymbol{K} \in \mathcal{T}$.

Theorem 1 is a direct consequence of the following Propositions 3, 4, 5, and 6, which provide the realizability-preserving properties of the explicit update (Eq. (62a)), the realizability-enforcing limiter (Eq. (62b)), the moment conversion (Eq. (19)), and the implicit update (Eq. (62c)), respectively. In these propositions, the notion of cell-averaged moments will come in handy. Given $\mathcal{U}_h^n$, the cell-averaged moments $\mathcal{U}_{\boldsymbol{K}} := (\mathcal{N}_{\boldsymbol{K}}, \mathcal{G}_{\boldsymbol{K}})$ are defined as

$$\mathcal{U}_{\boldsymbol{K}} = \left(\mathcal{U}_h\right)_{\boldsymbol{K}} / |\boldsymbol{K}|. \tag{64}$$

Proposition 3 *(Explicit advection update). Suppose (i) Assumption 1 holds, (ii) $v_h < 1$ for all $\boldsymbol{K} \in \mathcal{T}$, and (iii) Δt in Eq. (62a) satisfies the restriction (63). Let $\widehat{\boldsymbol{\mathcal{U}}}_{\boldsymbol{K}}^{n+1/2} := (\widehat{\mathcal{N}}_{\boldsymbol{K}}^{n+1/2}, \widehat{\mathcal{G}}_{\boldsymbol{K}}^{n+1/2})$ denote the element average of the moment $\widehat{\boldsymbol{\mathcal{U}}}_h^{n+1/2}$ (as defined in Eq. (64)) updated by Eq. (62a) from $\boldsymbol{\mathcal{U}}_h^n$. Then, it is guaranteed that, $\forall \boldsymbol{K} \in \mathcal{T}$, $\widehat{\mathcal{N}}_{\boldsymbol{K}}^{n+1/2} > 0$, provided $\boldsymbol{\mathcal{U}}_h^n \in \mathcal{R}$ on $S_{\otimes}^{\boldsymbol{K}}$, $\forall \boldsymbol{K} \in \mathcal{T}$. Further, when a reduced one-dimensional planar geometry[3] is considered, it is guaranteed that, $\forall \boldsymbol{K} \in \mathcal{T}$, $\widehat{\boldsymbol{\mathcal{U}}}_{\boldsymbol{K}}^{n+1/2} \in \mathcal{R}$ when an additional time-step restriction (101) is satisfied.*

Proposition 4 *(Realizability-enforcing limiter). Suppose $\widehat{\mathcal{N}}_{\boldsymbol{K}} > 0$ on element $\boldsymbol{K}$, applying the realizability-enforcing limiter given in Algorithm 1 (see Section 5.2) to the moments $\widehat{\boldsymbol{\mathcal{U}}}_h$ on $\boldsymbol{K}$ leads to realizable moments $\boldsymbol{\mathcal{U}}_h \in \mathcal{R}$ on $S_{\otimes}^{\boldsymbol{K}} \cup \widehat{S}_{\otimes}^{\boldsymbol{K}}$ in $\boldsymbol{K}$.*

Proposition 5 *(Moment conversion). Suppose that Assumption 1 holds and that $v_h < 1$ on all $\boldsymbol{K} \in \mathcal{T}$, then the conversion between conserved and primitive moments following the relation in Eq. (19) preserves realizability, i.e., for a pair of conserved and primitive moments $(\boldsymbol{\mathcal{U}}, \boldsymbol{\mathcal{M}})$ satisfying Eq. (19), $\boldsymbol{\mathcal{U}} \in \mathcal{R}$ if and only if $\boldsymbol{\mathcal{M}} \in \mathcal{R}$. Further, given $\boldsymbol{\mathcal{U}} \in \mathcal{R}$, the iterative solver (57) in Section 4.3.1 converges to the unique $\boldsymbol{\mathcal{M}} \in \mathcal{R}$ that satisfies Eq. (19).*

Proposition 6 *(Implicit collision solve). Suppose Assumption 1 holds and $v_h < 1$ on all $\boldsymbol{K} \in \mathcal{T}$. Let $\boldsymbol{\mathcal{U}}_h^{n+1/2} \in \mathcal{R}$ on $S_{\otimes}^{\boldsymbol{K}}$ for all $\boldsymbol{K} \in \mathcal{T}$, then solving Eq. (62c) with the iterative solvers considered in Section 4.3 gives $\widehat{\boldsymbol{\mathcal{U}}}_h^{n+1} \in \mathcal{R}$ on $S_{\otimes}^{\boldsymbol{K}}$ for all $\boldsymbol{K} \in \mathcal{T}$.*

These propositions form a basis for the proof of Theorem 1. Specifically, Proposition 3 guarantees that the updated moments $\widehat{\boldsymbol{\mathcal{U}}}_h^{n+1/2}$ from Eq. (62a) have a nonnegative cell-averaged density $\widehat{\mathcal{N}}_{\boldsymbol{K}}^{n+1/2}$ for each $\boldsymbol{K} \in \mathcal{T}$. It follows from Proposition 4 that the limited moments $\boldsymbol{\mathcal{U}}_h^{n+1/2}$ are realizable on $S_{\otimes}^{\boldsymbol{K}}$ for all $\boldsymbol{K} \in \mathcal{T}$. Solving Eq. (62c) on each nodal point in $S_{\otimes}^{\boldsymbol{K}}$ for all $\boldsymbol{K} \in \mathcal{T}$ gives the updated moment $\widehat{\boldsymbol{\mathcal{U}}}_h^{n+1}$, which is guaranteed to be realizable on $S_{\otimes}^{\boldsymbol{K}}$, $\forall \boldsymbol{K} \in \mathcal{T}$, by Proposition 6. Applying the realizability-enforcing limiter again to $\widehat{\boldsymbol{\mathcal{U}}}_h^{n+1}$ on every $\boldsymbol{K} \in \mathcal{T}$ leads to $\boldsymbol{\mathcal{U}}_h^{n+1}$, which is realizable on $\widehat{S}_{\otimes}^{\boldsymbol{K}}$, $\forall \boldsymbol{K} \in \mathcal{T}$, again from Proposition 4.

In Sections 5.1, 5.2, 5.3, and 5.4, we prove Propositions 3, 4, 5, and 6, respectively. These results together lead to the main realizability-preserving property of the numerical scheme Eqs. (62a)–(62c) given in Theorem 1, under the exact closure assumption, Assumption 1. In Section 5.5, we extend the realizability-preserving and convergence results in Propositions 5 and 6 to the case of evaluating the closure with the approximate Eddington factor ψ_{a} in Eq. (27), which is often used in practice to reduce computational cost.

5.1. Explicit advection update

In this section, we prove Proposition 3 by deriving the time-step restriction (63) under which the updated cell-averaged number density $\widehat{\mathcal{N}}_{\boldsymbol{K}}^{n+1/2} > 0$. In a one-dimensional planar geometry, we show that $\widehat{\boldsymbol{\mathcal{U}}}_{\boldsymbol{K}}^{n+1/2} \in \mathcal{R}$ under an additional time-step restriction given in Eq. (101).

Since constant functions are in the approximation space $\mathbb{V}_h^k(\boldsymbol{K})$, we start with deriving the update formula for cell-averaged moments by setting $\varphi_h = 1$ in Eq. (62a), which leads to

$$
\begin{aligned}
\widehat{\boldsymbol{\mathcal{U}}}_{\boldsymbol{K}}^{n+1/2} &= \boldsymbol{\mathcal{U}}_{\boldsymbol{K}}^{n} + \frac{\Delta t}{|\boldsymbol{K}|} \mathcal{B}_h\big(\boldsymbol{\mathcal{U}}_h^n, \boldsymbol{v}_h\big)_{\boldsymbol{K}} \\
&= \gamma^{\boldsymbol{x}} \left\{ \boldsymbol{\mathcal{U}}_{\boldsymbol{K}}^{n} + \frac{\Delta t}{\gamma^{\boldsymbol{x}} |\boldsymbol{K}|} \mathcal{B}_h^{\boldsymbol{x}}\big(\boldsymbol{\mathcal{U}}_h^n, \boldsymbol{v}_h\big)_{\boldsymbol{K}} \right\} + \gamma^{\varepsilon} \left\{ \boldsymbol{\mathcal{U}}_{\boldsymbol{K}}^{n} + \frac{\Delta t}{\gamma^{\varepsilon} |\boldsymbol{K}|} \mathcal{B}_h^{\varepsilon}\big(\boldsymbol{\mathcal{U}}_h^n, \boldsymbol{v}_h\big)_{\boldsymbol{K}} \right\} \\
&\quad + \gamma^{\mathcal{S}} \left\{ \boldsymbol{\mathcal{U}}_{\boldsymbol{K}}^{n} + \frac{\Delta t}{\gamma^{\mathcal{S}} |\boldsymbol{K}|} \big(\mathcal{S}(\boldsymbol{\mathcal{U}}_h^n, \boldsymbol{v}_h)\big)_{\boldsymbol{K}} \right\}, \\
&=: \gamma^{\boldsymbol{x}} \widehat{\boldsymbol{\mathcal{U}}}_{\boldsymbol{K}}^{n+1/2,\boldsymbol{x}} + \gamma^{\varepsilon} \widehat{\boldsymbol{\mathcal{U}}}_{\boldsymbol{K}}^{n+1/2,\varepsilon} + \gamma^{\mathcal{S}} \widehat{\boldsymbol{\mathcal{U}}}_{\boldsymbol{K}}^{n+1/2,\mathcal{S}},
\end{aligned}
\tag{65}
$$

where we have defined $\gamma^{\boldsymbol{x}}, \gamma^{\varepsilon}, \gamma^{\mathcal{S}} > 0$, satisfying $\gamma^{\boldsymbol{x}} + \gamma^{\varepsilon} + \gamma^{\mathcal{S}} = 1$. In the following subsections, we show that, when $\boldsymbol{\mathcal{U}}_h^n \in \mathcal{R}$ on $\widehat{S}_{\otimes}^{\boldsymbol{K}}$ for all $\boldsymbol{K} \in \mathcal{T}$, $\widehat{\boldsymbol{\mathcal{U}}}_{\boldsymbol{K}}^{n+1/2,\boldsymbol{x}}$ and $\widehat{\boldsymbol{\mathcal{U}}}_{\boldsymbol{K}}^{n+1/2,\varepsilon}$ are realizable under time-step restrictions given in Eq. (63) (Sections 5.1.1 and 5.1.2) and that $\widehat{\mathcal{N}}_{\boldsymbol{K}}^{n+1/2,\mathcal{S}} > 0$ (Section 5.1.3) for all $\boldsymbol{K} \in \mathcal{T}$. Further, we show in Section 5.1.3 that $\widehat{\boldsymbol{\mathcal{U}}}_{\boldsymbol{K}}^{n+1/2,\mathcal{S}}$ is realizable in one-dimensional, planar geometry under an additional time-step restriction given in Eq. (101). Since the realizable set $\mathcal{R}$ is convex and $\widehat{\boldsymbol{\mathcal{U}}}_{\boldsymbol{K}}^{n+1/2}$ is written as a convex combination of $\widehat{\boldsymbol{\mathcal{U}}}_{\boldsymbol{K}}^{n+1/2,\boldsymbol{x}}$, $\widehat{\boldsymbol{\mathcal{U}}}_{\boldsymbol{K}}^{n+1/2,\varepsilon}$, and $\widehat{\boldsymbol{\mathcal{U}}}_{\boldsymbol{K}}^{n+1/2,\mathcal{S}}$ in Eq. (65), we thus conclude that, under the time-step restrictions in Eqs. (63) and (101), (i) $\widehat{\mathcal{N}}_{\boldsymbol{K}}^{n+1/2} > 0$ and (ii) $\widehat{\boldsymbol{\mathcal{U}}}_{\boldsymbol{K}}^{n+1/2} \in \mathcal{R}$ in a planar geometry.

[3] An example of this one-dimensional geometry is the reduced case of the full three-dimensional geometry when the fluxes in two of the three spatial dimensions are assumed to be zero. See Section 5.1.3 for further discussions.

5.1.1. Position space fluxes

For transport in position space we follow the approach in [35] and write

$$\widehat{\boldsymbol{\mathcal{U}}}_{\boldsymbol{K}}^{n+1/2,\boldsymbol{x}} = \boldsymbol{\mathcal{U}}_{\boldsymbol{K}}^{n} + \frac{\Delta t_{\boldsymbol{x}}}{|\boldsymbol{K}|}\,\mathcal{B}_h^{\boldsymbol{x}}\big(\boldsymbol{\mathcal{U}}_h^n, \boldsymbol{v}_h\big)_{\boldsymbol{K}} \quad (\Delta t_{\boldsymbol{x}} = \Delta t/\gamma^{\boldsymbol{x}}). \tag{66}$$

To find sufficient conditions such that $\widehat{\boldsymbol{\mathcal{U}}}_{\boldsymbol{K}}^{n+1/2,\boldsymbol{x}} \in \mathcal{R}$, we define (cf. [35])

$$\Gamma^i\big[\boldsymbol{\mathcal{U}}_h^n\big](\tilde{\boldsymbol{z}}^i) = \frac{1}{|K_{\boldsymbol{x}}^i|}\Big[\int_{K_{\boldsymbol{x}}^i} \boldsymbol{\mathcal{U}}_h^n\, dx^i - \frac{\Delta t_{\boldsymbol{x}}}{\beta^i}\Big(\widehat{\mathcal{F}^i}\big(\boldsymbol{\mathcal{U}}_h^n, \boldsymbol{v}_h\big)|_{x_{\mathrm{H}}^i} - \widehat{\mathcal{F}^i}\big(\boldsymbol{\mathcal{U}}_h^n, \boldsymbol{v}_h\big)|_{x_{\mathrm{L}}^i}\Big)\Big], \tag{67}$$

so that

$$\widehat{\boldsymbol{\mathcal{U}}}_{\boldsymbol{K}}^{n+1/2,\boldsymbol{x}} = \sum_{i=1}^{d_{\boldsymbol{x}}} \frac{\beta^i}{|\tilde{\boldsymbol{K}}^i|}\int_{\tilde{\boldsymbol{K}}^i} \Gamma^i\big[\boldsymbol{\mathcal{U}}_h^n\big](\tilde{\boldsymbol{z}}^i)\,\tau\, d\tilde{\boldsymbol{z}}^i, \tag{68}$$

where we have defined the set of positive constants $\{\beta^i\}_{i=1}^{d_{\boldsymbol{x}}}$ satisfying $\sum_{i=1}^{d_{\boldsymbol{x}}}\beta^i = 1$.

If a quadrature rule $\tilde{\boldsymbol{\mathcal{Q}}}^i : C^0(\tilde{\boldsymbol{K}}^i) \to \mathbb{R}$ with positive weights, e.g., the tensor product of one-dimensional LG quadrature, is used to approximate the integral in Eq. (68), it is sufficient to show that, under the assumptions in Proposition 3, $\Gamma^i\big[\boldsymbol{\mathcal{U}}_h^n\big](\tilde{\boldsymbol{z}}^i) \in \mathcal{R}$ holds for $\tilde{\boldsymbol{z}}^i \in \tilde{\mathcal{S}}^i \subset \tilde{\boldsymbol{K}}^i$, where $\tilde{\mathcal{S}}^i$ denotes the set of quadrature points given by $\tilde{\boldsymbol{\mathcal{Q}}}^i$. We prove this sufficient condition in the remainder of this subsection.

Let $\hat{\mathcal{Q}}^i : C^0(K_{\boldsymbol{x}}^i) \to \mathbb{R}$ denote the $\hat{k}$-point LGL quadrature rule on $K_{\boldsymbol{x}}^i$ with points $\hat{\mathcal{S}}_i^{\boldsymbol{K}} = \{x_{\mathrm{L}}^i = \hat{x}_1^i, \ldots, \hat{x}_{\hat{k}}^i = x_{\mathrm{H}}^i\}$ as defined in Section 4.1 and strictly positive weights $\{\hat{w}_q\}_{q=1}^{\hat{k}}$, normalized such that $\sum_{q=1}^{\hat{k}}\hat{w}_q = 1$. Since $\hat{k} \geq \frac{k+5}{2}$, this quadrature integrates $\boldsymbol{\mathcal{U}}_h^n$ exactly, and thus we have

$$\int_{K_{\boldsymbol{x}}^i} \boldsymbol{\mathcal{U}}_h^n(x^i)\, dx^i = \hat{\mathcal{Q}}^i\big[\boldsymbol{\mathcal{U}}_h^n\big] = |K_{\boldsymbol{x}}^i| \sum_{q=1}^{\hat{k}} \hat{w}_q\, \boldsymbol{\mathcal{U}}_h^n(\hat{x}_q^i), \tag{69}$$

where, for notational convenience, we have suppressed explicit dependence on $\tilde{\boldsymbol{z}}^i$ in writing $\boldsymbol{\mathcal{U}}_h^n(\hat{x}_q^i, \tilde{\boldsymbol{z}}^i) = \boldsymbol{\mathcal{U}}_h^n(\hat{x}_q^i)$. Similarly, $\boldsymbol{\mathcal{U}}_h^n(x_{\mathrm{L}}^{i,\pm}, \tilde{\boldsymbol{z}}^i) = \boldsymbol{\mathcal{U}}_h^n(x_{\mathrm{L}}^{i,\pm})$ and $\boldsymbol{\mathcal{U}}_h^n(x_{\mathrm{H}}^{i,\pm}, \tilde{\boldsymbol{z}}^i) = \boldsymbol{\mathcal{U}}_h^n(x_{\mathrm{H}}^{i,\pm})$. Then, using the quadrature rule in Eq. (69) and the LF flux in Eq. (42), we can write Eq. (67) as a convex combination

$$\begin{aligned}
&\Gamma^i\big[\boldsymbol{\mathcal{U}}_h^n\big](\tilde{\boldsymbol{z}}^i) \\
&= \sum_{q=2}^{\hat{k}-1} \hat{w}_q\, \boldsymbol{\mathcal{U}}_h^n(\hat{x}_q^i) + \hat{w}_1\, \Phi_1^i\big[\boldsymbol{\mathcal{U}}_h^n(x_{\mathrm{L}}^{i,-}), \boldsymbol{\mathcal{U}}_h^n(x_{\mathrm{L}}^{i,+}), \hat{\boldsymbol{v}}(x_{\mathrm{L}}^i)\big] + \hat{w}_{\hat{k}}\, \Phi_{\hat{k}}^i\big[\boldsymbol{\mathcal{U}}_h^n(x_{\mathrm{H}}^{i,-}), \boldsymbol{\mathcal{U}}_h^n(x_{\mathrm{H}}^{i,+}), \hat{\boldsymbol{v}}(x_{\mathrm{H}}^i)\big],
\end{aligned} \tag{70}$$

where

$$\Phi_1^i\big[\boldsymbol{\mathcal{U}}_a, \boldsymbol{\mathcal{U}}_b, \hat{\boldsymbol{v}}\big] = \boldsymbol{\mathcal{U}}_b + \lambda_{\boldsymbol{x}}^i\, \mathcal{F}_{\mathrm{LF}}^i\big(\boldsymbol{\mathcal{U}}_a, \boldsymbol{\mathcal{U}}_b, \hat{\boldsymbol{v}}\big), \tag{71}$$

$$\Phi_{\hat{k}}^i\big[\boldsymbol{\mathcal{U}}_a, \boldsymbol{\mathcal{U}}_b, \hat{\boldsymbol{v}}\big] = \boldsymbol{\mathcal{U}}_a - \lambda_{\boldsymbol{x}}^i\, \mathcal{F}_{\mathrm{LF}}^i\big(\boldsymbol{\mathcal{U}}_a, \boldsymbol{\mathcal{U}}_b, \hat{\boldsymbol{v}}\big), \tag{72}$$

and $\lambda_{\boldsymbol{x}}^i = \Delta t_{\boldsymbol{x}}/(\beta^i\, \hat{w}_{\hat{k}}\, |K_{\boldsymbol{x}}^i|)$. Since Eq. (70) is a convex combination, it is sufficient to show the realizability of each term independently to obtain $\Gamma^i\big[\boldsymbol{\mathcal{U}}_h^n\big](\tilde{\boldsymbol{z}}^i) \in \mathcal{R}$. For the first term on the right-hand side of Eq. (70), it is sufficient that $\boldsymbol{\mathcal{U}}_h^n(\hat{x}_q^i) \in \mathcal{R}$, which holds under the assumption that $\boldsymbol{\mathcal{U}}_h^n \in \mathcal{R}$ on $S_{\otimes}^{\boldsymbol{K}}$ for all $\boldsymbol{K} \in \mathcal{T}$. It remains to find conditions for which $\Phi_1^i, \Phi_{\hat{k}}^i \in \mathcal{R}$, which we summarize in the following lemmas.

Lemma 1. *Define*

$$\Theta_{\pm}^i(\boldsymbol{\mathcal{U}}, \hat{\boldsymbol{v}}) = \boldsymbol{\mathcal{U}}[\hat{\boldsymbol{v}}^i] \pm \mathcal{F}^i(\boldsymbol{\mathcal{U}}, \hat{\boldsymbol{v}}), \tag{73}$$

where $\boldsymbol{\mathcal{U}}[\hat{\boldsymbol{v}}^i]$ and $\mathcal{F}^i(\boldsymbol{\mathcal{U}}, \hat{\boldsymbol{v}})$ are defined as in Eqs. (15) and (16), respectively, and $\hat{\boldsymbol{v}}^i = \big(\delta^{i1}\,\hat{v}^1, \delta^{i2}\,\hat{v}^2, \delta^{i3}\,\hat{v}^3\big)^{\intercal}$ as defined in Remark 3. Suppose that $\boldsymbol{\mathcal{U}} \in \mathcal{R}$ and $\hat{v} = |\hat{\boldsymbol{v}}| < 1$. Then $\Theta_{\pm}^i(\boldsymbol{\mathcal{U}}, \hat{\boldsymbol{v}}) \in \mathcal{R}$.

Proof. The first component of $\Theta_{\pm}^i(\boldsymbol{\mathcal{U}}, \hat{\boldsymbol{v}})$ can be written as

$$\frac{1}{4\pi}\int_{\mathbb{S}^2} (1 \pm \hat{v}^i)(1 \pm \ell^i)\, f\, d\omega, \tag{74}$$

while the remaining components can be written as

$$\frac{1}{4\pi}\int_{\mathbb{S}^2}(1\pm\hat{v}^i)(1\pm\ell^i)\,f\,\ell_j\,d\omega,\quad (j=1,2,3). \tag{75}$$

Since $(1\pm\hat{v}^i)(1\pm\ell^i)\,f\in\mathfrak{R}$, the result follows. □

Lemma 2. *Let Φ^i_1 and $\Phi^i_{\hat{k}}$ be defined as in Eqs. (71) and (72), respectively. Assume that the following holds*

(a) *$\mathcal{U}_a,\mathcal{U}_b\in\mathcal{R}$, defined as in Eq. (15) as the moments of distributions $f_a,f_b\in\mathfrak{R}$.*
(b) *The three-velocity in Eq. (43) satisfies $\hat{v}=|\hat{\boldsymbol{v}}|<1$.*
(c) *The time step $\Delta t_{\boldsymbol{x}}$ is chosen such that $\lambda^i_{\boldsymbol{x}}\le(1-\hat{v})$.*

Then $\Phi^i_1\left[\mathcal{U}_a,\mathcal{U}_b,\hat{\boldsymbol{v}}\right],\Phi^i_{\hat{k}}\left[\mathcal{U}_a,\mathcal{U}_b,\hat{\boldsymbol{v}}\right]\in\mathcal{R}$.

Proof. Define

$$\Theta^i_0(\mathcal{U},\hat{\boldsymbol{v}})=\frac{\mathcal{U}[\hat{\boldsymbol{v}}]-\lambda^i_{\boldsymbol{x}}\mathcal{U}[\hat{\boldsymbol{v}}^i]}{1-\lambda^i_{\boldsymbol{x}}}. \tag{76}$$

Then, using the LF flux in Eq. (42), we can write

$$\Phi^i_1\left[\mathcal{U}_a,\mathcal{U}_b,\hat{\boldsymbol{v}}\right]=(1-\lambda^i_{\boldsymbol{x}})\,\Theta^i_0(\mathcal{U}_b,\hat{\boldsymbol{v}})+\frac{1}{2}\lambda^i_{\boldsymbol{x}}\,\Theta^i_+(\mathcal{U}_a,\hat{\boldsymbol{v}})+\frac{1}{2}\lambda^i_{\boldsymbol{x}}\,\Theta^i_+(\mathcal{U}_b,\hat{\boldsymbol{v}}), \tag{77}$$

which is a convex combination for $\lambda^i_{\boldsymbol{x}}<1$. From assumptions (a) and (b) above, it follows from Lemma 1 that $\Theta^i_+(\mathcal{U}_a,\hat{\boldsymbol{v}}),\Theta^i_+(\mathcal{U}_b,\hat{\boldsymbol{v}})\in\mathcal{R}$. It remains to show that $\Theta^i_0(\mathcal{U}_b,\hat{\boldsymbol{v}})\in\mathcal{R}$. The first component of $\Theta^i_0(\mathcal{U}_b,\hat{\boldsymbol{v}})$ can be written as

$$\frac{1}{4\pi}\int_{\mathbb{S}^2}\mathsf{f}(\omega)\,d\omega,\quad\text{where}\quad\mathsf{f}(\omega)=\frac{[(1-\hat{\boldsymbol{v}}\cdot\boldsymbol{\ell})-\lambda^i_{\boldsymbol{x}}(1+\hat{v}^i\,\ell^i)]}{(1-\lambda^i_{\boldsymbol{x}})}\,f, \tag{78}$$

while the remaining components can be written as

$$\frac{1}{4\pi}\int_{\mathbb{S}^2}\mathsf{f}(\omega)\,\ell_j(\omega)\,d\omega,\quad(j=1,2,3). \tag{79}$$

From assumptions (b) and (c), it follows that $\mathsf{f}\in\mathfrak{R}$, which implies $\Theta^i_0(\mathcal{U}_b,\hat{\boldsymbol{v}})\in\mathcal{R}$. The proof for $\Phi^i_{\hat{k}}\left[\mathcal{U}_a,\mathcal{U}_b,\hat{\boldsymbol{v}}\right]$ is analogous and is omitted. □

To this end, the results of Lemma 2 lead to $\widehat{\mathcal{U}}^{n+1/2,\boldsymbol{x}}_{\boldsymbol{K}}\in\mathcal{R}$ under the assumptions therein. It is straightforward to verify that these assumptions are fulfilled for each $\boldsymbol{K}\in\mathcal{T}$ when the assumptions in Proposition 3 hold. In particular, from Eq. (43), it is clear that $\hat{v}<1$ is implied by $v_h<1$. Also, by defining

$$C^i:=\gamma^{\boldsymbol{x}}\beta^i\hat{w}_{\hat{k}}\,, \tag{80}$$

the time-step restriction in Eq. (63) guarantees $\lambda^i_{\boldsymbol{x}}\le(1-\hat{v})$ for all $\boldsymbol{K}\in\mathcal{T}$. Therefore, we have shown that, under the assumptions of Proposition 3, $\widehat{\mathcal{U}}^{n+1/2,\boldsymbol{x}}_{\boldsymbol{K}}\in\mathcal{R}$ for all $\boldsymbol{K}\in\mathcal{T}$.

5.1.2. Energy space fluxes

For energy space advection, we define

$$\widehat{\mathcal{U}}^{n+1/2,\varepsilon}_{\boldsymbol{K}}=\mathcal{U}^n_{\boldsymbol{K}}+\frac{\Delta t_\varepsilon}{|\boldsymbol{K}|}\,\mathcal{B}^\varepsilon_h\left(\mathcal{U}^n_h,\boldsymbol{v}_h\right)_{\boldsymbol{K}}\quad(\Delta t_\varepsilon=\Delta t/\gamma^\varepsilon), \tag{81}$$

and seek to find sufficient conditions such that $\widehat{\mathcal{U}}^{n+1/2,\varepsilon}_{\boldsymbol{K}}\in\mathcal{R}$. We proceed in a fashion similar to that in Section 5.1.1, and define

$$\Gamma^\varepsilon[\mathcal{U}_h](\boldsymbol{x})=\frac{1}{|K_\varepsilon|}\left[\int_{K_\varepsilon}\mathcal{U}_h\,\varepsilon^2 d\varepsilon-\Delta t_\varepsilon\left(\varepsilon^3\,\widehat{\mathcal{F}}^\varepsilon(\mathcal{U}_h,\boldsymbol{v}_h)|_{\varepsilon_{\mathrm{H}}}-\varepsilon^3\,\widehat{\mathcal{F}}^\varepsilon(\mathcal{U}_h,\boldsymbol{v}_h)|_{\varepsilon_{\mathrm{L}}}\right)\right] \tag{82}$$

so that

$$\widehat{\mathcal{U}}^{n+1/2,\varepsilon}_{\boldsymbol{K}}=\frac{1}{|\boldsymbol{K}_{\boldsymbol{x}}|}\int_{\boldsymbol{K}_{\boldsymbol{x}}}\Gamma^\varepsilon[\mathcal{U}_h](\boldsymbol{x})\,d\boldsymbol{x}. \tag{83}$$

Evaluating the integrals in the energy dimension using the same $\hat{k}$-point LGL quadrature rule leads to

$$\Gamma^\epsilon[\boldsymbol{\mathcal{U}}_h](\boldsymbol{x}) = \sum_{q=2}^{\hat{k}-1} \hat{w}_q\, \hat{\varepsilon}_q^2\, \boldsymbol{\mathcal{U}}_h(\hat{\varepsilon}_q) + \hat{w}_1\, \varepsilon_{\mathrm{L}}^2\, \Phi_1^\epsilon\left[\boldsymbol{\mathcal{U}}_h(\varepsilon_{\mathrm{L}}^-), \boldsymbol{\mathcal{U}}_h(\varepsilon_{\mathrm{L}}^+), \boldsymbol{v}_h\right] + \hat{w}_{\hat{k}}\, \varepsilon_{\mathrm{H}}^2\, \Phi_{\hat{k}}^\epsilon\left[\boldsymbol{\mathcal{U}}_h(\varepsilon_{\mathrm{H}}^-), \boldsymbol{\mathcal{U}}_h(\varepsilon_{\mathrm{H}}^+), \boldsymbol{v}_h\right], \tag{84}$$

where the integral of the moments is exact when $\hat{k} \geq \frac{k+5}{2}$, i.e.,

$$\int_{K_\varepsilon} \boldsymbol{\mathcal{U}}_h(\varepsilon)\, \varepsilon^2 d\varepsilon = \hat{\mathcal{Q}}^\epsilon\left[\boldsymbol{\mathcal{U}}_h\right] = |K_\varepsilon| \sum_{q=1}^{\hat{k}} \hat{w}_q\, \boldsymbol{\mathcal{U}}_h(\hat{\varepsilon}_q)\, \hat{\varepsilon}_q^2\,. \tag{85}$$

Since $\Gamma^\epsilon[\boldsymbol{\mathcal{U}}_h](\boldsymbol{x})$ is written as a convex combination in Eq. (84), the realizability of each term on the right-hand side gives the realizability of $\Gamma^\epsilon[\boldsymbol{\mathcal{U}}_h](\boldsymbol{x})$. Since $\boldsymbol{\mathcal{U}}_h(\hat{\varepsilon}_q) \in \mathcal{R}$ for each $\hat{\varepsilon}_q$ under the assumption that $\boldsymbol{\mathcal{U}}_h \in \mathcal{R}$ on $\widehat{S}^{\boldsymbol{K}}_{\otimes}$ for all $\boldsymbol{K} \in \mathcal{T}$, we focus on proving realizability of Φ_1^ϵ and $\Phi_{\hat{k}}^\epsilon$, which are defined as

$$\begin{aligned}
\Phi_1^\epsilon\left[\boldsymbol{\mathcal{U}}_a, \boldsymbol{\mathcal{U}}_b, \boldsymbol{v}_h\right] &= \boldsymbol{\mathcal{U}}_b + \lambda_{\mathrm{L}}^\epsilon\, \mathscr{F}_{\mathrm{LF}}^\epsilon\left(\boldsymbol{\mathcal{U}}_a, \boldsymbol{\mathcal{U}}_b, \boldsymbol{v}_h\right) \qquad (86)\\
&= (1 - \alpha^\epsilon\, \lambda_{\mathrm{L}}^\epsilon)\, \Theta_{0,\mathrm{L}}^\epsilon(\boldsymbol{\mathcal{U}}_b, \boldsymbol{v}_h) + \frac{1}{2}\, \alpha^\epsilon\, \lambda_{\mathrm{L}}^\epsilon\, \Theta_+^\epsilon(\boldsymbol{\mathcal{U}}_a, \boldsymbol{v}_h) + \frac{1}{2}\, \alpha^\epsilon\, \lambda_{\mathrm{L}}^\epsilon\, \Theta_+^\epsilon(\boldsymbol{\mathcal{U}}_b, \boldsymbol{v}_h),\\
\Phi_{\hat{k}}^\epsilon\left[\boldsymbol{\mathcal{U}}_a, \boldsymbol{\mathcal{U}}_b, \boldsymbol{v}_h\right] &= \boldsymbol{\mathcal{U}}_a - \lambda_{\mathrm{H}}^\epsilon\, \mathscr{F}_{\mathrm{LF}}^\epsilon\left(\boldsymbol{\mathcal{U}}_a, \boldsymbol{\mathcal{U}}_b, \boldsymbol{v}_h\right) \qquad (87)\\
&= (1 - \alpha^\epsilon\, \lambda_{\mathrm{H}}^\epsilon)\, \Theta_{0,\mathrm{H}}^\epsilon(\boldsymbol{\mathcal{U}}_a, \boldsymbol{v}_h) + \frac{1}{2}\, \alpha^\epsilon\, \lambda_{\mathrm{H}}^\epsilon\, \Theta_-^\epsilon(\boldsymbol{\mathcal{U}}_a, \boldsymbol{v}_h) + \frac{1}{2}\, \alpha^\epsilon\, \lambda_{\mathrm{H}}^\epsilon\, \Theta_-^\epsilon(\boldsymbol{\mathcal{U}}_b, \boldsymbol{v}_h),
\end{aligned}$$

where we used the definition of $\mathscr{F}_{\mathrm{LF}}^\epsilon$ given in Eq. (46) and defined $\lambda_{\mathrm{L/H}}^\epsilon = \varepsilon_{\mathrm{L/H}}\, \Delta t_\epsilon / (\hat{w}_{\hat{k}}\, |K_\epsilon|)$,

$$\Theta_{0,\mathrm{L/H}}^\epsilon(\boldsymbol{\mathcal{U}}, \boldsymbol{v}_h) = \frac{\boldsymbol{\mathcal{U}}[\boldsymbol{v}_h] - \alpha^\epsilon\, \lambda_{\mathrm{L/H}}^\epsilon\, \boldsymbol{\mathcal{M}}}{1 - \alpha^\epsilon\, \lambda_{\mathrm{L/H}}^\epsilon}, \quad \text{and} \quad \Theta_\pm^\epsilon(\boldsymbol{\mathcal{U}}, \boldsymbol{v}_h) = \boldsymbol{\mathcal{M}} \pm \frac{1}{\alpha^\epsilon}\, \mathscr{F}^\epsilon(\boldsymbol{\mathcal{U}}, \boldsymbol{v}_h)\,. \tag{88}$$

Similar to the approach in Section 5.1.1, the following two lemmas show realizability of $\Theta_\pm^\epsilon$ and $\Theta_{0,\mathrm{L/H}}^\epsilon$.

Lemma 3. *Let $\Theta_\pm^\epsilon(\boldsymbol{\mathcal{U}}, \boldsymbol{v}_h)$ be given as in Eq. (88). Assume that $\boldsymbol{\mathcal{U}} \in \mathcal{R}$. Then $\Theta_\pm^\epsilon \in \mathcal{R}$.*

Proof. The first component of $\Theta_\pm^\epsilon$ can be written as

$$\frac{1}{4\pi} \int_{\mathbb{S}^2} \mathsf{f}_\pm[\boldsymbol{v}_h, \alpha^\epsilon](\omega)\, d\omega, \quad \text{where} \quad \mathsf{f}_\pm[\boldsymbol{v}_h, \alpha^\epsilon](\omega) = \left(1 \pm Q(\boldsymbol{v}_h)/\alpha^\epsilon\right) f(\omega), \tag{89}$$

and where $Q(\boldsymbol{v}_h)$ is the quadratic form in Eq. (47). Similarly, the remaining components of $\Theta_\pm^\epsilon$ can be written as

$$\frac{1}{4\pi} \int_{\mathbb{S}^2} \mathsf{f}_\pm[\boldsymbol{v}_h, \alpha^\epsilon](\omega)\, \boldsymbol{\ell}(\omega)\, d\omega. \tag{90}$$

Since $|Q(\boldsymbol{v}_h)|/\alpha^\epsilon \leq 1$, it follows that $\mathsf{f}_\pm[\boldsymbol{v}_h, \alpha^\epsilon](\omega) \in \mathfrak{R}$ and $\Theta_\pm^\epsilon \in \mathcal{R}$. □

Lemma 4. *Consider $\Theta_{0,\mathrm{L/H}}^\epsilon(\boldsymbol{\mathcal{U}}, \boldsymbol{v}_h)$ as defined in Eq. (88). Assume that $\boldsymbol{\mathcal{U}}, \boldsymbol{\mathcal{M}} \in \mathcal{R}$, $v_h = |\boldsymbol{v}_h| < 1$, and $\eta_{\mathrm{L/H}}^\epsilon := \alpha^\epsilon\, \lambda_{\mathrm{L/H}}^\epsilon < (1 - v_h)$. Then, $\Theta_{0,\mathrm{L/H}}^\epsilon(\boldsymbol{\mathcal{U}}, \boldsymbol{v}_h) \in \mathcal{R}$.*

Proof. The first component of $\Theta_{0,\mathrm{L/H}}^\epsilon(\boldsymbol{\mathcal{U}}, \boldsymbol{v}_h)$ can be written as

$$\frac{1}{4\pi} \int_{\mathbb{S}^2} \mathsf{f}[\boldsymbol{v}_h, \eta_{\mathrm{L/H}}^\epsilon](\omega)\, d\omega, \quad \text{where} \quad \mathsf{f}[\boldsymbol{v}_h, \eta_{\mathrm{L/H}}^\epsilon](\omega) = \frac{(1 - \boldsymbol{v}_h \cdot \boldsymbol{\ell} - \eta_{\mathrm{L/H}}^\epsilon)}{(1 - \eta_{\mathrm{L/H}}^\epsilon)}\, f(\omega). \tag{91}$$

The remaining components of $\Theta_{0,\mathrm{L/H}}^\epsilon(\boldsymbol{\mathcal{U}}, \boldsymbol{v}_h)$ can be written as

$$\frac{1}{4\pi} \int_{\mathbb{S}^2} \mathsf{f}[\boldsymbol{v}_h, \eta_{\mathrm{L/H}}^\epsilon](\omega)\, \boldsymbol{\ell}(\omega)\, d\omega. \tag{92}$$

Since $v_h < 1$ and $\eta_{\mathrm{L/H}}^\epsilon < 1 - v$, we have $(1 - \boldsymbol{v}_h \cdot \boldsymbol{\ell} - \eta_{\mathrm{L/H}}^\epsilon) \geq (1 - v_h) - \eta_{\mathrm{L/H}}^\epsilon > 0$. This, together with $f \in \mathfrak{R}$, implies that $\mathsf{f}[\boldsymbol{v}_h, \eta_{\mathrm{L/H}}^\epsilon](\omega) \in \mathfrak{R}$ and $\Theta_{0,\mathrm{L/H}}^\epsilon(\boldsymbol{\mathcal{U}}, \boldsymbol{v}_h) \in \mathcal{R}$. □

Analogous to the spatial advection case, the assumptions in Lemma 4 are fulfilled for all $\boldsymbol{K} \in \mathcal{T}$ under assumptions of Proposition 3, when

$$C^\epsilon := \gamma^\epsilon\, \alpha^\epsilon\, \hat{w}_{\hat{k}} \tag{93}$$

is used in the time-step restriction (63). Under these assumptions, $\eta^{\epsilon}_{\mathrm{L/H}} := \alpha^{\epsilon}\lambda^{\epsilon}_{\mathrm{L/H}} < (1-v_h) \leq 1$. Therefore $\Phi^{\epsilon}_{\mathrm{l}}$ and $\Phi^{\epsilon}_{\hat{k}}$ are convex combinations of realizable terms, and are thus realizable. We have shown that, under the assumptions of Proposition 3, $\widehat{\boldsymbol{\mathcal{U}}}^{n+1/2,\epsilon}_{\boldsymbol{K}} \in \mathcal{R}$ for all $\boldsymbol{K} \in \mathcal{T}$.

5.1.3. Sources

The last part of the explicit update involves the source term in the number flux equation. We define

$$\begin{aligned}\widehat{\boldsymbol{\mathcal{U}}}^{n+1/2,\mathcal{S}}_{\boldsymbol{K}} &= \boldsymbol{\mathcal{U}}^{n}_{\boldsymbol{K}} + \frac{\Delta t_{\mathcal{S}}}{|\boldsymbol{K}|}\big(\,\boldsymbol{\mathcal{S}}(\boldsymbol{\mathcal{U}}^{n}_{h},\boldsymbol{v}_h)\,\big)_{\boldsymbol{K}} \quad (\Delta t_{\mathcal{S}} = \Delta t/\gamma^{\mathcal{S}}) \\ &= \frac{1}{|\boldsymbol{K}|}\int_{\boldsymbol{K}}\Big[\,\boldsymbol{\mathcal{U}}^{n}_{h} + \Delta t_{\mathcal{S}}\,\boldsymbol{\mathcal{S}}(\boldsymbol{\mathcal{U}}^{n}_{h},\boldsymbol{v}_h)\Big]\,\tau\,d\boldsymbol{z}. \end{aligned} \tag{94}$$

From the definition of the source term $\boldsymbol{\mathcal{S}}$ in Eq. (17), the number density is not affected in the source update. Thus we have $\widehat{\mathcal{N}}^{n+1/2,\mathcal{S}}_{\boldsymbol{K}} = \mathcal{N}^{n}_{\boldsymbol{K}} > 0$, which, together with the results obtained in Sections 5.1.1 and 5.1.2, concludes the proof of the first claim in Proposition 3.

Ideally, one would expect to show that $\widehat{\boldsymbol{\mathcal{U}}}^{n+1/2,\mathcal{S}}_{\boldsymbol{K}} \in \mathcal{R}$ under time-step restrictions similar to the ones in Sections 5.1.1 and 5.1.2. Unfortunately, this is not true in the three-dimensional case considered in this paper. In the rest of this section, we will show that (i) realizability of $\widehat{\boldsymbol{\mathcal{U}}}^{n+1/2,\mathcal{S}}_{\boldsymbol{K}}$ is preserved by the semi-discrete equation, i.e., without time discretization, and (ii) with the forward Euler discretization in Eq. (62a), $\widehat{\boldsymbol{\mathcal{U}}}^{n+1/2,\mathcal{S}}_{\boldsymbol{K}} \in \mathcal{R}$ in a reduced, one-dimensional planar geometry.

Proposition 7 *(Semi-discrete source update). Given a quadrature rule $\boldsymbol{Q} : C^0(\boldsymbol{K}) \to \mathbb{R}$ with positive weights and points given by the set $\boldsymbol{S}^{\boldsymbol{K}}_{\otimes}$, we show that, for all $\boldsymbol{z} \in \boldsymbol{S}^{\boldsymbol{K}}_{\otimes} \subset \boldsymbol{K}$, the solution $\boldsymbol{\mathcal{U}}_h(\boldsymbol{z},t)$ to the semi-discrete equation*

$$\partial_t \boldsymbol{\mathcal{U}}_h(\boldsymbol{z},t) = \boldsymbol{\mathcal{S}}(\boldsymbol{\mathcal{U}}_h(\boldsymbol{z},t),\boldsymbol{v}_h(\boldsymbol{z})) \tag{95}$$

remains in the realizable set $\mathcal{R}$ for all $t \geq t_0$, provided that $\boldsymbol{\mathcal{U}}_h(\boldsymbol{z},t_0)$ is realizable.

This semi-discrete equation is consistent with the source update portion in Eq. (18) and results in Eq. (94) after applying forward Euler discretization and cell-averaging.

Proof. To show that $\boldsymbol{\mathcal{U}}_h(\boldsymbol{z},t) \in \mathcal{R}$ for $t \geq t_0$, we first observe that since the first component of $\boldsymbol{\mathcal{S}}(\boldsymbol{\mathcal{U}},\boldsymbol{v})$ is zero (see Eq. (17)), the source update does not affect $\mathcal{N}_h$. Thus, showing $\boldsymbol{\mathcal{U}}_h(\boldsymbol{z},t) \in \mathcal{R}$ is equivalent to proving that $\mathcal{G}_h(\boldsymbol{z},t) \leq \mathcal{N}_h(\boldsymbol{z})$, where $\mathcal{G}_h(\boldsymbol{z},t) = |\boldsymbol{\mathcal{G}}_h(\boldsymbol{z},t)|$ with $\boldsymbol{\mathcal{G}}_h$ the number flux governed by Eq. (95). Due to the continuity of $\mathcal{G}_h(\boldsymbol{z},t)$ in time, it suffices to show that if $\mathcal{G}_h(\boldsymbol{z},\hat{t}) = \mathcal{N}_h(\boldsymbol{z})$ for some $\hat{t} \geq t_0$, then $\mathcal{G}_h(\boldsymbol{z},t) = \mathcal{N}_h(\boldsymbol{z})$ for all $t \geq \hat{t}$, i.e., the number flux magnitude does not exceed the number density. Indeed, the number flux portion of Eq. (95) is given by

$$\partial_t \mathcal{G}_{h,j} = \mathcal{Q}^{i}{}_{kj}(\partial_i v^k)_h - \mathcal{I}^{i}(\partial_i v_j)_h = \frac{1}{4\pi}\int_{\mathbb{S}^2}\big(\,\ell^i(\omega)\ell_k(\omega)\ell_j(\omega)(\partial_i v^k)_h - \ell^i(\omega)(\partial_i v_j)_h\,\big)\,f(\omega,t)\,d\omega. \tag{96}$$

Suppose $\mathcal{G}_h(\boldsymbol{z},\hat{t}) = \mathcal{N}_h(\boldsymbol{z})$ for some $\hat{t} \geq t_0$, it is known [51] that the distribution function $f(\omega)$ takes the form of a Dirac delta function, i.e., $f(\omega) = c\,\delta(\hat{\omega})$ for some $c > 0$, $\hat{\omega} \in \mathbb{S}^2$. Therefore, at $t = \hat{t}$, we have $\mathcal{G}^{j}_{h} = c\,\big(1 + v^k\ell_k(\hat{\omega})\big)\,\ell^j(\hat{\omega})$ and

$$\partial_t \mathcal{G}_{h,j} = \frac{c}{4\pi}\big(\,\ell^i(\hat{\omega})\ell_k(\hat{\omega})\ell_j(\hat{\omega})(\partial_i v^k)_h - \ell^i(\hat{\omega})(\partial_i v_j)_h\,\big)\,. \tag{97}$$

Thus,

$$\begin{aligned}\frac{1}{2}\partial_t(\mathcal{G}_h)^2 &= \mathcal{G}^{j}_{h}\partial_t\mathcal{G}_{h,j} \\ &= \frac{c^2}{4\pi}\big(\,1 + v^k\ell_k(\hat{\omega})\,\big)\,\ell^j(\hat{\omega})\big(\,\ell^i(\hat{\omega})\ell_k(\hat{\omega})\ell_j(\hat{\omega})(\partial_i v^k)_h - \ell^i(\hat{\omega})(\partial_i v_j)_h\,\big) = 0\,,\end{aligned} \tag{98}$$

where the fact $\ell^i\ell_i = 1$ is used in the last equality. Eq. (98) indicates that the number flux magnitude does not change once $\mathcal{G}_h(\boldsymbol{z},\hat{t}) = \mathcal{N}_h(\boldsymbol{z})$ for some $\hat{t} \geq t_0$ and implies that $\boldsymbol{\mathcal{U}}_h(\boldsymbol{z},t) \in \mathcal{R}$ for $t \geq t_0$. □

Remark 6. The result in Eq. (98) also explains why the discretized source update (94) cannot guarantee realizability of the updated moments. Specifically, Eq. (98) suggests that, for moments on the realizable boundary ($\mathcal{G}_h = \mathcal{N}_h$), the continuous source update (95) moves the moments tangentially with the boundary of the realizable set. Once explicit discretization is applied, e.g., Eq. (94), the update may result in unrealizable moments, regardless of the time-step size.

Next, we show that, in a one-dimensional planar geometry [2, Section 6.5], the discretized source update Eq. (94) preserves realizability of the moments when a time-step restriction is satisfied. In the planar geometry, the spatial fluxes are zero in two of the three spatial dimensions (e.g., $\partial_{x^2} f = \partial_{x^3} f = 0$) with the angular direction reduced from $\omega = (\vartheta,\varphi) \in \mathbb{S}^2$ to $\mu = \cos\vartheta \in [-1,1]$. In the

remainder of this subsection, we use $x(=x^1)$ to denote the only spatial dimension that has nonzero fluxes, and use a scalar function v to denote the velocity which varies only in the x direction. Moreover, the primitive moments in the planar geometry are given by

$$\{\mathcal{D}, \mathcal{I}, \mathcal{K}, Q\}(\varepsilon, x, t) = \frac{1}{2}\int_{-1}^{1} f(\omega, \varepsilon, x, t)\{1, \mu, \mu^2, \mu^3\}\, d\mu, \tag{99}$$

and the conserved moments are $\mathcal{N} = \mathcal{D} + v\mathcal{I}$ and $\mathcal{G} = \mathcal{I} + v\mathcal{K}$. In this case, the semi-discrete source update Eq. (95) reduces to

$$\partial_t \mathcal{N} = 0\,, \quad \text{and} \quad \partial_t \mathcal{G} = \frac{1}{2}\int_{-1}^{1} \mu(\mu^2 - 1)(\partial_x v)_h\, f(\mu) d\mu\,. \tag{100}$$

The following proposition shows that the discretized version of this source update preserves moment realizability under a time step restriction.

Proposition 8. *In the planar geometry, suppose Assumption 1 holds, $v_h < 1$, and the time step satisfies*

$$\Delta t \le \frac{1}{2}\gamma^{\mathcal{S}}\frac{1 - v_h}{|(\partial_x v)_h|}, \quad \left(\text{ i.e., } \Delta t_{\mathcal{S}} \le \frac{1}{2}\frac{1 - v_h}{|(\partial_x v)_h|}\right). \tag{101}$$

Then the discretized source update Eq. (94) gives a realizable cell-averaged moment $\widehat{\boldsymbol{\mathcal{U}}}_{\boldsymbol{K}}^{n+1/2,\mathcal{S}}$ for all $\boldsymbol{K} \in \mathcal{T}$, provided $\boldsymbol{\mathcal{U}}_h^n \in \mathcal{R}$ on all $\boldsymbol{K} \in \mathcal{T}$.

Proof. In this proof, we show the realizability of $\widehat{\boldsymbol{\mathcal{U}}}_h^{n+1/2,\mathcal{S}} := \boldsymbol{\mathcal{U}}_h^n + \Delta t_{\mathcal{S}}\,\boldsymbol{\mathcal{S}}(\boldsymbol{\mathcal{U}}_h^n, \boldsymbol{v}_h)$, which leads to the realizability of $\widehat{\boldsymbol{\mathcal{U}}}_{\boldsymbol{K}}^{n+1/2,\mathcal{S}}$ when the element integral in Eq. (94) is evaluated using quadrature rules with positive weights in both the spatial and energy dimensions.

We start with denoting $\widehat{\boldsymbol{\mathcal{U}}}_h^{n+1/2,\mathcal{S}} =: (\widehat{\mathcal{N}}_h^{n+1/2,\mathcal{S}}, \widehat{\mathcal{G}}_h^{n+1/2,\mathcal{S}})$. In the planar geometry, the number density $\widehat{\mathcal{N}}_h^{n+1/2,\mathcal{S}}$ and number flux $\widehat{\mathcal{G}}_h^{n+1/2,\mathcal{S}}$ are both scalar-valued. From Assumption 1 and the definition of the source terms $\boldsymbol{\mathcal{S}}$ in Eq. (17), we can write

$$\widehat{\mathcal{N}}_h^{n+1/2,\mathcal{S}} = \frac{1}{2}\int_{-1}^{1}\left(1 + v_h\mu\right) f(\mu)\, d\mu, \tag{102}$$

$$\begin{aligned}\widehat{\mathcal{G}}_h^{n+1/2,\mathcal{S}} &= \frac{1}{2}\int_{-1}^{1}\left[\left(1 + v_h\mu\right)\mu + \Delta t_{\mathcal{S}}\left(\mu^3(\partial_x v)_h - \mu(\partial_x v)_h\right)\right] f(\mu)\, d\mu \\ &= \frac{1}{2}\int_{-1}^{1}\left(1 + v_h\mu\right) f(\mu)\left[\mu - \Delta t_{\mathcal{S}}\,(\partial_x v)_h\mu\,\frac{1 - \mu^2}{1 + v_h\mu}\right] d\mu,\end{aligned} \tag{103}$$

where $f \in \mathfrak{R}$. Since $v_h < 1$ and $\mu \in [-1, 1]$, it is clear that $\widehat{\mathcal{N}}_h^{n+1/2,\mathcal{S}} > 0$. We next prove $\widehat{\mathcal{N}}_h^{n+1/2,\mathcal{S}} - |\widehat{\mathcal{G}}_h^{n+1/2,\mathcal{S}}| \ge 0$ when $\Delta t_{\mathcal{S}}$ satisfies Eq. (101). By Cauchy-Schwartz inequality,

$$|\widehat{\mathcal{G}}_h^{n+1/2,\mathcal{S}}|^2 \le \frac{1}{4}\int_{-1}^{1}(1 + v_h\mu) f(\mu)\, d\mu \int_{-1}^{1}(1 + v_h\mu) f(\mu)\left[\mu - \Delta t_{\mathcal{S}}\,(\partial_x v)_h\mu\,\frac{1 - \mu^2}{1 + v_h\mu}\right]^2 d\mu. \tag{104}$$

We then show that, under Eq. (101), $\left[\mu - \Delta t_{\mathcal{S}}\,(\partial_x v)_h\mu\,\frac{1-\mu^2}{1+v_h\mu}\right]^2 \le 1$ for $\mu \in [-1, 1]$. This inequality clearly holds when $\mu = \pm 1$ and $\mu = 0$. We thus focus on the case when $\mu \in (-1, 1)$ and $\mu \ne 0$. Since $\Delta t_{\mathcal{S}} > 0$ and $\frac{1-\mu^2}{1+v_h\mu} \ge 0$, the inequality holds when

$$\Delta t_{\mathcal{S}} \le \frac{1 + v_h\mu}{(\partial_x v)_h\mu\,(1 - \mu)} \text{ if } (\partial_x v)_h\,\mu > 0, \text{ and } \Delta t_{\mathcal{S}} \le \frac{1 + v_h\mu}{(-(\partial_x v)_h\mu)(1 + \mu)} \text{ if } (\partial_x v)_h\,\mu < 0. \tag{105}$$

It is straightforward to verify that Eq. (101) gives a sufficient condition to the two time-step restrictions above. □

5.2. Realizability-enforcing limiter

It has been shown in Proposition 3 that, when starting from realizable moments $\boldsymbol{\mathcal{U}}_h^n$, the explicit update in Eq. (62a) is guaranteed to provide updated cell-averaged moments $\widehat{\boldsymbol{\mathcal{U}}}_{\boldsymbol{K}}^{n+1/2}$ with number density $\widehat{\mathcal{N}}_{\boldsymbol{K}}^{n+1/2} > 0$ for every $\boldsymbol{K}$ under a reasonable time-step restriction. In this section, we discuss how the realizability-enforcing limiter proposed in [35] is used here in Eq. (62b) to enforce realizability of moments $\boldsymbol{\mathcal{U}}_h$ at a point set $\widetilde{S}_{\otimes}^{\boldsymbol{K}}$ defined in Eq. (52), which covers all DG nodal points as well as the auxiliary points in element $\boldsymbol{K}$.

In [35], the realizability-enforcing limiter was formulated following the approach considered in [55,56] for constructing bound-preserving limiters for high-order DG schemes. The limiter enforces moment realizability at each quadrature point in a DG element by

relaxing unrealizable moments towards the realizable cell-averaged moments. Specifically, this limiter replaces unrealizable moments with their convex combinations with the cell-averaged moment, which preserves the Eulerian-frame particle number in each element (but not the energy; see Section 6.2 for further discussions) when the same convex combination factor ($\theta^{\mathcal{N}}_{\boldsymbol{K}}$ and $\theta^{\boldsymbol{\mathcal{U}}}_{\boldsymbol{K}}$) is applied to all moments within the element. For completeness, the steps taken in this realizability-enforcing limiter are summarized in Algorithm 1. We refer to [35] and references therein for detailed discussions.

Algorithm 1: $\boldsymbol{\mathcal{U}}_h = \texttt{RealizabilityLimiter}(\widehat{\boldsymbol{\mathcal{U}}}_h)$.

Inputs: Discretized moments $\widehat{\boldsymbol{\mathcal{U}}}_h$ with $\widehat{\mathcal{N}}_{\boldsymbol{K}} > 0$ for all $\boldsymbol{K} \in \mathcal{T}$.
Parameter: $0 < \delta \ll 1$.
for *each element* $\boldsymbol{K}$ **do**
  if $\widehat{\boldsymbol{\mathcal{U}}}_{\boldsymbol{K}} \in \mathcal{R}$ **then**
    /* limit number density */
    $\tilde{\mathcal{N}}_h \leftarrow \theta^{\mathcal{N}}_{\boldsymbol{K}} \widehat{\mathcal{N}}_h + (1-\theta^{\mathcal{N}}_{\boldsymbol{K}}) \widehat{\mathcal{N}}_{\boldsymbol{K}}$ with $\theta^{\mathcal{N}}_{\boldsymbol{K}} \leftarrow \min\{\frac{\widehat{\mathcal{N}}_{\boldsymbol{K}}}{\widehat{\mathcal{N}}_{\boldsymbol{K}} - \min_{\boldsymbol{z}\in\widetilde{S}^{\boldsymbol{K}}_{\otimes}} \widehat{\mathcal{N}}_h(\boldsymbol{z})}, 1\}$;
    /* build intermediate moments */
    $\tilde{\boldsymbol{\mathcal{U}}}_h \leftarrow (\tilde{\mathcal{N}}_h, \widehat{\mathcal{G}}_h)$;
    /* limit full moments */
    $\boldsymbol{\mathcal{U}}_h \leftarrow \theta^{\boldsymbol{\mathcal{U}}}_{\boldsymbol{K}} \tilde{\boldsymbol{\mathcal{U}}}_h + (1-\theta^{\boldsymbol{\mathcal{U}}}_{\boldsymbol{K}}) \widehat{\boldsymbol{\mathcal{U}}}_{\boldsymbol{K}}$ where
    $\theta^{\boldsymbol{\mathcal{U}}}_{\boldsymbol{K}} \leftarrow \arg\min_\theta \{\theta \in [0,1] : \gamma(\theta \tilde{\boldsymbol{\mathcal{U}}}_h(\boldsymbol{z}) + (1-\theta)\widehat{\boldsymbol{\mathcal{U}}}_{\boldsymbol{K}}) \geq 0, \forall \boldsymbol{z} \in \widetilde{S}^{\boldsymbol{K}}_{\otimes}\}$ with γ defined in Eq. (22).
  else
    /* replace number densities with the cell average and shrink number fluxes accordingly */
    $\boldsymbol{\mathcal{U}}_h \leftarrow (\mathcal{N}_h, \mathcal{G}_h)$ with $\mathcal{N}_h = \widehat{\mathcal{N}}_{\boldsymbol{K}}$ and $\mathcal{G}_h = (1-\delta)\widehat{\mathcal{N}}_{\boldsymbol{K}} \frac{\widehat{\mathcal{G}}_h}{|\widehat{\mathcal{G}}_h|}$;

As seen in Algorithm 1, starting from discretized moment $\widehat{\boldsymbol{\mathcal{U}}}_h$ with positive cell-averaged number density $\widehat{\mathcal{N}}_{\boldsymbol{K}}$, the limiter enforces realizability of the resulting moments $\boldsymbol{\mathcal{U}}_h$ in the point set $\widetilde{S}^{\boldsymbol{K}}_{\otimes}$ by limiting toward the cell-averaged moments. The limiter is guaranteed to provide realizable outputs at the point set when the starting moment has a positive cell-averaged number density, thus Proposition 4 holds.

We note that, when approximate closures are considered, the explicit update may not result in moments with positive cell-averaged number density (since Assumption 1 does not hold). If a negative cell-averaged number density is observed in element $\boldsymbol{K}$, we set the moments in $\boldsymbol{K}$ to be an isotropic moment with close to zero but positive number density and zero number flux. This safeguard affects the conservation property of the scheme, however, we do not observe a negative cell-averaged number density in any of the numerical experiments presented in Section 8.

5.3. Conversion between conserved and primitive moments

In this section, we prove Proposition 5 by showing that, under Assumption 1 and assuming $v_h < 1$, (i) the conversion between conserved and primitive moments preserves realizability and (ii) the iterative solver in Eq. (57) is guaranteed to converge to a unique $\boldsymbol{\mathcal{M}} \in \mathcal{R}$ that satisfies Eq. (19) given $\boldsymbol{\mathcal{U}} \in \mathcal{R}$.

In the following two lemmas, we show that the realizability is preserved in the conversion between conserved and primitive moments.

Lemma 5. *Suppose Assumption 1 holds and $v < 1$. Let $\boldsymbol{\mathcal{U}}$ be given as in Eq. (19) with $\boldsymbol{\mathcal{M}} \in \mathcal{R}$, then $\boldsymbol{\mathcal{U}} \in \mathcal{R}$.*

Proof. Let $f \in \mathfrak{R}$ be the underlying distribution for $\boldsymbol{\mathcal{M}} \in \mathcal{R}$. Then, from Eq. (19), the components of $\boldsymbol{\mathcal{U}}$ can be written as

$$\left(\mathcal{N}, \mathcal{G}_j\right)^{\intercal} = \frac{1}{4\pi} \int_{\mathbb{S}^2} \left(1 + v^i \ell_i(\omega)\right) f(\omega) \left(1, \ell_j(\omega)\right)^{\intercal} d\omega := \frac{1}{4\pi} \int_{\mathbb{S}^2} \mathsf{f}(\omega) \left(1, \ell_j(\omega)\right)^{\intercal} d\omega. \tag{106}$$

Since $f \in \mathfrak{R}$ and $v^i \ell_i \in (-1, 1)$, it follows that $\mathsf{f}(\omega) := \left(1 + v^i \ell_i(\omega)\right) f(\omega) \in \mathfrak{R}$ and thus $\boldsymbol{\mathcal{U}} \in \mathcal{R}$. □

Lemma 6. *Suppose Assumption 1 holds, $v < 1$, and $\boldsymbol{\mathcal{U}} \in \mathcal{R}$. Then there exists some $\boldsymbol{\mathcal{M}} \in \mathcal{R}$ that satisfies Eq. (19).*

Proof. Let $\mathsf{f} \in \mathfrak{R}$ denote the underlying distribution for $\boldsymbol{\mathcal{U}} \in \mathcal{R}$. Then the components of $\boldsymbol{\mathcal{U}}$ can be written as

$$\left(\mathcal{N}, \mathcal{G}_j\right)^{\intercal} = \frac{1}{4\pi} \int_{\mathbb{S}^2} \mathsf{f}(\omega) \left(1, \ell_j(\omega)\right)^{\intercal} d\omega. \tag{107}$$

Since $\mathsf{f} \in \mathfrak{R}$ and $v^i \ell_i \in (-1, 1)$, it follows that $f(\omega) := \left(1 + v^i \ell_i(\omega)\right)^{-1} \mathsf{f}(\omega) \in \mathfrak{R}$. Taking the moments of f leads to $\boldsymbol{\mathcal{M}} \in \mathcal{R}$. Using the relation between f and f it is then straightforward to verify that $\boldsymbol{\mathcal{M}}$ satisfies Eq. (19). □

Lemma 6 shows the existence of realizable primitive moments corresponding to given conserved moments. However, it does not provide guarantees on the convergence of the iterative solver we use to find the primitive moments. In the remainder of this subsection, we prove that the iterative solver in Eq. (57) guarantees the convergence to a realizable moment $\boldsymbol{\mathcal{M}}$. To start, in the following lemma we show that realizability is guaranteed at each iteration of the solver in Eq. (57).

Lemma 7. *Let $\boldsymbol{\mathcal{U}} \in \mathcal{R}$ and $\lambda \leq \frac{1}{1+v}$ in Eq. (56). Then, the solver in Eq. (57) guarantees that $\boldsymbol{\mathcal{M}}^{[k+1]} = (\mathcal{D}^{[k+1]}, \mathcal{I}^{[k+1]})^\intercal \in \mathcal{R}$, provided that $\boldsymbol{\mathcal{M}}^{[k]} = (\mathcal{D}^{[k]}, \mathcal{I}^{[k]})^\intercal \in \mathcal{R}$.*

Proof. We write the iterative update in Eq. (57) as

$$\begin{aligned}\boldsymbol{\mathcal{M}}^{[k+1]} = \begin{pmatrix} \mathcal{D}^{[k+1]} \\ \mathcal{I}_j^{[k+1]} \end{pmatrix} &= (1-\lambda) \begin{pmatrix} \mathcal{D}^{[k]} - \frac{\lambda}{1-\lambda} v^i \mathcal{I}_i^{[k]} \\ \mathcal{I}_j^{[k]} - \frac{\lambda}{1-\lambda} v^i \mathsf{k}_{ij}^{[k]} \mathcal{D}^{[k]} \end{pmatrix} + \lambda \begin{pmatrix} \mathcal{N} \\ \mathcal{G}_j \end{pmatrix} \\ &=: (1-\lambda) \widetilde{\boldsymbol{\mathcal{M}}}^{[k]} + \lambda \boldsymbol{\mathcal{U}}.\end{aligned} \tag{108}$$

Since the realizable set $\mathcal{R}$ is convex and $\boldsymbol{\mathcal{M}}^{[k+1]}$ is a convex combination of $\widetilde{\boldsymbol{\mathcal{M}}}^{[k]}$ and $\boldsymbol{\mathcal{U}} \in \mathcal{R}$, it suffices to show that $\widetilde{\boldsymbol{\mathcal{M}}}^{[k]} \in \mathcal{R}$. We observe that the entries in $\widetilde{\boldsymbol{\mathcal{M}}}^{[k]}$ take the exact same form as the ones on the right-hand side of Eq. (19), except with $\boldsymbol{v}$ replaced by $-\frac{\lambda}{1-\lambda}\boldsymbol{v}$. It then follows from Lemma 5 that $\widetilde{\boldsymbol{\mathcal{M}}}^{[k]} \in \mathcal{R}$ if $\frac{\lambda}{1-\lambda} v \leq 1$, i.e., $\lambda \leq \frac{1}{1+v}$. □

It is well-known that, when solving a fixed-point problem defined by a contraction operator, the Picard iteration converges to the unique fixed point (see, e.g., [54]). We show below in Proposition 9 that the fixed-point operator $\boldsymbol{\mathcal{H}}_{\boldsymbol{\mathcal{U}}}$ defined in Eq. (56) is a contraction under mild assumptions on v_h, which thus guarantees the convergence of the iterative solver in Eq. (57). The proof of Proposition 9 uses results from the following two technical lemmas.

Lemma 8. *For any $\boldsymbol{\mathcal{M}} \in \mathcal{R}$, $\|\partial_{\mathcal{D}}(v^i \mathsf{k}_{ij} \mathcal{D})\| \leq v$.*

Proof. See Appendix B.2 for the proof. □

Lemma 9. *For any $\boldsymbol{\mathcal{M}} \in \mathcal{R}$, $\|\nabla_{\mathcal{I}}(v^i \mathsf{k}_{ij} \mathcal{D})\| \leq 2v$.*

Proof. See Appendix B.3 for the proof. □

We now state and prove Proposition 9.

Proposition 9. *Suppose $v < \sqrt{2} - 1$ and $\lambda \in (0, 1]$. Then, $\boldsymbol{\mathcal{H}}_{\boldsymbol{\mathcal{U}}}$ defined in Eq. (56) is a contraction operator, i.e., there exists some $L < 1$ such that*

$$\|\boldsymbol{\mathcal{H}}_{\boldsymbol{\mathcal{U}}}(\boldsymbol{\mathcal{M}}^{(1)}) - \boldsymbol{\mathcal{H}}_{\boldsymbol{\mathcal{U}}}(\boldsymbol{\mathcal{M}}^{(2)})\| \leq L \|\boldsymbol{\mathcal{M}}^{(1)} - \boldsymbol{\mathcal{M}}^{(2)}\|, \quad \forall \boldsymbol{\mathcal{M}}^{(1)}, \boldsymbol{\mathcal{M}}^{(2)} \in \mathcal{R}. \tag{109}$$

Proof. First, for convenience, we denote $\Delta\mathcal{D} = \mathcal{D}^{(1)} - \mathcal{D}^{(2)}$ and $\Delta\mathcal{I}_j = \mathcal{I}_j^{(1)} - \mathcal{I}_j^{(2)}$. It then follows from the definition of $\boldsymbol{\mathcal{H}}_{\boldsymbol{\mathcal{U}}}$ and the triangle inequality that

$$\|\boldsymbol{\mathcal{H}}_{\boldsymbol{\mathcal{U}}}(\boldsymbol{\mathcal{M}}^{(1)}) - \boldsymbol{\mathcal{H}}_{\boldsymbol{\mathcal{U}}}(\boldsymbol{\mathcal{M}}^{(2)})\| \leq (1-\lambda) \left\| \begin{pmatrix} \Delta\mathcal{D} \\ \Delta\mathcal{I}_j \end{pmatrix} \right\| + \lambda \left\| \begin{pmatrix} v^i \Delta\mathcal{I}_i \\ v^i (\mathsf{k}_{ij}^{(1)} \mathcal{D}^{(1)} - \mathsf{k}_{ij}^{(2)} \mathcal{D}^{(2)}) \end{pmatrix} \right\| \tag{110}$$

Thus, it suffices to show that, there exists some $\tilde{L} < 1$ such that

$$\left\| \begin{pmatrix} v^i \Delta\mathcal{I}_i \\ v^i (\mathsf{k}_{ij}^{(1)} \mathcal{D}^{(1)} - \mathsf{k}_{ij}^{(2)} \mathcal{D}^{(2)}) \end{pmatrix} \right\| \leq \tilde{L} \left\| \begin{pmatrix} \Delta\mathcal{D} \\ \Delta\mathcal{I}_j \end{pmatrix} \right\|, \quad \forall \boldsymbol{\mathcal{M}}^{(1)}, \boldsymbol{\mathcal{M}}^{(2)} \in \mathcal{R}. \tag{111}$$

Lemmas 8 and 9 imply that the gradients of $v^i \mathsf{k}_{ij} \mathcal{D}$ in the $\mathcal{D}$ and $\mathcal{I}$ directions are bounded. Thus, we have

$$\begin{aligned}\|v^i (\mathsf{k}_{ij}^{(1)} \mathcal{D}^{(1)} - \mathsf{k}_{ij}^{(2)} \mathcal{D}^{(2)})\| &\leq \|\partial_{\mathcal{D}}(v^i \mathsf{k}_{ij} \mathcal{D})\| \|\Delta\mathcal{D}\| + \|\nabla_{\mathcal{I}}(v^i \mathsf{k}_{ij} \mathcal{D})\| \|\Delta\mathcal{I}_j\| \\ &\leq v \|\Delta\mathcal{D}\| + 2v \|\Delta\mathcal{I}_j\|,\end{aligned} \tag{112}$$

which leads to

$$\begin{aligned}\|v^i (\mathsf{k}_{ij}^{(1)} \mathcal{D}^{(1)} - \mathsf{k}_{ij}^{(2)} \mathcal{D}^{(2)})\|^2 &\leq v^2 \|\Delta\mathcal{D}\|^2 + 4v^2 \|\Delta\mathcal{D}\| \|\Delta\mathcal{I}_j\| + 4v^2 \|\Delta\mathcal{I}_j\|^2 \\ &\leq (3 + 2\sqrt{2}) v^2 \|\Delta\mathcal{D}\|^2 + (2 + 2\sqrt{2}) v^2 \|\Delta\mathcal{I}_j\|^2,\end{aligned} \tag{113}$$

where the second inequality follows from the inequality, $2ab \le (\sqrt{2}+1)a^2 + (\sqrt{2}-1)b^2$, with $a = \sqrt{2}v\|\Delta\mathcal{D}\|$ and $b = \sqrt{2}v\|\Delta\mathcal{I}_j\|$. Taking the square of the left-hand side in Eq. (111) and applying the inequality in Eq. (113) gives

$$\begin{aligned}\left\| \begin{pmatrix} v^i \Delta\mathcal{I}_i \\ v^i(\mathsf{k}_{ij}^{(1)}\mathcal{D}^{(1)} - \mathsf{k}_{ij}^{(2)}\mathcal{D}^{(2)}) \end{pmatrix} \right\|^2 &= v^2\|\Delta\mathcal{I}_j\|^2 + \|v^i(\mathsf{k}_{ij}^{(1)}\mathcal{D}^{(1)} - \mathsf{k}_{ij}^{(2)}\mathcal{D}^{(2)})\|^2 \\ &\le (3+2\sqrt{2})v^2(\|\Delta\mathcal{D}\|^2 + \|\Delta\mathcal{I}_j\|^2)\,. \end{aligned} \tag{114}$$

Let $\tilde{L} := \sqrt{3+2\sqrt{2}}\,v$, the claim then holds when $v < \left(\sqrt{3+2\sqrt{2}}\right)^{-1} = \sqrt{2}-1$. □

Theorem 2. *Suppose $v < \sqrt{2}-1$ and $\lambda \le \frac{1}{1+v}$ in Eq. (56). Then, for any given conserved moment $\boldsymbol{\mathcal{U}} = (\mathcal{N}, \boldsymbol{\mathcal{G}})^\intercal \in \mathcal{R}$ and initial primitive moment $\boldsymbol{\mathcal{M}}^{[0]} \in \mathcal{R}$, the iterative solver in Eq. (57) converges to the unique realizable primitive moment $\boldsymbol{\mathcal{M}}$ that satisfies Eq. (19) as $k \to \infty$.*

Proof. This theorem is a direct consequence from the realizability-preserving property of the solver shown in Lemma 7 and the contraction property of $\mathcal{H}_{\boldsymbol{\mathcal{U}}}$ proved in Proposition 9. □

The results in Theorem 2 lead to the following corollary on the uniqueness of realizable primitive moments associated with realizable conserved moments.

Corollary 1. *Suppose $v < \sqrt{2}-1$. For any conserved moment $\boldsymbol{\mathcal{U}} \in \mathcal{R}$, there exists a unique realizable primitive moment $\boldsymbol{\mathcal{M}}$ that satisfies Eq. (19).*

5.4. Implicit collision update

In this section, we prove Proposition 6, which states that, under Assumption 1, the implicit update in Eq. (62c) preserves realizability when the iterative solver in Eq. (61) is used. We first show in the following lemma that realizability is preserved in each iteration when starting from a realizable moment.

Lemma 10. *Let $\boldsymbol{\mathcal{U}}^{(*)} = (\mathcal{N}^{(*)}, \boldsymbol{\mathcal{G}}^{(*)})^\intercal \in \mathcal{R}$, $\lambda \le \frac{1}{1+v}$, and $\kappa \ge \chi \ge 0$ in Eq. (60). Then, the solver in Eq. (61) guarantees that $\boldsymbol{\mathcal{M}}^{[k+1]} = (\mathcal{D}^{[k+1]}, \mathcal{I}^{[k+1]})^\intercal \in \mathcal{R}$, provided that $\boldsymbol{\mathcal{M}}^{[k]} = (\mathcal{D}^{[k]}, \mathcal{I}^{[k]})^\intercal \in \mathcal{R}$.*

Proof. We follow the approach in the proof of Lemma 7 and write Eq. (61) as

$$\begin{aligned}\boldsymbol{\mathcal{M}}^{[k+1]} &= (1-\lambda)\,\Lambda \begin{pmatrix} \mathcal{D}^{[k]} - \frac{\lambda}{1-\lambda} v^i \mathcal{I}_i^{[k]} \\ \mathcal{I}_j^{[k]} - \frac{\lambda}{1-\lambda} v^i \mathsf{k}_{ij}^{[k]} \mathcal{D}^{[k]} \end{pmatrix} + \lambda\,\Lambda \begin{pmatrix} \mathcal{N}^{(*)} + \Delta t \chi\, \mathcal{D}_0 \\ \mathcal{G}_j^{(*)} \end{pmatrix} \\ &=: (1-\lambda)\,\Lambda\,\tilde{\boldsymbol{\mathcal{M}}}^{[k]} + \lambda\,\Lambda\,\tilde{\boldsymbol{\mathcal{U}}}^{(*)}. \end{aligned} \tag{115}$$

In the proof of Lemma 7, we have shown that $\tilde{\boldsymbol{\mathcal{M}}}^{[k]} \in \mathcal{R}$ when $\lambda \le \frac{1}{1+v}$. Also, it is clear that $\tilde{\boldsymbol{\mathcal{U}}}^{(*)} \in \mathcal{R}$ because $\boldsymbol{\mathcal{U}}^{(*)} \in \mathcal{R}$, $\chi \ge 0$ and $\mathcal{D}_0 \ge 0$. Since $\kappa \ge \chi \ge 0$, we have $\mu_\chi \ge \mu_\kappa \ge 0$, implying that $\Lambda\boldsymbol{\mathcal{M}} \in \mathcal{R}$ for all $\boldsymbol{\mathcal{M}} \in \mathcal{R}$ based on the definition of $\mathcal{R}$ in Eq. (22). Therefore, $\Lambda\tilde{\boldsymbol{\mathcal{M}}}^{[k]}$ and $\Lambda\tilde{\boldsymbol{\mathcal{U}}}^{(*)}$ are both realizable, which, together with the convexity of $\mathcal{R}$, completes the proof. □

Similar to the moment conversion problem considered in Section 5.3, we next show in the following proposition that the operator $\boldsymbol{\mathcal{Q}}$ in Eq. (60) is a contraction, which implies convergence of the Picard iteration method in Eq. (61).

Proposition 10. *Suppose $v < \sqrt{2}-1$, $\lambda \in (0,1]$, and $\kappa \ge \chi \ge 0$. Then, $\boldsymbol{\mathcal{Q}}$ is a contraction operator, i.e., there exists some $L < 1$ such that*

$$\|\boldsymbol{\mathcal{Q}}(\boldsymbol{\mathcal{M}}^{(1)}) - \boldsymbol{\mathcal{Q}}(\boldsymbol{\mathcal{M}}^{(2)})\| \le L\|\boldsymbol{\mathcal{M}}^{(1)} - \boldsymbol{\mathcal{M}}^{(2)}\|\,, \quad \forall \boldsymbol{\mathcal{M}}^{(1)}, \boldsymbol{\mathcal{M}}^{(2)} \in \mathcal{R}\,. \tag{116}$$

Proof. From the definitions of $\mathcal{H}_{\boldsymbol{\mathcal{U}}}$ and $\boldsymbol{\mathcal{Q}}$ in Eqs. (56) and (60), we observe that

$$\begin{aligned}\|\boldsymbol{\mathcal{Q}}(\boldsymbol{\mathcal{M}}^{(1)}) - \boldsymbol{\mathcal{Q}}(\boldsymbol{\mathcal{M}}^{(2)})\| &= \|\Lambda(\mathcal{H}_{\boldsymbol{\mathcal{U}}}(\boldsymbol{\mathcal{M}}^{(1)}) - \mathcal{H}_{\boldsymbol{\mathcal{U}}}(\boldsymbol{\mathcal{M}}^{(2)}))\| \\ &\le \|\Lambda\|\|\mathcal{H}_{\boldsymbol{\mathcal{U}}}(\boldsymbol{\mathcal{M}}^{(1)}) - \mathcal{H}_{\boldsymbol{\mathcal{U}}}(\boldsymbol{\mathcal{M}}^{(2)})\| \end{aligned} \tag{117}$$

for all $\boldsymbol{\mathcal{M}}^{(1)}, \boldsymbol{\mathcal{M}}^{(2)} \in \mathcal{R}$. Since $\kappa \ge \chi \ge 0$ and $\lambda > 0$, we have $0 \le \mu_\kappa \le \mu_\chi \le 1$, i.e., $\|\Lambda\| \le 1$. The claim is thus a direct consequence of Proposition 9. □

Theorem 3. *Suppose* $\upsilon < \sqrt{2}-1$, $\lambda \leq \frac{1}{1+\upsilon}$, *and* $\kappa \geq \chi \geq 0$ *in Eq.* (60). *Then, for any given conserved moment* $\boldsymbol{\mathcal{U}}^{(*)} = (\mathcal{N}^{(*)}, \mathcal{G}^{(*)})^\intercal \in \mathcal{R}$ *and initial primitive moment* $\boldsymbol{\mathcal{M}}^{[0]} \in \mathcal{R}$, *the iterative solver in Eq.* (61) *converges to a unique realizable primitive moment* $\boldsymbol{\mathcal{M}}$ *as* $k \to \infty$. *Further, the conserved moment* $\boldsymbol{\mathcal{U}}$ *associated to* $\boldsymbol{\mathcal{M}}$ *via Eq.* (19) *is also realizable and solves the implicit system in Eq.* (58).

Proof. The convergence to a unique $\boldsymbol{\mathcal{M}} \in \mathcal{R}$ is given by the realizability-preserving property in Lemma 10 and the contraction property in Proposition 10. Realizability of $\boldsymbol{\mathcal{U}}$ follows from Lemma 5, and the formulation of the fixed-point problem in Eq. (60) guarantees that $\boldsymbol{\mathcal{U}}$ solves the implicit system in Eq. (58). □

5.5. Extension to approximate moment closures

In the earlier sections, we have shown the realizability-preserving property of the numerical scheme (62a)–(62c) under Assumption 1, in which the use of exact moment closures is assumed. As discussed in Sections 2 and 3, the approximate Minerbo closure is often used in practice to reduce the computational cost, where the approximate Eddington factor ψ_{a} and heat-flux factor ζ_{a}, defined respectively in Eqs. (27) and (28), are considered. In this section, we show that the realizability-preserving and convergence analyses for the conserved-to-primitive moment conversion (Eq. (19)) and the implicit update (Eq. (62c)) given in Sections 5.3 and 5.4 can be extended to the case when the approximate Minerbo closure is used.

When the approximate Eddington factor ψ_{a} in Eq. (27) is used, we replace Lemma 5 with the following lemma.

Lemma 11. *Suppose* $\upsilon < 1$. *Let* $\boldsymbol{\mathcal{U}}$ *be given as in Eq.* (19) *with* $\boldsymbol{\mathcal{M}} \in \mathcal{R}$, *then* $\boldsymbol{\mathcal{U}} \in \mathcal{R}$.

Proof. Since $\boldsymbol{\mathcal{M}} = \left(\mathcal{D}, \mathcal{I}\right)^\intercal \in \mathcal{R}$, we know that $\mathcal{D} > 0$ and $\mathcal{D} - \mathcal{I} \geq 0$. To show $\boldsymbol{\mathcal{U}} = \left(\mathcal{N}, \mathcal{G}\right)^\intercal \in \mathcal{R}$, we first prove $\mathcal{N} > 0$. By definition,

$$\mathcal{N} = \mathcal{D} + v^i \mathcal{I}_i \geq \mathcal{D} - \upsilon \mathcal{I} > \mathcal{D} - \mathcal{I} \geq 0\,, \tag{118}$$

where the Cauchy-Schwartz inequality and the assumption that $\upsilon < 1$ are used. We next prove that $\mathcal{N}^2 - \mathcal{G}^2 \geq 0$ with $\mathcal{G} := |\mathcal{G}|$, which implies $\mathcal{N} - \mathcal{G} \geq 0$. Writing $\mathcal{N}^2$ and $\mathcal{G}^2$ in terms of the primitive moments leads to

$$\begin{aligned}
\mathcal{N}^2 &= \mathcal{D}^2 + 2(v^i \mathcal{I}_i)\mathcal{D} + (v^i \mathcal{I}_i)^2\,, \\
&= \mathcal{D}^2\left(1 + 2\, v^i\, \hat{n}_i h + (v^i\, \hat{n}_i)^2 h^2\right), \\
\mathcal{G}^2 &= \mathcal{I}^2 + 2\mathcal{I}^j (v^i\, \mathsf{k}_{ij})\mathcal{D} + (v_\ell\, \mathsf{k}^{\ell j})(v^i\, \mathsf{k}_{ij})\mathcal{D}^2\,. \\
&= \mathcal{D}^2\left(h^2 + 2\, \hat{n}^j\, v^i\, \mathsf{k}_{ij} h + v_\ell\, \mathsf{k}^{\ell j}\, v^i\, \mathsf{k}_{ij}\right).
\end{aligned} \tag{119}$$

Using the definition of k_{ij} in Eq. (10) we obtain

$$\hat{n}^j\, v^i\, \mathsf{k}_{ij} = \hat{n}^j\, v^i\, \frac{1}{2}\left((1-\psi_{\mathsf{a}})\delta_{ij} + (3\psi_{\mathsf{a}} - 1)\hat{n}_i \hat{n}_j\right) = \psi_{\mathsf{a}}\, v^i\, \hat{n}_i\,, \tag{120}$$

$$v_\ell\, \mathsf{k}^{\ell j}\, v^i\, \mathsf{k}_{ij} = \frac{1}{4}\left((1-\psi_{\mathsf{a}})^2 v^2 + (1+\psi_{\mathsf{a}})(3\psi_{\mathsf{a}} - 1)(v_i\, \hat{n}^i)^2\right). \tag{121}$$

Plugging these terms into Eq. (119), denoting $s := v^i \hat{n}_i$, and using the assumption that $\upsilon < 1$ leads to a sufficient condition for $\mathcal{N}^2 - \mathcal{G}^2 \geq 0$: $\forall s \in [-1, 1]$ and $\forall h \in [0, 1]$,

$$\left(1 - h^2 - \frac{1}{4}(1-\psi_{\mathsf{a}})^2\right) + 2(1-\psi_{\mathsf{a}})sh + \left(h^2 - \frac{1}{4}(1+\psi_{\mathsf{a}})(3\psi_{\mathsf{a}} - 1)\right)s^2 \geq 0\,. \tag{122}$$

From Lemma 12 (e), we have $1 - \psi_{\mathsf{a}} \geq 0$. Thus, by applying the inequality $2sh \geq -1 - s^2 h^2$ to the second term above, it suffices to show that

$$\left[\psi_{\mathsf{a}} - h^2 - \frac{1}{4}(1-\psi_{\mathsf{a}})^2\right] + \left[\psi_{\mathsf{a}} h^2 - \frac{1}{4}(1+\psi_{\mathsf{a}})(3\psi_{\mathsf{a}} - 1)\right]s^2 \geq 0\,. \tag{123}$$

Since $\psi_{\mathsf{a}} - h^2 - \frac{1}{4}(1-\psi_{\mathsf{a}})^2 \geq 0$ (Lemma 12 (f)) and $s^2 \in [0, 1]$,

$$\left[\psi_{\mathsf{a}} - h^2 - \frac{1}{4}(1-\psi_{\mathsf{a}})^2 + \psi_{\mathsf{a}} h^2 - \frac{1}{4}(1+\psi_{\mathsf{a}})(3\psi_{\mathsf{a}} - 1)\right]s^2 \geq 0\,, \tag{124}$$

which then becomes

$$(\psi_{\mathsf{a}} - h^2)(1-\psi_{\mathsf{a}})s^2 \geq 0\,. \tag{125}$$

With $h^2 \leq \psi_{\mathsf{a}} \leq 1$ from Lemma 12 (e), the proof is complete. □

Lemma 11 extends Lemma 5 by showing that the mapping from primitive moments to conserved moments via Eq. (19) preserves realizability even when the approximate Minerbo closure is used. However, there is not an analogous extension of Lemma 6 to the case of approximate closures. To show that the map from conserved to primitive moments is realizability-preserving with the

approximate closure, we verify that the analysis from Lemma 7 to Corollary 1 is still valid when the approximate closure is considered. Specifically, when ψ is replaced by ψ_{a}, the result of Lemma 7 can be obtained by invoking Lemma 11 rather than Lemma 5 in the proof, the results of Lemmas 8 and 9 hold since it is shown in Appendix B.1 that ψ_{a} also satisfies the required properties of ψ, and the remainder of the analysis stays identical to the exact closure case considered in Section 5.3. Therefore, we have shown that, when ψ is replaced by ψ_{a}, the iterative solver in Eq. (57) converges to the unique, realizable primitive moment that satisfies Eq. (19) for the given conserved moment, which implies that the conserved to primitive moment map from Eq. (19) still preserves realizability when $\upsilon < \sqrt{2} - 1$. Further, we also verified that the convergence and realizability-preserving properties of the iterative solver in Eq. (61) for the implicit system in Eq. (60) given in Theorem 3 also hold in the approximate closure case by applying the same arguments to the analysis in Section 5.4.

6. Conservation property

6.1. Simultaneous number and energy conservation of the DG scheme

It has been shown in Proposition 1 that the two-moment model in Eqs. (2)–(3) conserves the Eulerian-frame energy up to $O(\upsilon)$. In this section, we discuss the simultaneous Eulerian-frame number and energy conservation properties of the two-moment model with the discontinuous Galerkin phase-space discretization presented in Section 4.1. We are primarily concerned with consistency with the Eulerian-frame energy equation for the phase-space advection problem. For this reason, we consider the collisionless case.

Eulerian-frame number conservation follows from the first component of the semi-discrete DG scheme in Eq. (36) (treating the general case with $d_{\boldsymbol{x}} = 3$),

$$\begin{aligned}\left(\partial_t \mathcal{N}_h, \varphi_h\right)_{\boldsymbol{K}} = &-\sum_{i=1}^{3} \int_{\tilde{\boldsymbol{K}}^i} \left[\widehat{\mathcal{F}^i_{\mathcal{N}}}\left(\boldsymbol{\mathcal{U}}_h, \boldsymbol{v}_h\right) \varphi_h|_{x^i_{\mathrm{H}}} - \widehat{\mathcal{F}^i_{\mathcal{N}}}\left(\boldsymbol{\mathcal{U}}_h, \boldsymbol{v}_h\right) \varphi_h|_{x^i_{\mathrm{L}}}\right] \tau \, d\tilde{\boldsymbol{z}}^i + \sum_{i=1}^{3} \left(\mathcal{F}^i_{\mathcal{N}}(\boldsymbol{\mathcal{U}}_h, \boldsymbol{v}_h), \partial_i \varphi_h\right)_{\boldsymbol{K}} \\ &- \int_{\boldsymbol{K}_{\boldsymbol{x}}} \left[\varepsilon^3 \widehat{\mathcal{F}^{\varepsilon}_{\mathcal{N}}}\left(\boldsymbol{\mathcal{U}}_h, \boldsymbol{v}_h\right) \varphi_h|_{\varepsilon_{\mathrm{H}}} - \varepsilon^3 \widehat{\mathcal{F}^{\varepsilon}_{\mathcal{N}}}\left(\boldsymbol{\mathcal{U}}_h, \boldsymbol{v}_h\right) \varphi_h|_{\varepsilon_{\mathrm{L}}}\right] d\boldsymbol{x} + \left(\varepsilon \mathcal{F}^{\varepsilon}_{\mathcal{N}}(\boldsymbol{\mathcal{U}}_h, \boldsymbol{v}_h), \partial_{\varepsilon} \varphi_h\right)_{\boldsymbol{K}},\end{aligned} \tag{126}$$

where $\mathcal{F}^i_{\mathcal{N}}$ and $\mathcal{F}^{\varepsilon}_{\mathcal{N}}$, respectively, are the first component of the position and energy space fluxes, defined in Eq. (16), and $\widehat{\mathcal{F}^i_{\mathcal{N}}}$ and $\widehat{\mathcal{F}^{\varepsilon}_{\mathcal{N}}}$ are the corresponding numerical fluxes, defined in Eqs. (41) and (45). Setting $\varphi_h = 1$ as the test function in Eq. (126) results in the equation for the element-integrated Eulerian-frame number density. Note that the volume terms (the second and fourth terms) on the right-hand side of Eq. (126) vanish when $\varphi_h = 1$. Then, because the numerical fluxes $\widehat{\mathcal{F}^i_{\mathcal{N}}}$ and $\widehat{\mathcal{F}^{\varepsilon}_{\mathcal{N}}}$ are continuous on element interfaces, summation over all phase-space elements $\boldsymbol{K} \in D$ results in cancellation of all interior fluxes, and the resulting rate of change in the total Eulerian-frame particle number is only due to the flow of particles though the boundary of the domain D. That is, the DG scheme for the Eulerian-frame particle number is conservative by construction.

As for Eulerian-frame energy conservation, similar to Eq. (7) in Proposition 1, the element-integrated Eulerian-frame energy equation can be derived by adding the Eulerian-frame number equation in Eq. (126), with $\varphi_h = \varepsilon$, and the sum of the three number flux equations in Eq. (36) with test functions $\varphi_h = \varepsilon v^j_h$. To accommodate this choice of test functions, the approximation space $\mathbb{V}^k_h$ must include the piecewise linear function in the energy dimension, i.e., $k \geq 1$. Let $\mathcal{E}_h := \varepsilon\left(\mathcal{N}_h + v^j_h \mathcal{G}_{h,j}\right)$ denote the discretized Eulerian-frame energy density. Then, the resulting equation for the element-integrated Eulerian-frame energy takes the form

$$\begin{aligned}&\left(\partial_t \mathcal{E}_h\right)_{\boldsymbol{K}} := \left(\partial_t \mathcal{N}_h, \varepsilon\right)_{\boldsymbol{K}} + \left(\partial_t \mathcal{G}_{j,h}, \varepsilon v^j_h\right)_{\boldsymbol{K}} \\ &= -\sum_{i=1}^{3} \int_{\tilde{\boldsymbol{K}}^i} \left[\widehat{\mathcal{F}^i_{\mathcal{E}}}\left(\boldsymbol{\mathcal{U}}_h, \boldsymbol{v}_h\right)|_{x^i_{\mathrm{H}}} - \widehat{\mathcal{F}^i_{\mathcal{E}}}\left(\boldsymbol{\mathcal{U}}_h, \boldsymbol{v}_h\right)|_{x^i_{\mathrm{L}}}\right] \tau \, d\tilde{\boldsymbol{z}}^i - \int_{\boldsymbol{K}_{\boldsymbol{x}}} \left[\varepsilon^3 \widehat{\mathcal{F}^{\varepsilon}_{\mathcal{E}}}\left(\boldsymbol{\mathcal{U}}_h, \boldsymbol{v}_h\right)|_{\varepsilon_{\mathrm{H}}} - \varepsilon^3 \widehat{\mathcal{F}^{\varepsilon}_{\mathcal{E}}}\left(\boldsymbol{\mathcal{U}}_h, \boldsymbol{v}_h\right)|_{\varepsilon_{\mathrm{L}}}\right] d\boldsymbol{x} \\ &\quad + \left(\varepsilon \mathcal{F}^{\varepsilon}_{\mathcal{N}}(\boldsymbol{\mathcal{U}}_h, \boldsymbol{v}_h)\right)_{\boldsymbol{K}} + \sum_{i=1}^{3} \left(\mathcal{F}^i_{\mathcal{G}_j}(\boldsymbol{\mathcal{U}}_h, \boldsymbol{v}_h), \varepsilon \partial_i v^j_h\right)_{\boldsymbol{K}} + O(\upsilon^2_h),\end{aligned} \tag{127}$$

where we have defined the position space numerical fluxes,

$$\widehat{\mathcal{F}^i_{\mathcal{E}}}\left(\boldsymbol{\mathcal{U}}_h, \boldsymbol{v}_h\right)|_{x^i_{\mathrm{H/L}}} = \varepsilon \left[\widehat{\mathcal{F}^i_{\mathcal{N}}}\left(\boldsymbol{\mathcal{U}}_h, \boldsymbol{v}_h\right) + v^j_h \widehat{\mathcal{F}^i_{\mathcal{G}_j}}\left(\boldsymbol{\mathcal{U}}_h, \boldsymbol{v}_h\right)\right]|_{x^i_{\mathrm{H/L}}}, \tag{128}$$

the energy space numerical fluxes,

$$\widehat{\mathcal{F}^{\varepsilon}_{\mathcal{E}}}\left(\boldsymbol{\mathcal{U}}_h, \boldsymbol{v}_h\right)|_{\varepsilon_{\mathrm{H/L}}} = \varepsilon \widehat{\mathcal{F}^{\varepsilon}_{\mathcal{N}}}\left(\boldsymbol{\mathcal{U}}_h, \boldsymbol{v}_h\right)|_{\varepsilon_{\mathrm{H/L}}}, \tag{129}$$

and $\upsilon^2_h := |\boldsymbol{v}_h|^2$. Here, $\mathcal{F}^i_{\mathcal{G}_j}$ and $\widehat{\mathcal{F}^i_{\mathcal{G}_j}}$ represent the fluxes and the corresponding numerical fluxes for the number flux equation, defined in Eqs. (16) and (41), respectively. The third and fourth term on the right-hand side of Eq. (127), which emanate from the energy derivative term of the number equation and the spatial derivative of the number flux equations, respectively, can be written as

$$\left(\varepsilon\mathcal{F}^{\varepsilon}_{\mathcal{N}}(\boldsymbol{\mathcal{U}}_h,\boldsymbol{v}_h)\right)_{\boldsymbol{K}}+\sum_{i=1}^{3}\left(\mathcal{F}^{i}_{\mathcal{G}_j}(\boldsymbol{\mathcal{U}}_h,\boldsymbol{v}_h),\varepsilon\partial_i v^j_h\right)_{\boldsymbol{K}}=\int_{\boldsymbol{K}}\varepsilon\mathcal{K}^i{}_j\left[\partial_i v^j_h-(\partial_i v^j)_h\right]\tau d\boldsymbol{z}+O(v_h^2), \tag{130}$$

where $(\partial_i v^j)_h$ is the discretized velocity derivative that satisfies Eq. (44).

Provided (i) the numerical flux in Eq. (128) is uniquely defined on element interfaces and (ii) the first term on the right-hand side of Eq. (130) vanishes, Eq. (127) is, to $O(v_h^2)$, a phase-space conservation law for the element-integrated Eulerian-frame energy, in accordance with Proposition 1. These requirements — which generally require the discrete velocity $\boldsymbol{v}_h$ to be continuous across the elements — are not satisfied exactly by the DG scheme proposed here. Since the components of $\boldsymbol{v}_h$ are represented by piecewise polynomials, which are discontinuous on element boundaries, the violation in Eulerian-frame energy conservation may be larger than what would be expected from $O(v_h^2)$ contributions alone. We will investigate the simultaneous conservation of Eulerian-frame number and energy numerically in Section 8.

6.2. Spectral redistribution

In addition to the potential violations of Eulerian-frame energy conservation, beyond the $O(v^2)$ violations inherent to the model, from discontinuous $\boldsymbol{v}_h$ as discussed in Section 6.1, the realizability-enforcing limiter introduced in Section 5.2 can also affect the conservation of energy. In fact, for small velocities (including $v=0$, when the total energy should be preserved to machine precision) the realizability-enforcing limiter is the dominant source of non-conservation of the Eulerian-frame energy. To improve Eulerian-frame energy conservation, we propose a "spectral redistribution" scheme that corrects the change of energy induced by the realizability-enforcing limiter via a redistribution of particles between energy elements through a sweeping procedure. This approach maintains Eulerian-frame number and energy conservation across all energy elements *for a given spatial element*, at the expense of local number conservation in each energy element. The spectral redistribution does not correct for Eulerian-frame energy conservation violations inherent to the $O(v)$ two-moment model or due to discontinuous $\boldsymbol{v}_h$ (see Section 6.1).

To facilitate the discussion, we denote the element integrated Eulerian-frame number and energy by N and E, respectively. Given evolved moments $\boldsymbol{\mathcal{U}}_h=(\mathcal{N}_h,\boldsymbol{\mathcal{G}}_h)$ on element $\boldsymbol{K}$, the element-integrated number and energy can be computed by

$$\mathsf{N}_{\boldsymbol{K}}=\int_{\boldsymbol{K}}\mathcal{N}_h\varepsilon^2 d\boldsymbol{z}:=|\boldsymbol{K}|\sum_{\boldsymbol{k}=1}^{|S^{\boldsymbol{K}}_{\otimes}|}w^{(2)}_{\boldsymbol{k}}\mathcal{N}_{\boldsymbol{k}},\quad\text{and} \tag{131}$$

$$\mathsf{E}_{\boldsymbol{K}}=\int_{\boldsymbol{K}}(\mathcal{N}_h+v^j\mathcal{G}_{h,j})\varepsilon^3 d\boldsymbol{z}:=|\boldsymbol{K}|\sum_{\boldsymbol{k}=1}^{|S^{\boldsymbol{K}}_{\otimes}|}w^{(3)}_{\boldsymbol{k}}(\mathcal{N}_{\boldsymbol{k}}+v^j\mathcal{G}_{\boldsymbol{k},j}), \tag{132}$$

where $S^{\boldsymbol{K}}_{\otimes}$ denote the set of local DG nodes as defined in Eq. (48), $\mathcal{N}_{\boldsymbol{k}}$ and $\mathcal{G}_{\boldsymbol{k}}$ denotes the nodal values at points in $S^{\boldsymbol{K}}_{\otimes}$, and the weights $w^{(2)}_{\boldsymbol{k}}$ and $w^{(3)}_{\boldsymbol{k}}$ are given by the tensor product of the $(k+1)$-point one-dimensional LG quadrature rules introduced in Section 4.1, weighted by ε^2 and ε^3, respectively. Let $\widetilde{\boldsymbol{\mathcal{U}}}_h:=\texttt{RealizabilityLimiter}(\widehat{\boldsymbol{\mathcal{U}}}_h)$ be the output of the realizability-enforcing limiter given a potentially non-realizable solution $\widehat{\boldsymbol{\mathcal{U}}}_h$, and let $(\widetilde{\mathsf{N}}_{\boldsymbol{K}},\widetilde{\mathsf{E}}_{\boldsymbol{K}})$ and $(\widehat{\mathsf{N}}_{\boldsymbol{K}},\widehat{\mathsf{E}}_{\boldsymbol{K}})$ denote the element-integrated number and energy, defined in Eqs. (131) and (132), for $\widetilde{\boldsymbol{\mathcal{U}}}_h$ and $\widehat{\boldsymbol{\mathcal{U}}}_h$, respectively. As discussed in Section 5.2, the realizability-enforcing limiter gives a solution $\widetilde{\boldsymbol{\mathcal{U}}}_h$ that is realizable on $\widetilde{S}^{\boldsymbol{K}}_{\otimes}$ while maintaining number conservation in each element; i.e., $\widetilde{\mathsf{N}}_{\boldsymbol{K}}=\widehat{\mathsf{N}}_{\boldsymbol{K}}$. However, in part because of the additional factor of ε in the definition of the element-integrated energy in Eq. (132), the limiter results in energy changes (i.e., $\widetilde{\mathsf{E}}_{\boldsymbol{K}}\neq\widehat{\mathsf{E}}_{\boldsymbol{K}}$), which can lead to $\sum_{\boldsymbol{K}\in\mathcal{T}}\widetilde{\mathsf{E}}_{\boldsymbol{K}}\neq\sum_{\boldsymbol{K}\in\mathcal{T}}\widehat{\mathsf{E}}_{\boldsymbol{K}}$; i.e., a change in the global Eulerian-frame energy.

The proposed spectral redistribution corrects Eulerian-frame energy conservation violations by redistributing particles via a sweeping procedure in the energy dimension to produce $\boldsymbol{\mathcal{U}}_h:=\texttt{SpectralRedistribution}(\widetilde{\boldsymbol{\mathcal{U}}}_h)$, as detailed in Algorithm 2. Here we let $\mathcal{T}_{\boldsymbol{x}}$ denote the collection of all spatial elements $\boldsymbol{K}_{\boldsymbol{x}}$ and let $\mathcal{T}_{\varepsilon}=\{K_{\varepsilon,n}\}_{n=1}^{N_{\varepsilon}}$ denote the collection of all energy elements K_{ε} that cover the energy domain D_{ε}. For a given spatial element $\boldsymbol{K}_{\boldsymbol{x}}\in\mathcal{T}_{\boldsymbol{x}}$, the proposed spectral redistribution scheme sweeps through elements $\boldsymbol{K}=K_{\varepsilon}\times\boldsymbol{K}_{\boldsymbol{x}}$ for all $K_{\varepsilon}\in\mathcal{T}_{\varepsilon}$ in a user-prescribed order to redistribute particles in a way that the number and energy are both conserved for the given spatial element $\boldsymbol{K}_{\boldsymbol{x}}\in\mathcal{T}_{\boldsymbol{x}}$, i.e.,

$$\sum_{K_{\varepsilon}\in\mathcal{T}_{\varepsilon}}\mathsf{N}_{K_{\varepsilon}\times\boldsymbol{K}_{\boldsymbol{x}}}=\sum_{K_{\varepsilon}\in\mathcal{T}_{\varepsilon}}\widehat{\mathsf{N}}_{K_{\varepsilon}\times\boldsymbol{K}_{\boldsymbol{x}}}\quad\text{and}\quad\sum_{K_{\varepsilon}\in\mathcal{T}_{\varepsilon}}\widetilde{\mathsf{E}}_{K_{\varepsilon}\times\boldsymbol{K}_{\boldsymbol{x}}}=\sum_{K_{\varepsilon}\in\mathcal{T}_{\varepsilon}}\widehat{\mathsf{E}}_{K_{\varepsilon}\times\boldsymbol{K}_{\boldsymbol{x}}}, \tag{133}$$

which then leads to global number and energy conservation, $\sum_{\boldsymbol{K}\in\mathcal{T}}\mathsf{N}_{\boldsymbol{K}}=\sum_{\boldsymbol{K}\in\mathcal{T}}\widehat{\mathsf{N}}_{\boldsymbol{K}}$ and $\sum_{\boldsymbol{K}\in\mathcal{T}}\mathsf{E}_{\boldsymbol{K}}=\sum_{\boldsymbol{K}\in\mathcal{T}}\widehat{\mathsf{E}}_{\boldsymbol{K}}$, by summing over all spatial elements.

The details of the sweeping procedure and the spectral redistribution strategy are given in Algorithm 2. Specifically, when the Eulerian-frame energy conservation violation $\delta\mathsf{E}$ is nonzero, the spectral redistribution scheme redistributes particles between elements in a pairwise manner to correct $\delta\mathsf{E}$. The pairwise redistribution strategy is detailed in Algorithm 3, where a pair of scaling coefficients (θ_1,θ_2) is computed by solving a linear system that requires the sum of scaled energies to correct $\delta\mathsf{E}$ while preserving the sum of particles (see Line 3). When at least one of the coefficients (θ_1 and θ_2) is less than a prescribed threshold $\theta_{\min}>-1$, a damping factor γ is applied so that $\theta_1>\theta_{\min}$ and $\theta_2>\theta_{\min}$, which preserves moment realizability. (The moment realizability property is invariant to scaling by a positive scalar.) When the linear system does not have a solution, or when $\min(\theta_1,\theta_2)<\theta_{\min}$, the output

of Algorithm 3 does not fully correct $\delta\mathsf{E}$, and the remainder is propagated to the next pair of elements in the sweeping procedure. As shown in Algorithm 2, beginning on Line 18, a backward sweep will be launched after the forward sweep when $|\delta\mathsf{E}| > \delta$; i.e., when the energy conservation violation is not fully corrected. Here $\delta > 0$ is a user-specified tolerance on the energy conservation violation. In the implementation, we choose to omit the condition in Line 19 and perform the full backward sweep in order to improve the computational efficiency on GPUs. Moreover, to avoid numerical issues, we restrict the damping factor γ in Algorithm 3 such that the resulting redistributed moment $\mathcal{U}_h$ remains a strictly positive number density; i.e., $\mathcal{N}_h > 0$. In the numerical results reported in Section 8, we choose $\theta_{\min} = -0.5$ and permute the energy elements in an ascending order based on the associated energy values. We observe that the forward and backward sweeping procedure is sufficient for correcting the energy conservation violations introduced by the realizability-enforcing limiter — i.e., $\delta\mathsf{E} \to 0$ during the sweeping procedure — and that the additional scaling introduced in this energy correction process has no noticeable adverse impact on the solution to the two-moment system.

Algorithm 2: $\mathcal{U}_h = \texttt{SpectralRedistribution}(\widetilde{\mathcal{U}}_h)$.

1 **Inputs:** Discretized moments before and after the realizability-enforcing limiter, i.e., $\widehat{\mathcal{U}}_h$ and $\widetilde{\mathcal{U}}_h$; a permutation of the energy elements, denoted as $K_{\varepsilon,n}$, $n = 1, \dots, N_\varepsilon$.
2 **Parameter:** δ (Energy conservation violation tolerance)
3 $\mathcal{U}_h \leftarrow \widetilde{\mathcal{U}}_h$; // Initialize the output moment
4 **for** *each spatial element* $\boldsymbol{K}_x \in \mathcal{T}_x$ **do**
5 **for** $n = 1, \dots, N_\varepsilon$ **do**
6 Compute $(\mathsf{N}_{K_{\varepsilon,n}\times\boldsymbol{K}_x}, \mathsf{E}_{K_{\varepsilon,n}\times\boldsymbol{K}_x})$ from $\mathcal{U}_h$ using Eqs. (131) and (132);
7 Compute $\widehat{\mathsf{E}}_{K_{\varepsilon,n}\times\boldsymbol{K}_x}$ from $\widehat{\mathcal{U}}_h$ using Eq. (132);
8 $\mathsf{N}_n \leftarrow \mathsf{N}_{K_{\varepsilon,n}\times\boldsymbol{K}_x}$, $\mathsf{E}_n \leftarrow \mathsf{E}_{K_{\varepsilon,n}\times\boldsymbol{K}_x}$, $\widehat{\mathsf{E}}_n \leftarrow \widehat{\mathsf{E}}_{K_{\varepsilon,n}\times\boldsymbol{K}_x}$;
9 **if** $\sum_{n=1}^{N_\varepsilon} E_n \neq \sum_{n=1}^{N_\varepsilon} \widehat{E}_n$ **then**
10 $\delta\mathsf{E} \leftarrow \mathsf{E}_1 - \widehat{\mathsf{E}}_1$;
  /* Forward sweep */
11 **for** $n = 1, \dots, N_\varepsilon - 1$ **do**
12 $\delta\mathsf{E} \leftarrow \delta\mathsf{E} + \mathsf{E}_{n+1} - \widehat{\mathsf{E}}_{n+1}$;
13 **if** $|\delta E| > \delta$ **then**
14 $(\theta_n, \theta_{n+1}) = \texttt{ComputeCorrection}(\mathsf{N}_n, \mathsf{E}_n, \mathsf{N}_{n+1}, \mathsf{E}_{n+1}, \delta\mathsf{E})$;
  /* Update corrected moments, numbers, and energies */
15 $\mathcal{U}_h \leftarrow (1+\theta_n)\mathcal{U}_h$ on $K_{\varepsilon,n} \times \boldsymbol{K}_x$, and $\mathcal{U}_h \leftarrow (1+\theta_{n+1})\mathcal{U}_h$ on $K_{\varepsilon,n+1} \times \boldsymbol{K}_x$;
16 $(\mathsf{N}_n, \mathsf{E}_n) \leftarrow (1+\theta_n)(\mathsf{N}_n, \mathsf{E}_n)$, and $(\mathsf{N}_{n+1}, \mathsf{E}_{n+1}) \leftarrow (1+\theta_{n+1})(\mathsf{N}_{n+1}, \mathsf{E}_{n+1})$;
17 $\delta\mathsf{E} \leftarrow \mathsf{E}_n + \mathsf{E}_{n+1} + \delta\mathsf{E}$;
  /* Backward sweep */
18 **for** $n = N_\varepsilon - 1, \dots, 2$ **do**
19 **if** $|\delta E| > \delta$ **then**
20 $(\theta_n, \theta_{n-1}) = \texttt{ComputeCorrection}(\mathsf{N}_n, \mathsf{E}_n, \mathsf{N}_{n-1}, \mathsf{E}_{n-1}, \delta\mathsf{E})$;
  /* Update corrected moments, numbers, and energies */
21 $\mathcal{U}_h \leftarrow (1+\theta_n)\mathcal{U}_h$ on $K_{\varepsilon,n} \times \boldsymbol{K}_x$, and $\mathcal{U}_h \leftarrow (1+\theta_{n-1})\mathcal{U}_h$ on $K_{\varepsilon,n-1} \times \boldsymbol{K}_x$;
22 $(\mathsf{N}_n, \mathsf{E}_n) \leftarrow (1+\theta_n)(\mathsf{N}_n, \mathsf{E}_n)$, and $(\mathsf{N}_{n-1}, \mathsf{E}_{n-1}) \leftarrow (1+\theta_{n-1})(\mathsf{N}_{n-1}, \mathsf{E}_{n-1})$;
23 $\delta\mathsf{E} \leftarrow \mathsf{E}_n + \mathsf{E}_{n-1} + \delta\mathsf{E}$;
24 **else**
25 break;

Algorithm 3: $(\theta_1, \theta_2) = \texttt{ComputeCorrection}(\mathsf{N}_1, \mathsf{E}_1, \mathsf{N}_2, \mathsf{E}_2, \delta\mathsf{E})$.

1 **Inputs:** $\mathsf{N}_1, \mathsf{E}_1, \mathsf{N}_2, \mathsf{E}_2, \delta\mathsf{E}$
2 **Parameter:** $\theta_{\min} > -1$
3 Compute (θ_1, θ_2) by solving $\begin{cases} \theta_1\mathsf{N}_1 + \theta_2\mathsf{N}_2 = 0 \\ \theta_1\mathsf{E}_1 + \theta_2\mathsf{E}_2 = -\delta\mathsf{E} \end{cases}$;
4 **if** *no solution* **then**
5 $(\theta_1, \theta_2) \leftarrow (0, 0)$;
6 **if** $\min(\theta_1, \theta_2) < \theta_{\min}$ **then**
7 $\gamma \leftarrow \frac{\theta_{\min}}{\min(\theta_1, \theta_2)}$;
8 $(\theta_1, \theta_2) \leftarrow (\gamma\theta_1, \gamma\theta_2)$; // Limit for realizability

7. Implementation, programming models, and portability

The DG-IMEX method proposed here has been implemented in the toolkit for high-order neutrino radiation hydrodynamics (THORNADO). Here we briefly discuss some considerations in this process.

Neutrino transport is only one component (along with, e.g., hydrodynamics, nuclear reaction kinetics, and gravity) of a broader, multiphysics simulation framework needed to model multiscale astrophysical systems, e.g., core-collapse supernova explosions. However, the number of evolved degrees of freedom is relatively high compared to other components. For example, simulations incorporating a two-moment model (four moments), evolving three independent neutrino flavors (six species), with 16 linear elements (k = 1) to discretize the energy dimension evolve $4 \times 6 \times 16 \times (k+1) = 768$ degrees of freedom per spatial point. As such, spectral neutrino radiation transport represents the bulk of the computational load in such scientific applications. With this in mind, node-level performance and portability for heterogeneous computing systems are prioritized in THORNADO development as a collection of modular physics components that can be incorporated into distributed multiphysics simulation codes (e.g., FLASH-X [57]), which are equipped with native infrastructure for distributed parallelism. In particular, we target frameworks that utilize adaptive mesh refinement, where simulation data is mapped to smaller grid blocks of relatively even size.

THORNADO uses a combination of compiler directives and optimized linear algebra libraries to accelerate all components of the DG-IMEX method. All of the solver components — e.g., the computation of numerical fluxes, evaluation of phase-space divergences, and limiters — are reduced to small, discrete kernels that can be executed either as collapsible, tightly-nested loops over phase-space dimensions or basic linear algebra operations. In addition to optimizing many key metrics for GPU performance (e.g., occupancy, register pressure, and memory coalescence), this strategy naturally exposes vector-level parallelism which also benefits performance on modern, multicore CPUs. This is especially important when invoking iterative solvers, such as those described in Sections 4.3.1 and 4.3.2, across many independent phase-space points. Since iteration counts can vary, assigning an even number of phase-space points to each thread can lead to severe load imbalance among GPU threads. We address this problem by tracking the convergence of each point independently, removing them from calculations in each kernel until all points have converged.

Our portability strategy focuses on maintaining a single code-base that can efficiently execute on different hardware architectures and software environments. THORNADO contains three distinct implementations of compiler directives that are managed with C preprocessor macros: traditional OpenMP (CPU multi-core), OpenMP offload (GPU), and OpenACC (GPU). We refer to code listings in [58] for specific examples.

Interfaces to optimized linear algebra routines are also written in a generic way for portability across different libraries. Currently, THORNADO has linear algebra interfaces supporting several LAPACK and BLAS [59] routines with GPU implementations from NVIDIA, AMD, Intel, and MAGMA [60]. This approach hides the complexities of managing different interfaces in a single THORNADO module that can be easily used throughout the code. In addition to the individual routine interfaces, each linear algebra package requires specific attention to interoperability with the compiler directives to ensure correct synchronization when using multiple execution streams per device. This is managed during initialization with compiler directives and C preprocessor macros.

We provide timing results and a breakdown of the computational cost associated with key solver components for one of the numerical examples in Section 8.

8. Numerical tests

In this section, we demonstrate the performance of our implementation of the DG-IMEX method to solve the $O(v)$ two-moment model. We consider problems with and without collisions. For problems with collisions, we use the IMEX scheme proposed in [35] (see also [36] for details). For problems without collisions, we use the optimal second- and third-order accurate strong stability-preserving Runge–Kutta methods from [61], referred to as SSPRK2 and SSPRK3, respectively. For the tests in Sections 8.2 and 8.3, unless stated otherwise, we set the time step to $\Delta t = 0.3 \times |K_{\boldsymbol{x}}^1|/(k+1)$, where k is the polynomial degree. For the tests in Sections 8.4-8.6, we enforce the time step restriction given in Theorem 1.

Collisions tend to drive the distribution towards isotropy in the angular dimensions of momentum space (i.e., $|\mathcal{I}| \to 0$), which places the comoving-frame moments $\mathcal{M}$ safely inside the realizable domain. Therefore, to emphasize the improved robustness resulting from our analysis, our main focus is on phase-space advection problems without collisions, where the moments evolve close to the boundary of the realizable domain.

8.1. Moment conversion solver

The solution of the conserved-to-primitive moment conversion problem in Eq. (19) and the implicit system in Eq. (58) contribute the majority of the computational cost of the realizability-preserving scheme. In this section, we test the iterative solver for solving the moment conversion problem Eq. (19) with various solver configurations, and the results reported provide guidance for selecting iterative solver configurations for this critical part of the algorithm.

As discussed in Section 4.3.1, we formulate the moment conversion problem in Eq. (19) as a fixed-point problem on the primitive moments $\mathcal{M} = (\mathcal{D}, \mathcal{I})^\intercal$ of the form stated in Eq. (56). In Lemma 7, we have shown that the moment realizability is preserved in the iterative procedure when Eq. (56) is solved with the Picard iteration method in Eq. (57) and $\lambda \leq (1+v)^{-1}$ in Eq. (56). The convergence of Picard iteration is guaranteed in Theorem 2 with the additional assumption that $v < \sqrt{2} - 1$.

We first compare the iteration counts required for convergence of the Picard iteration solver and an Anderson acceleration (AA) solver, using two different choices for λ. The AA technique was first proposed in [62] to accelerate the convergence of fixed-point iterations by accounting for the past iteration history to compute new iterates. Here we follow the formulation and implementation in [63,36] and apply the AA solver to the moment conversion problem Eq. (56). In Fig. 5, the iteration counts are reported for the two iterative solvers applied to solve Eq. (56) at varying fluid speed $v := |\boldsymbol{v}|$ and flux factor $h = |\mathcal{I}|/\mathcal{D}$, with λ chosen to be the largest

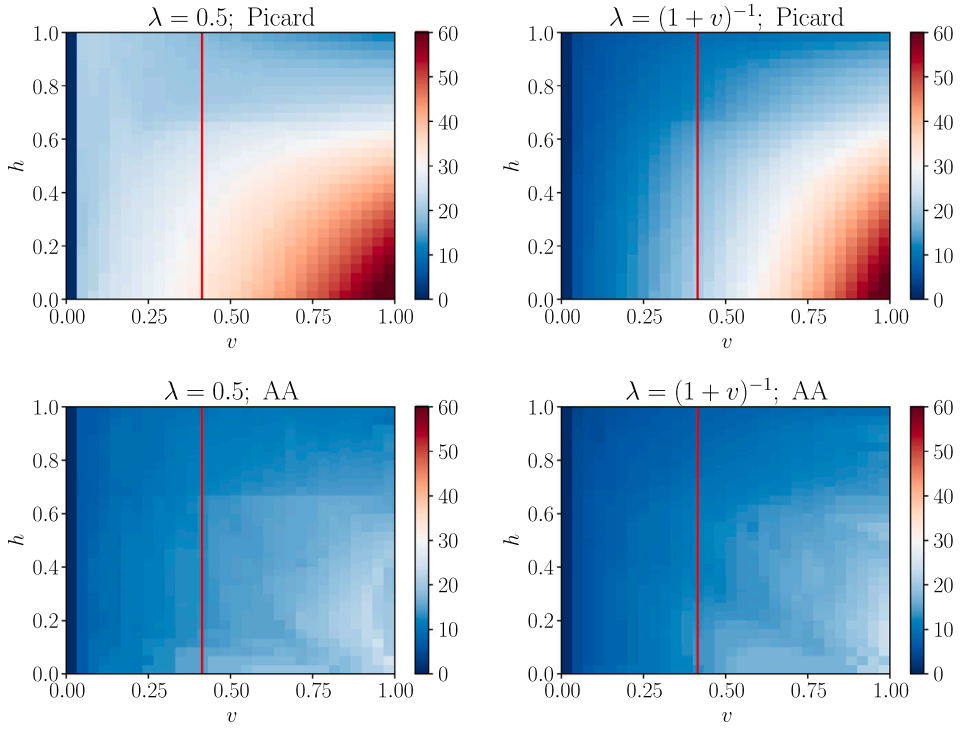


Fig. 5. Iteration counts for the Picard iteration (top panels) and AA (bottom panels) solvers with modified Richardson iteration parameter $\lambda = (1+v)^{-1}$ and $\lambda = 0.5$ (right and left columns, respectively), applied to the moment conversion problems with various velocity v and flux factor h. The reported iteration counts are the average over 100 randomly generated moment conversion problems at each (v, h), where the randomness is applied to the directions of $\boldsymbol{v}$ and $\mathcal{I}/\mathcal{D}$. In each panel, the red vertical line indicates the upper velocity bound for guaranteed convergence, $v = \sqrt{2} - 1$.

allowable value, i.e., $\lambda = (1+v)^{-1}$ and a more conservative value $\lambda = 0.5$. The AA solver uses the memory parameter $m = 1$ (defined in [36]), so that only information from the previous and current iterate is used. The stopping criteria for both solvers are given as

$$\|\boldsymbol{\mathcal{M}}^{[k]} - \boldsymbol{\mathcal{M}}^{[k-1]}\| \leq \texttt{tol}\, \|\boldsymbol{\mathcal{U}}\|, \tag{134}$$

where we consider the norms in the L^2 sense and the tolerance $\texttt{tol} = 10^{-8}$. For each choice of (v, h) in Fig. 5, the fixed-point problem is solved for 100 randomly generated $\boldsymbol{\mathcal{U}} \in \mathcal{R}$ (varying the direction of $\boldsymbol{v}$ and $\mathcal{I}/\mathcal{D}$ randomly), and the averaged iteration counts over these 100 problems are recorded. In each test, the initial guess takes the form $\boldsymbol{\mathcal{M}}^{[0]} = \boldsymbol{\mathcal{U}}$. The results in Fig. 5 illustrate that, for both the Picard iteration and the AA solvers, choosing the parameter λ to be the largest allowable value $(1+v)^{-1}$ indeed reduces the iteration counts from the more conservative choice $\lambda = 0.5$, particularly in the low velocity regime. In addition, it can be found from Fig. 5 that AA solver consistently outperforms the Picard iteration method, and the advantage of using AA grows as the velocity increases. We note that the realizability-preserving and convergence properties analyzed in Section 5.3 are only applicable to the Picard iteration solver, and not to the AA solver.[4] However, in the numerical results reported in Fig. 5, we have not observed convergence failure by any of the solvers, even when the velocity is larger than the upper bound ($v = \sqrt{2} - 1$; plotted as a red vertical line in each panel in Fig. 5) required in the convergence analysis in Theorem 2.

In Fig. 6, we show results from experimenting with two choices for the initial guess, $\boldsymbol{\mathcal{M}}^{[0]} = (\mathcal{N}, \mathbf{0})^\intercal$ and $\boldsymbol{\mathcal{M}}^{[0]} = \boldsymbol{\mathcal{U}} = (\mathcal{N}, \mathcal{G})^\intercal$, for the AA solver with $\lambda = (1+v)^{-1}$, which is the best performing configuration shown in Fig. 5. As shown in Fig. 6, initializing with the conserved moment $\boldsymbol{\mathcal{U}}$ generally outperforms the isotropic initial condition $(\mathcal{N}, \mathbf{0})$, except for the case when the flux factor $h = 0$, for which the isotropic initial condition is exactly the primitive moment. Since we expect moments with $h = 0$ to be rarely encountered in numerical simulations, adopting the AA solver with $\lambda = (1+v)^{-1}$ and initial guess $\boldsymbol{\mathcal{M}}^{[0]} = \boldsymbol{\mathcal{U}}$ appears to be the best choice. This conjecture is confirmed in the performance comparison reported in Section 8.7, where we observe a considerable improvement in terms of computational time by using the initial guess $\boldsymbol{\mathcal{M}}^{[0]} = \boldsymbol{\mathcal{U}}$.

[4] The realizability-preserving and convergence properties of the AA solver require additional conditions such as boundedness of extrapolation coefficients, which we do not enforce in the implementation.

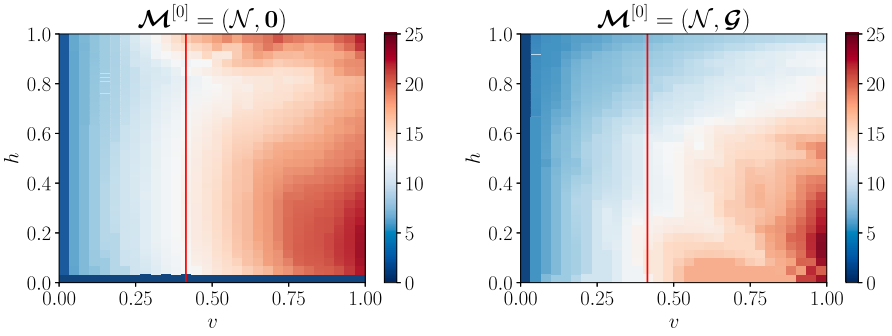


Fig. 6. Iteration counts for the AA solver with modified Richardson iteration parameters $\lambda = (1+v)^{-1}$ applied to the moment conversion problems with various velocity v and flux factor h for two different initial guesses. The reported iteration counts are the average over 100 randomly generated moment conversion problems at each (v, h).

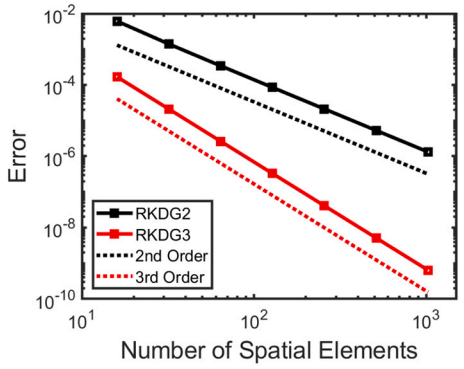


Fig. 7. Error in the L^2 norm versus number of spatial elements N for the Sine Wave Streaming test. Results obtained with second- and third-order schemes (using $k = 1$ polynomials and SSPRK2 time stepping and $k = 2$ polynomials and SSPRK3 time stepping, respectively), along with dotted reference lines proportional to $1/N^{k+1}$, are plotted using black and red lines, respectively.

8.2. Sine wave streaming

The first test we consider that evolves the two-moment system models free-streaming radiation through a background with a spatially (and temporally) constant velocity field in the x^1-direction. That is, we set $\chi = \sigma = 0$, while $\boldsymbol{v} = [v, 0, 0]^\intercal$, with $v = 0.1$. The purpose of this test is to verify (i) the correct radiation propagation speed in this idealized setting, and (ii) the expected order of accuracy of the implemented method. We consider a periodic one-dimensional unit spatial domain $D_{x^1} = [0, 1]$. The initial number density and flux are set to $\mathcal{D}(x^1, 0) = \mathcal{D}_0(x^1) = 0.5 + 0.49 \times \sin(2\pi x^1)$ and $\mathcal{I}^1(x^1, 0) = \mathcal{D}_0(x^1)$, respectively. Then, the flux factor is $h = 1$, and the analytic solution is given by $\mathcal{D}(x^1, t) = \mathcal{I}^1(x^1, t) = \mathcal{D}_0(x^1 - t)$; i.e., the initial profile propagates with unit speed, *independent* of v. (As noted by [1], dropping the velocity-dependent terms in the time derivatives of Eqs. (2) and (3), as is done in [4,5], the propagation speed becomes $1 + v$ for this test, which is unphysical.) Since the background velocity is constant, there is no coupling in the energy dimension. Therefore, this test is performed with a single energy. We run this test until $t = 1$, when the initial profile has crossed the grid once before returning to its initial position.

In Fig. 7, we plot the error in the L^2 norm versus the number of spatial elements for the second-order scheme, using linear polynomials ($k = 1$) and SSPRK2 time stepping, and the third-order scheme, using quadratic polynomials ($k = 2$) and SSPRK3 time stepping. The results confirm the expected convergence rate to the exact solution.

8.3. Gaussian diffusion

The next test we consider, adopted from [4], models the diffusion of particles through a medium moving with constant velocity in the x^1-direction. We consider a purely scattering medium and set $\sigma = 3.2 \times 10^3$ and $\chi = 0$, and let $\boldsymbol{v} = [v, 0, 0]^\intercal$, with $v = 0.1$. We let the spatial domain be periodic, $D_{x^1} = [0, 3]$, with initial conditions $\mathcal{D}_0(x^1) = \exp[-(x^1 - x_0^1)^2/(4t_0\kappa_\mathsf{D})]$ and $\mathcal{I}_0^1(x^1) = -\kappa_\mathsf{D}\partial_{x^1}\mathcal{D}_0$, where $\kappa_\mathsf{D} = (3\sigma)^{-1}$, and we set $x_0^1 = 1$ and $t_0 = 5$. Then, the evolution of the number density is approximately governed by the advection-diffusion equation

$$\partial_t \mathcal{D} + \partial_{x^1}\left(\mathcal{D}v - \kappa_\mathsf{D}\partial_{x^1}\mathcal{D}\right) = 0, \tag{135}$$

whose analytical solution is given by [4]

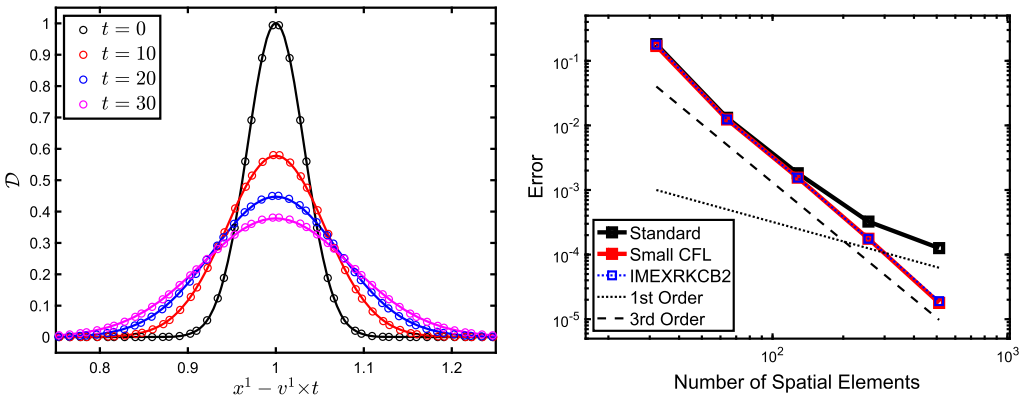


Fig. 8. Results for the Gaussian diffusion test. The left panel shows the numerical solution (open circles) versus the shifted coordinate $x^1 - vt$ for various times, compared with the analytic solution (solid lines) to the advection-diffusion equation in Eq. (135). The right panel shows the error in the L^2 norm versus the number of elements N. The error is computed with respect to a high-resolution reference run with $N = 8192$. Convergence results are shown for standard and reduced CFL number, using the first-order SSP-IMEX scheme from [35] (solid black and red, respectively; see text for details), and a second-order IMEX scheme from [64] with standard CFL number (dotted blue). The dotted and dashed black reference lines are proportional to N^{-1} and N^{-3}, respectively.

$$\mathcal{D}(x^1, t) = \sqrt{\frac{t_0}{t_0 + t}} \exp\left\{ -\frac{((x^1 - vt) - x_0^1)^2}{4(t_0 + t)\kappa_{\mathrm{D}}} \right\}. \tag{136}$$

Since there is no coupling in the energy dimension (v is constant), we perform this test with a single energy. We use quadratic elements ($k = 2$) and the IMEX time stepping scheme from [35], integrating the collision term implicitly. For this test, the time step is set to $\Delta t = C_{\mathrm{CFL}} \times |K_x^1|$, where C_{CFL} is specified below. The purpose of this test is to investigate the performance of the DG-IMEX scheme in a regime where both advection and diffusion contribute to the evolution of the number density. For $t > 0$, the Gaussian profile is advected with the flow, while the amplitude decreases and the width increases due to diffusion.

The left panel in Fig. 8 shows the number density versus $x^1 - vt$ for various times as the Gaussian profile propagates once across the periodic domain and returns to its initial position at $t = 30$. At this time the amplitude is reduced by a factor $\sqrt{5/35} \approx 0.378$. For this simulation, the spatial domain is discretized using 96 elements, so that the radio of the element width to the mean-free path is $|K_x^1|\sigma = 10^2$. The numerical solution (open circles) agrees well with the expression in Eq. (136) (solid lines).

The right panel in Fig. 8 shows the error in the L^2 norm at $t = 5$ versus the number of spatial elements N. Since the expression given by Eq. (136) is not an exact solution to the $O(v)$ two-moment model in the limit of high scattering opacity, we compare the numerical results to a high-resolution reference solution, computed with 8192 elements. (We have confirmed that for fixed N and t, and varying v, the difference between the numerical solution and the expression in Eq. (136) increases as $O(v^2)$.) The solid black curve with squares shows the error obtained with the standard CFL number $C_{\mathrm{CFL}} := C_{\mathrm{CFL}}^0 = 0.3/(k+1)$. For smaller N, the error falls off as N^{-3} (see dashed black reference line), consistent with a third-order convergence rate, while for larger N the convergence rate transitions to first-order (see dotted black reference line). The reason for this change in convergence rate is because the IMEX scheme, taken from [35], is formally only first-order accurate. Spatial discretization errors dominate for small N, but since these errors decrease with the third-order rate, temporal errors become dominant for large N. To verify this, we also plot convergence results obtained after reducing the time step by a factor of 25, $C_{\mathrm{CFL}} := C_{\mathrm{CFL}}^0/25$; see solid red curve with squares. For this case, the error decreases with the third-order rate for all N. We also show the error obtained with a second-order IMEX scheme (IMEXRKCB2 from [64]), using the standard CFL number. Due to better temporal accuracy, the error decreases with the third-order rate, but the scheme does not satisfy the convex-invariant conditions delineated in [35], and can therefore not be guaranteed to maintain moment realizability by our analysis.

8.4. Streaming Doppler shift

This test, adopted from [25] (see also [4,5]), models the propagation of free-streaming radiation along the x^1-direction through a background with a spatially varying velocity field. Because the two-moment model adopts momentum-space coordinates associated with a comoving observer, the radiation energy spectra will be Doppler shifted. We consider a one-dimensional spatial domain $D_{x^1} = [0, 10]$. Again, we set $\chi = \sigma = 0$, while the velocity field is set to $\boldsymbol{v} = (v, 0, 0)^{\mathsf{T}}$, where

$$v(x^1) = \begin{cases} 0, & x^1 \in [0, 2) \\ v_{\max} \times \sin^2[2\pi(x^1 - 2)/6], & x^1 \in [2, 3.5) \\ v_{\max}, & x^1 \in [3.5, 6.5) \\ v_{\max} \times \sin^2[2\pi(x^1 - 2)/6], & x^1 \in [6.5, 8) \\ 0, & x^1 \in [8, 10] \end{cases}, \tag{137}$$

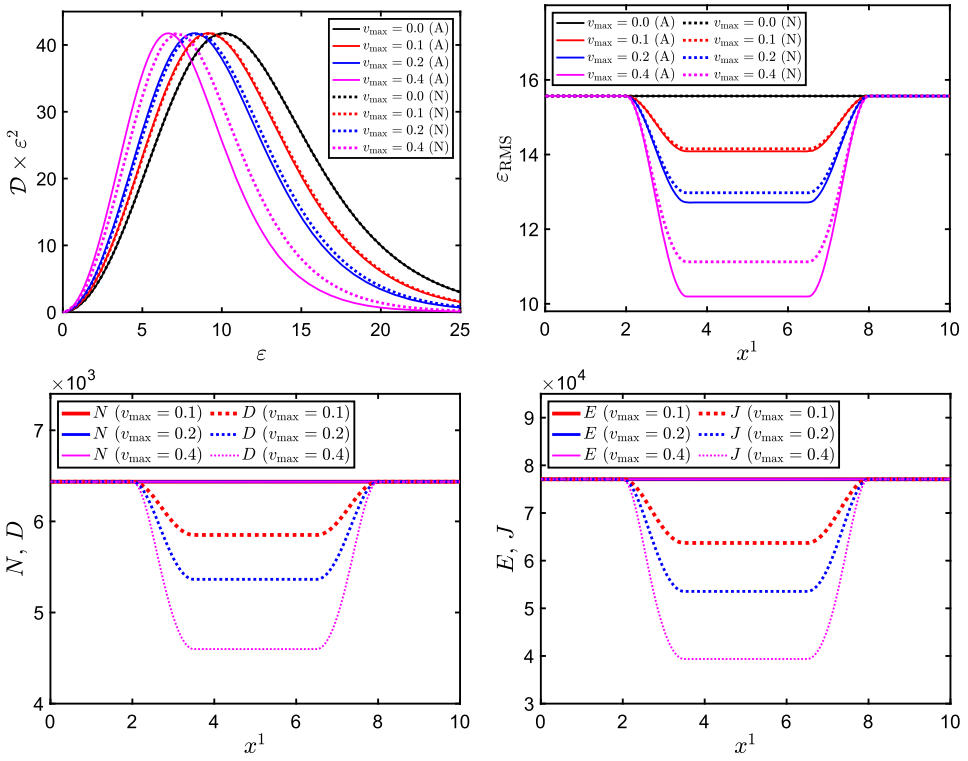


Fig. 9. Steady state solutions ($t = 20$) for the streaming Doppler shift problem for various $v_{\max} \in \{0.0, 0.1, 0.2, 0.4\}$. In the top panels, we plot spectra at $x^1 = 5$ ($\mathcal{D}\varepsilon^2$ versus ε; left) and the RMS energy, as defined in Eq. (139), versus position x^1 (right). In these panels, solid lines represent the analytic (A) solution from special relativistic considerations, given by Eq. (138), while dotted lines represent the numerical (N) results. In all panels, black, red, blue, and magenta curves represent runs with $v_{\max}$ set to 0.0, 0.1, 0.2, and 0.4, respectively. In the bottom left panel, the Eulerian-frame (solid lines) and comoving-frame (dotted lines) number densities are plotted versus position. In the bottom right panel, the Eulerian-frame (solid lines) and comoving-frame (dotted lines) energy densities are plotted versus position.

and where we will vary $v_{\max}$. We set the energy domain to $D_\varepsilon = [0, 50]$. In this test, we discretize the spatial and energy domains into 128 and 32 elements, respectively, and use quadratic elements ($k = 2$) and SSPRK3 time stepping. In the computational domain, the moments are initially set to $\mathcal{D} = 1 \times 10^{-40}$ and $\mathcal{I}^1 = 0$ for all $(x^1, \varepsilon) \in D_{x^1} \times D_\varepsilon$. At the inner spatial boundary, we impose an incoming, forward-peaked radiation field with a Fermi-Dirac spectrum; i.e., we set $\mathcal{D}(\varepsilon, x^1 = 0) = 1/[\exp(\varepsilon/3 - 3) + 1]$ and $\mathcal{I}^1(\varepsilon, x^1 = 0) = 0.999 \times \mathcal{D}(\varepsilon, x^1 = 0)$, so that the flux factor $h \approx 1$. (We impose outflow boundary conditions at $x^1 = 10$.) Then, for $t > 0$, a radiation front propagates through the computational domain, and a steady state is established for $t \gtrsim 10$, where the spectrum is Doppler-shifted according to the velocity field. From special relativistic considerations, similar to [4], the analytical spectral number density in the steady state can be written as

$$\mathcal{D}_{\mathrm{A}} = \frac{s^2}{\exp(s\varepsilon/3 - 3) + 1}, \tag{138}$$

where $s = \sqrt{(1+v)/(1-v)}$. The purpose of this test is to (i) compare steady state numerical solutions with the prediction from special relativity given by Eq. (138), and (ii) investigate the simultaneous Eulerian-frame number and energy conservation properties of the method as the initial conditions are evolved to steady state. To reach an approximate steady state, we run all models until $t = 20$.

8.4.1. *General solution characteristics*

Fig. 9 displays steady state solution characteristics for models where we have varied $v_{\max} \in \{0.0, 0.1, 0.2, 0.4\}$. From the top left panel, we see that the $O(v)$ spectra (dotted) — which for $v_{\max} > 0$ are red-shifted relative to the case with $v_{\max} = 0$ — agree well with the analytic, special relativistic results (solid) for the lower values of $v_{\max}$, while the difference between the $O(v)$ and the special relativistic results are larger when $v_{\max} = 0.4$, which is to be expected since $O(v^2)$ terms are no longer negligible. The top right panel shows the RMS energy, defined as

$$\varepsilon_{\mathrm{RMS}} = \sqrt{\int_{D_\varepsilon} \mathcal{D}\varepsilon^5 d\varepsilon \Big/ \int_{D_\varepsilon} \mathcal{D}\varepsilon^3 d\varepsilon}, \tag{139}$$

versus position, computed from the numerical, $O(v)$ solution and the analytic, special relativistic solution in Eq. (138). Indeed, at $x^1 = 5$, the relative difference in $\varepsilon_{\mathrm{RMS}}$ between the $O(v)$ and the special relativistic result for $v_{\max} = 0.1$ is 4.8×10^{-3}, while it is

2.0×10^{-2} and 9.1×10^{-2} for $v_{\max} = 0.2$ and $v_{\max} = 0.4$, respectively. That is, the relative error in the RMS energy increases roughly as $O(v^2)$.

The bottom panels of Fig. 9 show the Eulerian- and comoving-frame number densities (N and D, respectively; left), and the Eulerian- and comoving-frame energy densities (E and J, respectively; right) versus position. Here,

$$\{ N, D, E, J \} = 4\pi \int_{D_\varepsilon} \{ \mathcal{N}, \mathcal{D}, \mathcal{E}, \mathcal{D}\varepsilon \} \varepsilon^2 d\varepsilon, \tag{140}$$

where $\mathcal{N}$ and $\mathcal{E}$ are defined in Eqs. (4) and (5), respectively. Relative to where $v = 0$, both D and J are lower in the region where $v > 0$, which is expected from the redshifted spectra displayed in the top left panel in Fig. 9. The Eulerian-frame quantities, N and E, are practically unaffected by the velocity field, and remain relatively constant throughout the spatial domain.

8.4.2. Simultaneous number and energy conservation

Next, we investigate the simultaneous number and energy conservation properties of the scheme as a function of time for $t \in [0, 20]$. Here, a main challenge stems from the fact that, since the flux factor $h \approx 1$, the moments evolve close to the boundary of the realizable set $\mathcal{R}$. With high-order polynomials ($k \geq 1$), the solution can become non-realizable in one or more quadrature points in some elements, which then triggers the realizability-enforcing limiter discussed in Section 5.2. The realizability-enforcing limiter preserves the Eulerian-frame particle number, but not the Eulerian-frame energy, which is the reason for introducing the "spectral redistribution" scheme in Section 6.2. Here, we demonstrate the performance of the spectral redistribution, and its effect on the simultaneous number and energy conservation properties of the method. Recall from Proposition 1 that the $O(v)$ two-moment model is conservative for the Eulerian-frame energy only to $O(v^2)$.

In the context of the current test, the Eulerian-frame number density satisfies a conservation law of the form

$$\partial_t N + \partial_1 F_N^1 = 0. \tag{141}$$

Integration over space D_{x^1} and time $[t_0, t]$ gives

$$\underbrace{\int_{D_{x^1}} [\, N(x^1, t) - N(x^1, 0)]\, dx^1}_{\Delta N_{\text{int}}(t)} + \underbrace{\int_0^t [F_N^1|_{x^1=10} - F_N^1|_{x^1=0}]\, d\tau}_{\Delta N_{\text{ext}}(t)} = 0, \tag{142}$$

where $\Delta N_{\text{int}}(t)$ and $\Delta N_{\text{ext}}(t)$ represent the change in the total number of particles, from t_0 to t, *interior* and *exterior* to the domain D_{x^1}, respectively. Since there is no creation or destruction of particles, the sum vanishes. We can obtain a similar expression for the Eulerian-frame energy (with E replacing N in Eq. (142)), but for the $O(v)$ two-moment model considered here, by Proposition 1, one would in general expect

$$\Delta E_{\text{int}} + \Delta E_{\text{ext}} = O(v^2), \tag{143}$$

at the continuous level. At the discrete level, with consistent discretization of the left-hand side of the two-moment model, Eulerian-frame energy violations of $O(v^2)$ should be considered optimal. (Recall the discussion on this issue specific to the DG scheme from Section 6.1.) For this test, given our chosen spatial resolution and use of quadratic elements, velocity jumps across elements are small, and we expect to observe near optimal Eulerian-frame energy conservation properties. However, the acceptable level of Eulerian-frame energy nonconservation is application dependent, and should be considered on a case-by-case basis.

Fig. 10 plots the number and energy balance versus time for a run with $v_{\max} = 0.2$. Initially, there are essentially no particles in the computational domain D_{x^1}, and the flux at the outer boundary is zero. For $t > 0$, particles enter the domain through the inner boundary, and ΔN_{int} begins to increase linearly with time, while ΔN_{ext} decreases at the same rate, and the sum $\Delta N_{\text{int}} + \Delta N_{\text{ext}}$ remains zero to machine precision (see also Fig. 11). Around $t = 10$, the particles that entered the domain at $t = 0$ reach the outer boundary, establishing a balance between particles entering and leaving the domain, and the system reaches a steady state where both ΔN_{int} and ΔN_{ext} remain unchanged. The evolution observed for the Eulerian-frame energy quantities (ΔE_{ext} and ΔE_{ext}) is similar to that for the particle number, and, on the scale of the ordinate on the right panel in Fig. 10, the sum $\Delta E_{\text{ext}} + \Delta E_{\text{ext}}$ remains close to zero.

Fig. 11 shows select results from runs with $v_{\max} \in \{0.05, 0.1, 0.2, 0.3, 0.4\}$ and provides further details on the simultaneous number and energy conservation properties of the scheme when applied to the streaming Doppler shift problem. For comparison, we have also run all the models with the spectral redistribution turned off. The relative change in the total number of particles (top left panel) is at the level of machine precision for all values of $v_{\max}$ (with and without the spectral redistribution), which is expected since the two-moment model is formulated in number conservative form.

As mentioned earlier, the moments evolve close to the boundary of the realizable set in this test, and the realizability-enforcing limiter is frequently invoked to enforce pointwise realizability in all elements. From the lower left panel in Fig. 11, we observe that the minimum value of the limiter parameter $\theta_{\boldsymbol{K}}^{\boldsymbol{\mathcal{U}}}$ (solid lines) varies between 0 and 0.1 for $t \lesssim 10$, while the average value of $\theta_{\boldsymbol{K}}^{\boldsymbol{\mathcal{U}}}$ (dashed lines) grows from about 0.8 to 1 during this time. For $t \leq 10$, a discontinuity in the moments, driven by the inner boundary condition, propagates through the domain and is mainly responsible for triggering the realizability-enforcing limiter. Recall that $\theta_{\boldsymbol{K}}^{\boldsymbol{\mathcal{U}}} = 0$ implies

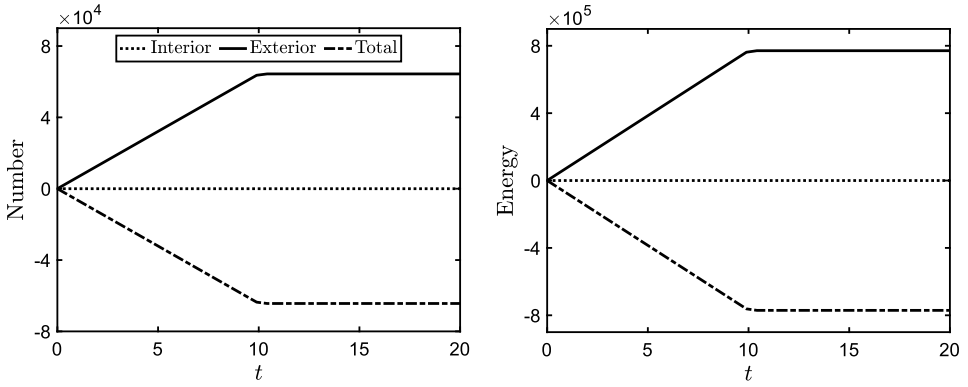


Fig. 10. Plot of Eulerian-frame number (left) and energy (right) balance versus time for the streaming Doppler shift problem for a run with $v_{\max} = 0.2$. In the left panel, ΔN_{int} (solid), ΔN_{ext} (dash-dotted), and $\Delta N_{\text{int}} + \Delta N_{\text{ext}}$ (dotted), see Eq. (142), are plotted. Similarly, in the right panel, ΔE_{int} (solid), ΔE_{ext} (dash-dotted), and $\Delta E_{\text{int}} + \Delta E_{\text{ext}}$ (dotted), see Eq. (143), are plotted.

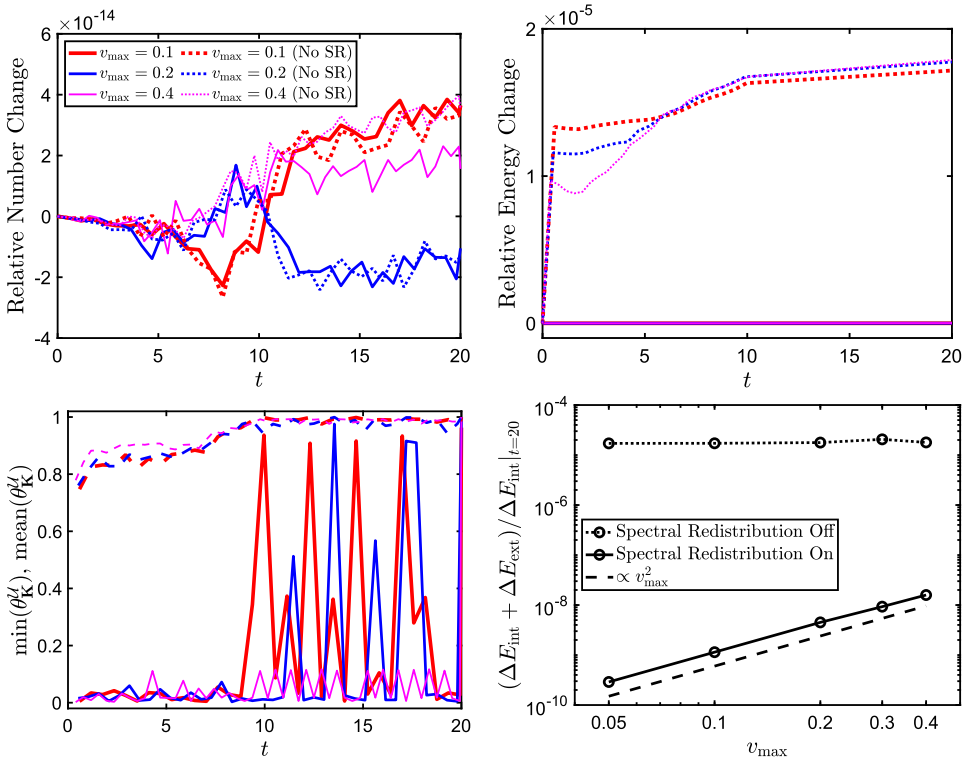


Fig. 11. Results from the streaming Doppler shift problem for models with various values of $v_{\max}$. In the upper left panel the relative change in the Eulerian-frame particle number, defined as $(\Delta N_{\text{int}} + \Delta N_{\text{ext}})/\Delta N_{\text{int}}(t = 20)$, is plotted versus time. Similarly, in the upper right panel, the relative change in the Eulerian-frame energy, defined as $(\Delta E_{\text{int}} + \Delta E_{\text{ext}})/\Delta E_{\text{int}}(t = 20)$, is plotted. In these panels, solid lines represent results obtained with the fiducial algorithm with the spectral redistribution on, while dotted lines represent results obtained with the spectral redistribution turned off. In the lower left panel, the minimum (solid) and mean (dashed) value of the limiter parameter $\theta_{\boldsymbol{K}}^{\boldsymbol{u}}$ (see Algorithm 1) are plotted versus time for the fiducial models with the spectral redistribution on. Results with $v_{\max}$ set to 0.1, 0.2, and 0.3 are plotted with red, blue, and magenta curves, respectively. In the lower right panel, the relative change in the Eulerian-frame energy at $t = 20$ is plotted versus $v_{\max}$ for models where the spectral redistribution is on (solid lines) and off (dotted lines) (the dashed black reference line is proportional to $v_{\max}^2$).

full limiting where all the moments within an element are set to their respective cell average, while $\theta_{\boldsymbol{K}}^{\boldsymbol{u}} = 1$ implies no limiting. After $t \approx 10$, the average $\theta_{\boldsymbol{K}}^{\boldsymbol{u}}$ value hovers around unity, but a few elements still require significant limiting when $t > 10$, especially for the $v_{\max} = 0.4$ model, with the minimum $\theta_{\boldsymbol{K}}^{\boldsymbol{u}}$ still dropping down close to zero. Closer inspection reveals that a few elements around the location of the velocity gradients ($x^1 \in [2, 3.5)$ and $x^1 \in [6.5, 8)$) require limiting beyond $t = 10$.

The relative change in the Eulerian-frame energy is plotted in the top right panel of Fig. 11, which reveals a significant improvement in conservation for the fiducial models with the spectral redistribution turned on (solid lines), when compared to models without the spectral redistribution (dotted lines). (We compute the relative change in number and energy by normalizing by interior values at

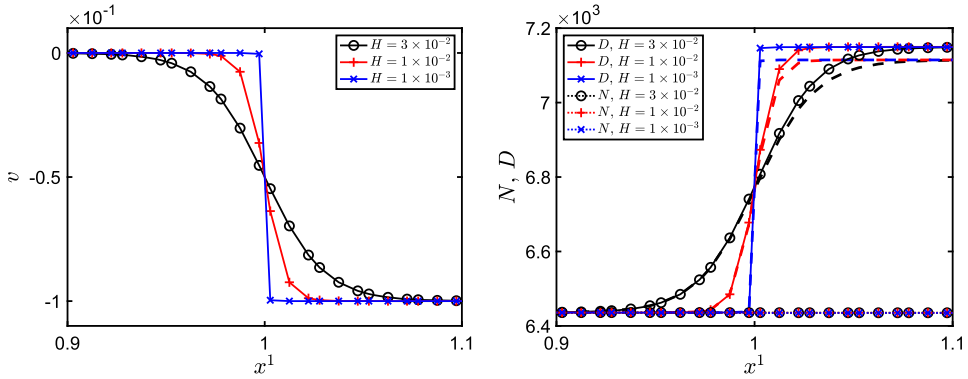


Fig. 12. Results from the transparent shock problem for $v_{\max} = -0.1$ and various values of the shock width parameter H, defined in Eq. (144). In the left panel, velocity profiles are plotted versus position for $H = 3\times10^{-2}$, 10^{-2}, and 10^{-3} (black, red, and blue curves, respectively). In the right panel, the numerical and analytic (special relativistic) comoving-frame number densities (solid lines with markers and dashed lines, respectively), and the numerical Eulerian-frame number densities (dotted lines with markers) are plotted versus position. For the results displayed in the right panel, the line colors correspond to the velocity profile with the matching color plotted in the left panel.

$t = 20$, when the system is in steady state, because the initial interior values are close to zero.) For the models without the spectral redistribution, the relative change in total energy immediately jumps to about 1×10^{-5}, and continues to grow for later times, while it remains relatively constant for the fiducial models. This implies that the realizability-enforcing limiter is the main driver of Eulerian-frame energy nonconservation for this test, and not the inherent nonconservation properties of the continuum $O(v)$ two-moment model. (Velocity jumps across elements are small, and we infer from our results that their contribution to energy nonconservation is negligible.) The relative change in the Eulerian-frame energy at $t = 20$ is plotted versus $v_{\max}$ in the lower right panel of Fig. 11. For the models with the spectral redistribution turned off, the relative change is essentially independent of $v_{\max}$, while Eulerian-frame energy violations grow as $v_{\max}^2$ for the fiducial models. Note that the spectral redistribution only recovers energy conservation violations caused by the realizability-enforcing limiter. Energy conservation violations caused by the use of the $O(v)$ two-moment model are unaffected by the spectral redistribution. Since we observe the $v_{\max}^2$ scaling when using the spectral redistribution, we posit that the DG discretization maintains consistency with the continuum model on this aspect.

8.5. *Transparent shock*

In this test, we investigate the performance of the method when the background velocity gradient is varied. We consider a one-dimensional spatial domain $D_{x^1} = [0, 2]$, set the opacities $\chi = \sigma = 0$, and the velocity field $\boldsymbol{v} = (v, 0, 0)^{\intercal}$, where

$$v(x^1) = \frac{1}{2} v_{\max} \times \left[1 + \tanh\left((x^1 - 1)/H \right) \right]. \tag{144}$$

We will vary both the velocity magnitude $v_{\max}$ and gradient, parameterized by the length scale H. The energy domain is again $D_\varepsilon = [0, 50]$. We discretize the spatial and energy domains using 80 and 32 elements, respectively, and use quadratic elements ($k = 2$) and SSPRK3 time stepping. Then, $\Delta x^1 = 0.025$, and $\Delta x^1/H = 5/6$, 2.5, and 25, for $H = 3\times10^{-2}$, 10^{-2}, and 10^{-3}, respectively. We use the same boundary conditions as in the Doppler shift test, and the moments are initially set to $\mathcal{D} = 1\times10^{-8}$ and $\mathcal{I}^1 = 0$ for all $(x^1, \varepsilon) \in D_{x^1} \times D_\varepsilon$. With the given initial and boundary conditions, the equations are integrated until $t = 3$, when an approximate steady state has been established.

Fig. 12 shows velocity profiles and comoving-frame and Eulerian-frame number densities versus position around the 'shock' ($x^1 \in [0.9, 1.1]$) for the different values of H for the case with $v_{\max} = -0.1$. (The markers indicate locations of LG quadrature points in each spatial element.) For $H = 3\times10^{-2}$, the shock is resolved by the spatial grid, for $H = 10^{-2}$ it is under-resolved, while for $H = 10^{-3}$, the velocity profile is discontinuous.

The comoving-frame number densities increase across the velocity gradient because of the Doppler effect, increasing the particle energy measured by the comoving observer, who is moving towards the inner boundary. Beyond the shock, the values for the computed comoving-frame number densities (solid lines) are about 0.5% higher than the analytic values obtained using Eq. (138) (dashed lines), and this fact is independent of the value of H. The Eulerian-frame number densities, which should remain unaffected by the presence of the background velocity, are essentially constant across the shock. These results indicate that the method is able to capture Doppler shifts correctly, even when velocity gradients are large.

In Fig. 13, the left panel displays the relative change in the Eulerian-frame energy, defined by the left-hand side of Eq. (143), versus $|v_{\max}|$ at $t = 3$, for various values of H. The right panel displays the relative error in ε_{RMS}. Both panels display results obtained with and without the spectral redistribution. (The relative change in the Eulerian-frame particle number, not shown, is at the level of machine precision for all models.) Considering the results obtained with the spectral redistribution active, in the left panel we observe that, for a given value of H, the relative change in the total energy increases with increasing $|v_{\max}|$, roughly proportional to $|v_{\max}|^2$. Also, for a given value of $|v_{\max}|$, the relative change in total energy increases with decreasing shock width H. For the models with the spectral redistribution turned off, the behavior is different in a large region of the $(v_{\max}, H)$-space. With $H = 3\times10^{-2}$

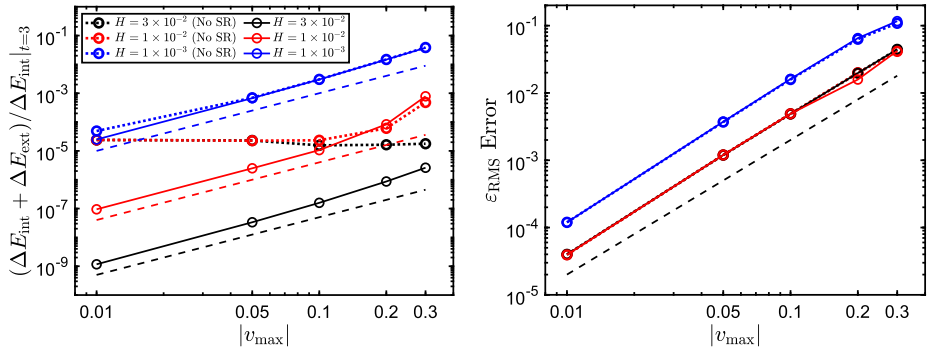


Fig. 13. Results for the transparent shock problem, plotted versus $|v_{\max}|$. The left panel displays the relative change in the Eulerian-frame energy at $t = 3$ for models where the spectral redistribution is on (solid lines) and off (dotted lines). The right panel displays the absolute relative difference in $\varepsilon_{\rm RMS}$ with respect to the exact solution from special relativity at $x^1 = 2$. In both panels, black, red, and blue curves correspond to $H = 3\times10^{-2}$, 10^{-2}, and 10^{-3}, respectively, and the dashed references line are proportional to $|v_{\max}|^2$.

(dotted black line), the relative change in the Eulerian-frame energy at $t = 3$ is around 2×10^{-5}; independent of $v_{\max}$. This can be attributed to the realizability-enforcing limiter. For the models with $H = 10^{-2}$ (dotted red line), the energy change is roughly constant until $|v_{\max}| = 0.1$, when the relative energy change begins to increase with $|v_{\max}|$ in a manner similar to the models with the spectral redistribution activated (solid red line). The models with the steepest velocity gradient ($H = 10^{-3}$; dotted blue line) follow the corresponding models with active spectral redistribution, and the relative change in the Eulerian-frame energy increases as $|v_{\max}|^2$ for all $|v_{\max}|$. From this we conclude that the spectral redistribution can help to recover the $O(v^2)$ Eulerian-frame energy conservation property of the $O(v)$ two-moment model for small velocities and velocity gradients. The relative error in $\varepsilon_{\rm RMS}$, which increases as $|v_{\max}|^2$ for all models, is essentially unaffected by the spectral redistribution.

8.6. Transparent vortex

The final test, inspired by the test in Section 4.2.3 of [4], considers evolution in a two-dimensional spatial domain $D_{x^1}\times D_{x^2} = [-5,5]\times[-5,5]$. We set the opacities $\chi = \sigma = 0$, and the velocity field is given by $\boldsymbol{v} = [v^1, v^2, 0]^\intercal$, where

$$v^1(x^1,x^2) = -v_{\max}\, x^2 \exp\left[(1-r^2)/2\right], \tag{145a}$$

$$v^2(x^1,x^2) = \quad v_{\max}\, x^1 \exp\left[(1-r^2)/2\right], \tag{145b}$$

and $r = \sqrt{(x^1)^2 + (x^2)^2}$. The energy domain is $D_\epsilon = [0,50]$, and we discretize the spatial and energy domains using 48×48 and 32 elements, respectively. We use quadratic elements ($k = 2$) and SSPRK3 time stepping. The upper left panel in Fig. 14 shows the velocity field for the case with $v_{\max} = 0.1$. The main purpose of this test is to investigate the robustness of the method in configurations where the radiation field propagates through a spatially variable velocity field with various relative angles between the radiation flux and velocity vectors. The moments are initially set to $\mathcal{D} = 1\times10^{-8}$ and $\mathcal{I}^1 = \mathcal{I}^2 = 0$ for all $(x^1,x^2,\varepsilon)\in D_{x^1}\times D_{x^2}\times D_\epsilon$. At the inner x^1 boundary, we impose an incoming, radiation field with a Fermi-Dirac spectrum: We set $\mathcal{D}(\varepsilon, x^1 = -5, x^2) = 0.05/[\exp(\varepsilon/3 - 3) + 1]$, $\mathcal{I}^1(\varepsilon, x^1 = -5, x^2) = 0.95\times\mathcal{D}(\varepsilon, x^1 = -5, x^2)$, and $\mathcal{I}^2(\varepsilon, x^1 = -5, x^2) = 0$, so that the flux factor is $h = 0.95$. With these initial and boundary conditions, the moment equations are integrated until a steady state is reached ($t = 20$).

In the upper right panel in Fig. 14, the solid lines represent numerical energy spectra at spatial locations indicated with solid markers of matching color in upper left panel. At the location of the black marker the velocity is close to zero, and thus the black line represents the spectrum of the incoming radiation. The red and blue markers are located where $\boldsymbol{v} = (v_{\max}, 0, 0)^\intercal$ and $\boldsymbol{v} = (-v_{\max}, 0, 0)^\intercal$, respectively, and the spectra at these locations are, respectively, red- and blue-shifted relative to the spectrum sampled at the location of the black marker. Analytic spectra at the locations of the black, red, and blue markers are plotted with dotted lines, which indicate good agreement between numerical and analytical solutions across all energies. At the locations of the black, red, and blue markers, we find that $\varepsilon_{\rm RMS}$ is approximately 15.6, 14.2, and 17.2, respectively. At the location of the magenta marker, which is placed on the opposite side of the vortex (relative to the black marker), the velocity is again close to zero, and it is expected that the spectrum at this location agrees with the spectrum at the location of the black marker. Comparing the solid black and magenta lines in the upper right panel, we observe that the spectral number density is consistently higher at the location of the magenta marker (by a constant factor of about 1.07). Comparing $\varepsilon_{\rm RMS}$ at the two locations, we find that the relative difference is less than 10^{-4}.

The lower left panel in Fig. 14 plots the relative change in the Eulerian-frame energy versus time for models with $v_{\max}\in\{0.01, 0.03, 0.1\}$. Results obtained with the spectral redistribution on are plotted with solid lines, while dotted lines correspond to results with the spectral redistribution turned off. For all models, the relative change in the Eulerian-frame energy is less than 10^{-4}. For the models with $v_{\max} = 0.1$, the relative change reaches the largest amplitudes for $t\in[4,7]$, when a radiation front, driven by the boundary condition at $x^1 = -5$, propagates through the vortex. The model with the spectral redistribution makes a better recovery than the corresponding model with the spectral redistribution turned off. For smaller $v_{\max}$, the relative change in the Eulerian-frame energy is clearly much smaller when the spectral redistribution is active. These results demonstrate the contribution to Eulerian-frame

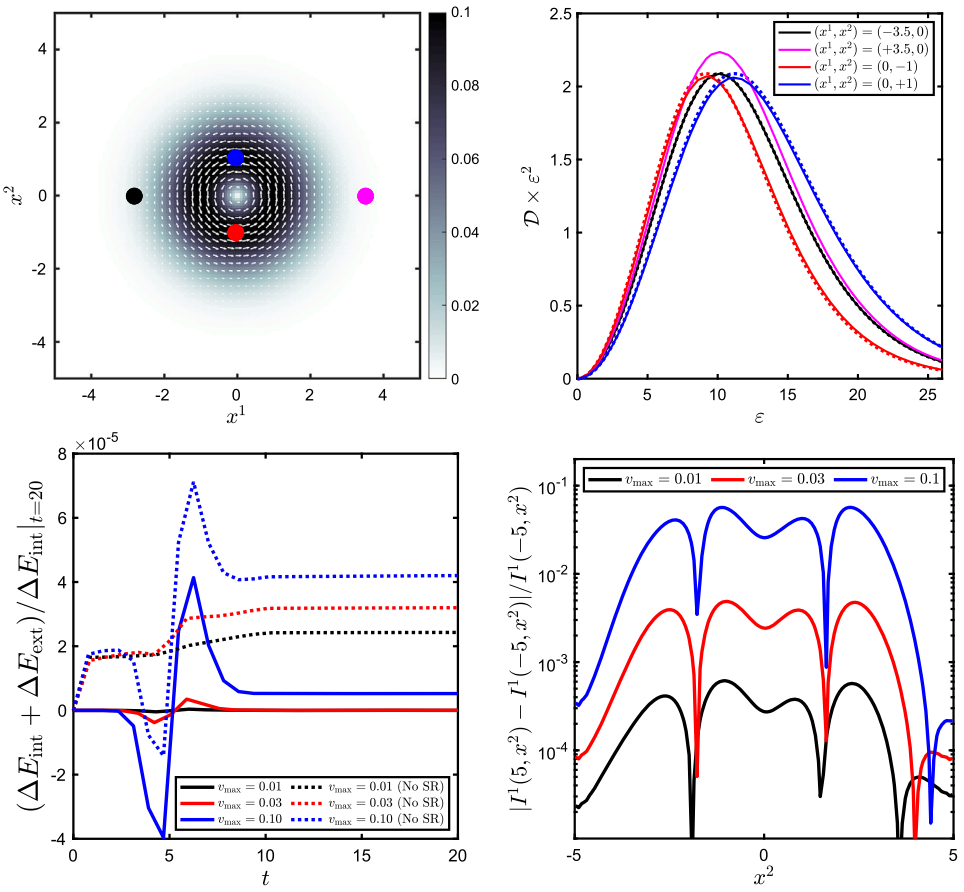


Fig. 14. Results for the transparent vortex problem. In the upper left panel, the magnitude of the velocity, for the case with $v_{\max} = 0.1$, is displayed in grayscale with velocity vectors overlaid. The black, magenta, red, and blue markers indicate spatial positions for which we plot numerical energy spectra in solid lines in the upper right panel, with line colors corresponding to the marker colors in the upper left panel. The dotted lines are analytic spectra obtained from special relativistic considerations using the local three-velocity. In the lower left panel, the relative change in Eulerian-frame energy is plotted versus time for models with $v_{\max}$ in Eq. (145) set to 0.01 (black lines), 0.03 (red lines), and 0.1 (blue lines). Results obtained with and without the spectral redistribution are plotted using solid and dotted lines, respectively. The lower right panel plots the relative difference between the incoming and outgoing particle fluxes in the x^1-direction.

energy nonconservation caused by the realizability-enforcing limiter. For both suites of models (spectral redistribution on or off), the relative change in the Eulerian-frame number (not shown) is at the level of machine precision for all models.

The lower right panel in Fig. 14, similar to Figure 6 (b) in [4], shows, for $t = 20$, the relative difference between the energy integrated x^1-component of the number flux densities evaluated at the inner and outer boundaries of D_{x^1}, defined as $|I^1(5, x^2) - I^1(-5, x^2)|/I^1(-5, x^2)$. As discussed by Just et al. [4], this quantity should vanish for exact calculations, while errors of $O(v^2)$ are to be expected for the $O(v)$ two-moment model. Comparing with their results, the curves plotted in our figure share similar features. Moreover, for $v_{\max} = 0.01$, the maximum relative difference is 6.15×10^{-4}, for $v_{\max} = 0.03$ it is 4.87×10^{-3}, while it is 5.68×10^{-2} for $v_{\max} = 0.1$; i.e., the maximum error grows as $v_{\max}^2$.

Despite the growing (with $v_{\max}$) relative difference between the number fluxes at the inner and outer boundaries in the x^1-direction, we point out that, due to number conservation, the integrated number fluxes through the inner and outer boundaries balance each other. That is, in the steady state at $t = 20$, $\int_{D_{x^2}} I^1(-5, x^2)\, dx^2 = \int_{D_{x^2}} I^1(5, x^2)\, dx^2$. However, the distribution of particles along the x^2-direction becomes nonuniform in the wake of the vortex, while a uniform distribution is expected as $|\boldsymbol{v}| \to 0$. We illustrate this further in Fig. 15. The left panel shows that, within the vortex ($r \lesssim 2$), the comoving-frame number density is higher than the reference value D_0 for $x^2 > 0$, and lower than D_0 for $x^2 < 0$, which is consistent with the Doppler shift of the spectra in the respective regions. In the wake of the vortex, the comoving-frame number density is relatively higher in the region centered around $x^2 = 0$, while it is lower further away (compare red and blue regions for $x^1 \gtrsim 2$ in the left panel in Fig. 15). The Eulerian-frame number density is relatively unaffected by the vortex for $x^1 < 0$, but exhibits a spatial distribution similar to the comoving-frame number density in the wake. In contrast, the spatial distribution of the RMS energy is more consistent with expectations: Within the vortex, $\varepsilon_{\rm RMS} > \varepsilon_{\rm RMS,0}$ for $x^2 > 0$, while $\varepsilon_{\rm RMS} < \varepsilon_{\rm RMS,0}$ for $x^2 < 0$. Moreover, the RMS energy returns to the reference value in the wake of the vortex, with almost uniform distribution along the x^2-direction. We do not have a complete theoretical explanation for the spatial distribution of the number densities in the wake of the vortex, but suspect that the two-moment approximation and the associated

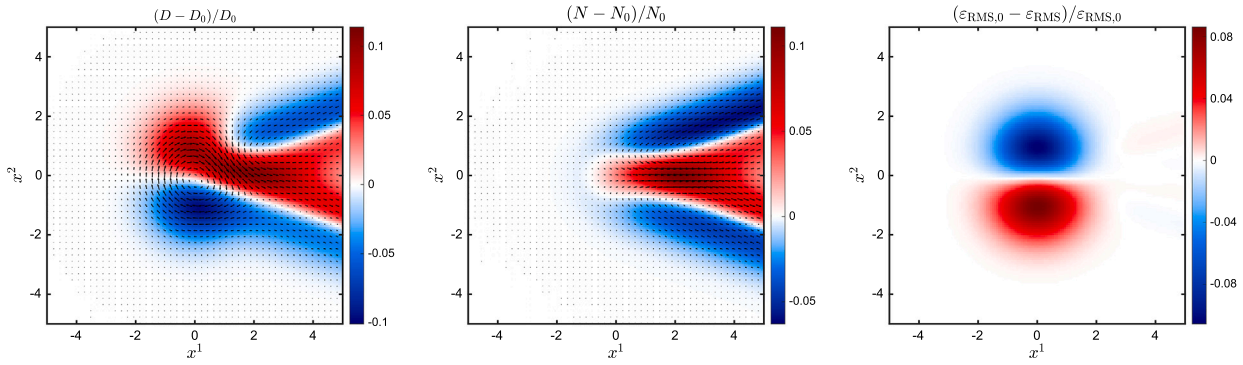


Fig. 15. Results for the transparent vortex problem at $t = 20$ for a model with $v_{\max} = 0.1$. The left panel shows the relative deviation in comoving-frame number density from $D_0 = D(x^1 = -5, x^2)$, $(D - D_0)/D_0$, with vectors of the comoving-frame number flux $(I^1 - I^1_0, I^2)^\intercal$ overlaid, where $I^1_0 = I^1(x^1 = -5, x^2)$ is the first component of the comoving-frame number flux density at the inner boundary in the x^1-direction, which is subtracted to better illustrate the flow, since $|I^2| \ll |I^1|$ generally holds. Similarly, the middle panel shows the corresponding relative deviation in the Eulerian-frame number density $(N - N_0)/N_0$, with vectors of the Eulerian-frame number flux $(F^1_N - F^1_{N,0}, F^2_N)^\intercal$. The right panel shows the relative deviation in the RMS energy, $(\varepsilon_{\mathrm{RMS},0} - \varepsilon_{\mathrm{RMS}})/\varepsilon_{\mathrm{RMS},0}$, where $\varepsilon_{\mathrm{RMS},0} = \varepsilon_{\mathrm{RMS}}(x^1 = -5, x^2)$.

closure, which assumes that the radiation field is axisymmetric about a preferred direction in momentum space [49], is insufficient for capturing relativistic aberration effects.

8.7. Performance evaluation

To demonstrate the GPU functionality and performance characteristics of the DG-IMEX method as implemented in THORNADO, we consider the Streaming Doppler Shift test, described in Section 8.4, with $v_{\max} = 0.1$. To more accurately capture a production workload, the tests are performed in three spatial dimensions, with the number of elements similar to what would be used for a single process invoking THORNADO in a multiphysics simulation. The benchmark is run in two configurations, using tensor product polynomials of degree $k = 1$ and $k = 2$, respectively. The SSPRK2 time stepper is used for both configurations. For $k = 1$, we use 16 energy elements and $96 \times 3 \times 3$ spatial elements, while 12 energy elements and $64 \times 2 \times 2$ spatial elements are used for $k = 2$ — thus keeping the total number of spatial degrees of freedom the same. Our goal is to provide a high-level demonstration of performance characteristics and the relative cost of main algorithmic components, while we defer a rigorous performance analysis to future work.

The tests are performed on a single node of the Summit computer at the Oak Ridge Leadership Computing Facility (OLCF). Each Summit node has 2 IBM POWER9 CPUs and 6 NVIDIA V100 GPUs, but here we limit our benchmarks to a single CPU or GPU. For the CPU runs, we use seven cores with one thread per core as this is the number of cores that would be available to one process if we divide the resources equally with one GPU per process. All runs use version 22.5 of the NVIDIA `nvfortran` compiler with standard `-O2` optimizations. Optimized linear algebra libraries are provided by IBM ESSL (v6.3.0) on the CPU and NVIDIA cuBLAS (v11.0.3) on the GPU. For the GPU runs, all computations are done on the GPU using OpenACC and libraries; the CPU process is only used to launch kernels and manage data transfer. In both cases, the salient metric is wall-time per time step (lower is better).

Fig. 16 shows a breakdown of the relative cost associated with evaluating the major components of the explicit phase-space advection operator. The polynomial degree has little effect on the absolute wall-time, especially for the GPU runs. For the CPU runs, the relative cost of linear algebra (`MatMul`) is somewhat higher when $k = 2$. As can be seen comparing the right and left panels, the initial guess in the conserved-to-primitive calculation can have a non-trivial impact on the total wall-time by reducing the total number of solver iterations. We measure a total speedup factor of 8–10 for the V100 relative to the multi-core CPU runs on the POWER9. Notably, the relative cost for linear algebra and limiters becomes negligible when using the GPU, and the majority of the computational cost is shifted to the iterative conserved-to-primitive calculations. We speculate that one approach to further improve the performance would be to combine the calculation of all of the primitive moments on the quadrature set $\widetilde{S}^{\boldsymbol{K}}_{\otimes}$, defined in Eq. (52), into a single kernel, rather than to calculate them separately for each evaluation of $\mathcal{F}^i$ and $\mathcal{F}^\epsilon$, defined in Eqs. (39)–(40), which results in some duplicate evaluations. While these savings may be significant for the phase-space advection problem considered here, refactoring will be considered in the context of a more physics-complete implementation. With more realistic collision terms included, the relative cost of the explicit phase-space advection part is expected to be small (see, e.g., [58,36]).

9. Summary and conclusions

We have proposed and analyzed a realizability-preserving numerical method for evolving a spectral two-moment model for neutral particles interacting with a moving background fluid. This number-conservative moment model is based on comoving-frame momentum coordinates, includes special relativistic corrections to $O(v)$, and, as a result, contains velocity-dependent terms accounting for spatial advection, Doppler shift, and angular aberration. The nonlinear two-moment model solves for comoving-frame angular moments, representing number density and components of the number flux density, and is closed by expressing higher-order moments (rank-two and rank-three tensors) in terms of the evolved moments using the maximum entropy closure (both exact and approximate)

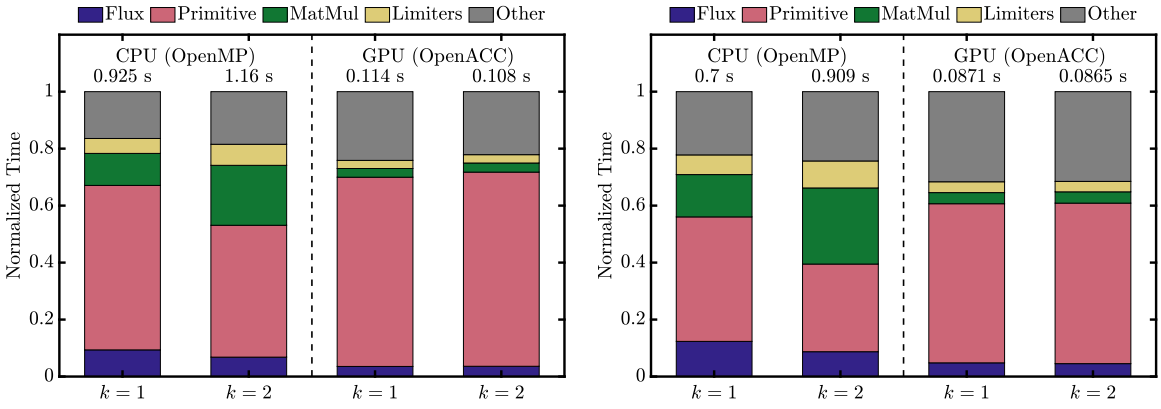


Fig. 16. Breakdown of normalized wall-time for components of the Streaming Doppler Shift test problem as implemented in THORNADO. The left panel uses an initial guess of $\mathcal{M}^{[0]} = (\mathcal{N}, \mathbf{0})^{\intercal}$ in the conserved-to-primitive calculation, and the right panel uses $\mathcal{M}^{[0]} = \mathcal{U} = (\mathcal{N}, \mathcal{G})^{\intercal}$. Absolute wall-clock times per time step are shown above each bar. `Flux` captures the calculation of fluxes $\mathcal{F}^i$ and $\mathcal{F}^\epsilon$ in Eqs. (39)–(40). `MatMul` represents the matrix-matrix multiplications used throughout the explicit operator, e.g., to evaluate polynomials $\mathcal{U}_h$ in quadrature points on all elements. `Primitive` captures the iterative conserved-to-primitive calculation described in Section 4.3.1. `Limiters` includes the application of the realizability-enforcing limiter described Section 5.2 and the spectral redistribution described in Section 6.2. `Other` is used to capture all remaining wall-time spent in the explicit step.

due to Minerbo [21]. The two-moment model is closely related to that promoted in [1], predicts wave speeds bounded by the speed of light (Proposition 2), and is consistent, to $O(v)$, with Eulerian-frame energy and momentum conservation (Proposition 1).

The numerical method is based on the DG method to discretize phase-space, and IMEX time stepping, where the phase-space advection part is integrated with explicit methods, and the collision term is integrated with implicit methods. The discretized spatial and energy derivative terms in the moment equations have been equipped with tailored numerical fluxes, which in the case of exact moment closure (Assumption 1) allow us to derive explicit time-step restrictions that guarantee realizable cell-averaged moments due to these terms, and $\mathcal{N} > 0$ overall. Unfortunately, a corresponding time-step restriction was not found for the source terms associated with phase-space advection in the number flux equation to guarantee the second moment realizability condition, in the sense of cell averages, for the evolved moments (i.e., $|\mathcal{G}| \leq \mathcal{N}$) in the general multidimensional case. However, an analysis in the semi-discrete setting revealed that the moments evolve tangentially to the boundary of the realizable domain when $|\mathcal{G}| = \mathcal{N}$, and we found a sufficient time-step restriction to guarantee realizable cell averages in the one-dimensional, planar geometry case. Given a positive cell-averaged number density, a realizability-enforcing limiter is proposed to recover pointwise moment realizability in each element. Specific properties of the IMEX scheme (i.e., convex-invariance, as defined in [35]) extend the applicability of our results beyond the forward-backward Euler sequence analyzed in detail.

Retention of specific $O(v)$ terms in the time derivative of the two-moment system, motivated by the desire to maintain wave speeds bounded by the speed of light and consistency with Eulerian-frame energy and momentum conservation equations, results in increased computational complexity of the numerical scheme in two (related) ways. First, since the evolved moments are nonlinear functions of the primitive moments used to close the moment equations, a nonlinear system must be solved to recover primitive moments from evolved moments. Second, because the collision operators are formulated in terms of primitive moments, the implicit collision update requires the solution of a similar nonlinear system. For both cases, solution methods have been formulated as fixed-point problems, and we have proposed tailored fixed-point operators in Eqs. (56) and (60), for the primitive recovery and implicit collision solve, respectively. The fixed-point operators are designed to preserve moment realizability in each iteration (subject to mild conditions on the step size), and we have proven convergence for cases with exact *and* approximate moment closures, subject to the additional constraint $|\boldsymbol{v}| \leq \sqrt{2} - 1$, which is mild when considering the applicability of the $O(v)$ model. Numerically, we did not observe convergence failures for the primitive recovery problem, even when violating the condition on the velocity, or when combining the algorithm with Anderson acceleration, which our analysis here did not consider.

The proposed algorithm has been implemented and tested against a series of benchmark problems. Using two problems with a constant background velocity — in the streaming and diffusion regimes, respectively — we demonstrate the expected rate of error convergence in the L^2 norm. Additional tests with spatially varying (smooth and discontinuous) background velocity fields — the Streaming Doppler Shift, Transparent Shock, and Transparent Vortex tests — were used to document the robustness of the proposed algorithm, and qualitative accuracy with respect to special relativistic considerations (e.g., correct Doppler shifts) for sufficiently small background velocities. In these tests, the moments evolve close to the boundary of the realizable domain, and the realizability-enforcing limiter is frequently triggered to recover pointwise realizability from (guaranteed) realizable cell averages. Without this recovery procedure, the algorithm fails invariably on these challenging problems.

We have analyzed the simultaneous Eulerian-frame number and energy conservation properties of the proposed method. While the DG method provides flexibility in the approximation spaces to capture conservation properties beyond those inherent to the model formulation (i.e., number conservation in the present setting), the approximation of the background velocity by piecewise polynomials from the DG approximation space, which accommodates discontinuities, can result in Eulerian-frame energy conservation errors that exceed the $O(v^2)$ scaling predicted by the continuum model. However, we found that the realizability-enforcing limiter is the main

contributor to Eulerian-frame energy conservation violations when the background velocity field is smooth and its magnitude is within the range of applicability of an $O(v)$ model. For this reason, a spectral redistribution scheme is proposed to recover conservation violations introduced by the realizability-enforcing limiter. This redistribution scheme trades (spectrally) local number conservation for number *and* energy conservation after integration over the phase-space energy dimension, and has no observed negative impact on solution accuracy, while improving Eulerian-frame energy conservation properties of the method. With the spectral redistribution active, we observe that energy conservation violations scale as $O(v^2)$, in accordance with the continuum model. We emphasize that the spectral redistribution introduces a rescaling of the moments, which does not impact moment realizability. However, the proposed strategy to promote Eulerian-frame energy conservation is not feasible without the realizability-preserving property.

Our goal is to apply the proposed algorithm to model neutrino transport in core-collapse supernova simulations. Several extensions are needed to achieve this goal. First, the collision term must be extended to include a complete set of neutrino weak interactions, and the model extended to include coupling to dynamical equations for the background fluid. Second, because neutrinos are fermions, for which the Pauli exclusion principle implies an upper bound on the phase-space density and associated bounds on the moments, the analysis should be extended to apply to moment closures based on Fermi-Dirac statistics. Third, because special *and* general relativistic effects contribute to the dynamics in nontrivial ways, further development and analysis is required to design realizability-preserving methods for fully relativistic moment models. In the context of the relativistic number-conservative two-moment model, consistency with the Eulerian-frame energy equation is exact at the continuum level, but it will be challenging to replicate this property at the discrete level. Within the framework of our approach, the realizability-enforcing limiter will be necessary in some situations also in the relativistic case, and, since this limiter is not inherently energy conservative, we anticipate the need for a spectral redistribution scheme in this case as well. We believe the methodologies developed in this paper can be helpful in these endeavors, and hope to present progress on addressing these challenges in future work.

CRediT authorship contribution statement

M. Paul Laiu: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis. **Eirik Endeve:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **J. Austin Harris:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation. **Zachary Elledge:** Writing – original draft, Visualization, Investigation. **Anthony Mezzacappa:** Writing – review & editing, Methodology, Conceptualization.

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

Research at Oak Ridge National Laboratory is supported under contract DE-AC05-00OR22725 from the U.S. Department of Energy to UT-Battelle, LLC. This research was supported by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration. This work was supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research via the Scientific Discovery through Advanced Computing (SciDAC) program. This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. This research was supported, in part, by the National Science Foundation's Gravitational Physics Program under grants NSF PHY 1806692 and 2110177.

Appendix A. Boltzmann equation in the $O(v)$ limit

In this appendix we provide the kinetic equation from which the moment equations presented in Section 2 can be derived in a straightforward manner. A derivation from first principles, which leverages the framework of relativistic kinetic theory, is too elaborate to include here. We refer the interested reader to, e.g., [45,2,46,47,65–67], and references therein. For our purpose, we consider the comprehensive work of Munier & Weaver [46,47], which provides an explicit listing of the kinetic equation for neutral

particle transport in the $O(v)$ limit.[5] Specifically, assuming Cartesian spatial coordinates and a time independent three-velocity, Eq. (142) in [47] for the distribution function $f(\omega,\varepsilon,\boldsymbol{x},t)$ takes the form

$$(1+v^i\ell_i)\,\partial_t f+(\ell^i+v^i)\,\partial_i f+\left[3f-\frac{1}{\varepsilon^2}\frac{\partial}{\partial\varepsilon}\left(\varepsilon^3 f\right)\right]\ell^i\ell^k\,\partial_i v_k+\frac{\partial f}{\partial\ell^i}\left[\ell^i\ell^j\ell^k\,\partial_j v_k-\ell^j\partial_j v^i\right]=C(f), \tag{A.1}$$

where $\{x^i\}$ and t are Eulerian space and time coordinates, respectively, while ε and $\omega=\{\vartheta,\varphi\}$ are spherical-polar momentum coordinates associated with the inertial frame whose Eulerian-frame three-velocity instantaneously coincides with the fluid three-velocity with components v^i. In Eq. (A.1) the Cartesian components of the propagation direction vector in the comoving frame, $\{\ell^1,\ell^2,\ell^3\}$, can be parameterized in terms of spherical-polar momentum space angles as $\{\cos\vartheta,\sin\vartheta\cos\varphi,\sin\vartheta\sin\varphi\}$.

On the right-hand side of Eq. (A.1), we adopt the simplified collision term

$$C(f)=\chi\left(f_0-f\right)+\sigma\left(\mathcal{D}-f\right), \tag{A.2}$$

where f_0, χ, and σ are the equilibrium distribution, absorption opacity, and scattering opacity, respectively (all assumed to be independent of ϑ and φ), and $\mathcal{D}$ is the zeroth moment defined in Eq. (1).

Straightforward manipulations, noting that, to $O(v)$, $\{\ell^i\}$ are independent of $\{x^i\}$ and $\partial\ell^i/\partial\ell^j=\delta^i{}_j$, brings Eq. (A.1) into conservative form

$$\partial_t\left((1+v^i\ell_i)f\right)+\partial_i\left((\ell^i+v^i)f\right)-\frac{1}{\varepsilon^2}\frac{\partial}{\partial\varepsilon}\left(\varepsilon^3 f\,\ell^i\ell_k\,\partial_i v^k\right)+\frac{\partial}{\partial\ell^i}\left(f\left(\ell^i\ell^k\ell_l\,\partial_k v^l-\ell^j\,\partial_j v^i\right)\right)=C(f). \tag{A.3}$$

Eq. (A.3) provides an easy path to the angular moment equations presented in Section 2. Taking the zeroth moment of Eq. (A.3), noting that $\frac{1}{4\pi}\int_{\mathbb{S}^2}\partial\left[\ldots\right]/\partial\ell^i d\omega=0$ [47], results in Eq. (2). Similarly, multiplying Eq. (A.3) by ℓ_j and integrating over angles results in Eq. (3).

Alternatively, the moment equations can be derived from the general relativistic, number-conservative two-moment model presented in Section 4.7.3 in [7], after taking the limit of flat spacetime and specializing to Cartesian spatial coordinates (i.e., setting the lapse function $\alpha\to1$, the shift vector components $\beta^i\to0$, and the components of the spatial metric $\gamma_{ij}\to\delta_{ij}$), and retaining velocity-dependent terms to $O(v)$.

Appendix B. Technical proofs

B.1. Various bounds for the exact and approximate Eddington factors

In the following lemma, we list several bounds on functions dependent on the exact or approximate Eddington factors (ψ or ψ_{a}). These bounds are used in the proofs of Lemmas 8 and 9 in Appendix B.2 and Appendix B.3, respectively, as well of the proof of Lemma 11 in Section 5.5.

Lemma 12. *Let ψ be the Eddington factor in the exact Minerbo closure as given in Eq. (26) and let*

$$\phi_1:=3\psi-1-3\psi' h\quad\text{and}\quad\phi_2:=(3\psi-1)h^{-1}. \tag{B.1}$$

Then, the following bounds hold when $h\in[0,1]$.

(a) $-4\le\phi_1\le0$,
(b) $\phi_2^2-\psi'\phi_2\ge0$,
(c) $3(\psi')^2-3\psi'\phi_2+\phi_2^2\ge0$,
(d) $\partial_h(\phi_2^2-\psi'\phi_2+(\psi')^2)>0$.

Moreover, let ψ_{a} be the approximate Eddington factor defined in Eq. (27) and let

$$\phi_{\mathsf{a},1}:=3\psi_{\mathsf{a}}-1-3\psi_{\mathsf{a}}' h\quad\text{and}\quad\phi_{\mathsf{a},2}:=(3\psi_{\mathsf{a}}-1)h^{-1}. \tag{B.2}$$

Then the bounds (a)–(d) *hold when (ψ,ϕ_1,ϕ_2) are replaced by $(\psi_{\mathsf{a}},\phi_{\mathsf{a},1},\phi_{\mathsf{a},2})$. In addition, the following two bounds hold for the approximate Eddington factor when $h\in[0,1)$.*

(e) $\psi_{\mathsf{a}}-h^2-\frac{1}{4}(1-\psi_{\mathsf{a}})^2\ge0$,
(f) $h^2\le\psi_{\mathsf{a}}\le1$.

Since both ψ and ψ_{a} are one-dimensional functions defined between 0 and 1, the proof of the bounds are straightforward but are rather tedious. Instead of giving rigorous proofs for these bounds, we plot the functions of interest in Fig. B.17, from which the bounds can be visually verified.

[5] Note that Munier & Weaver list their equation in terms of the specific intensity $I\propto\varepsilon^3 f$, while we prefer to work with the distribution function f.

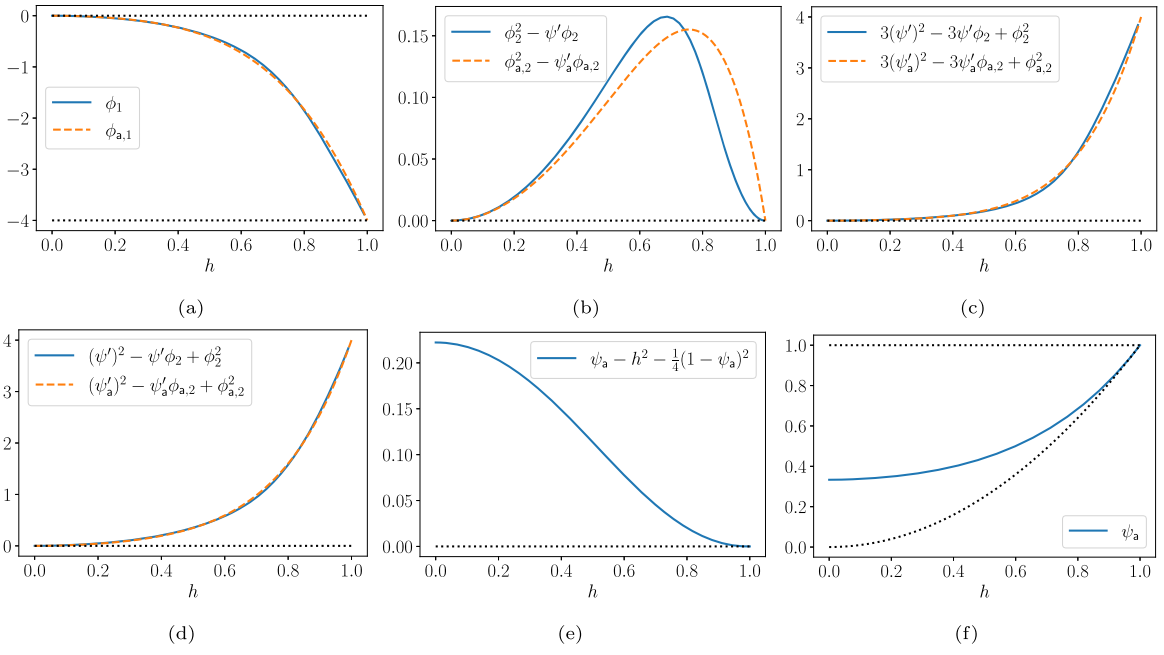


Fig. B.17. Verification of the bounds (a)–(f) in Lemma 12 in the cases when the exact Eddington factor ψ and approximate ψ_{a} are considered.

B.2. Proof of Lemma 8

Proof of Lemma 8. Using the definition of the closure terms k_{ij} in Eq. (10), we have from chain rule that

$$\begin{aligned} v^i \partial_{\mathcal{D}}(\mathsf{k}_{ij}\mathcal{D}) &= v^i \left(\frac{1}{2}\left[3\hat{n}_i\hat{n}_j - \delta_{ij}\right]\frac{\partial \psi}{\partial h}\frac{\partial h}{\partial \mathcal{D}}\mathcal{D} + \frac{1}{2}\left[(1-\psi)\delta_{ij} + (3\psi-1)\hat{n}_i\hat{n}_j\right]\right) \\ &= v^i \left(-\frac{1}{2}\left[3\hat{n}_i\hat{n}_j - \delta_{ij}\right]\psi' h + \frac{1}{2}\left[(1-\psi)\delta_{ij} + (3\psi-1)\hat{n}_i\hat{n}_j\right]\right) \\ &= \frac{1}{2}(3\psi - 1 - 3\psi' h)\left(v^i\hat{n}_i\hat{n}_j - \frac{1}{3}v_j\right) + \frac{1}{3}v_j = \frac{1}{2}\phi_1\left(v^i\hat{n}_i\hat{n}_j - \frac{1}{3}v_j\right) + \frac{1}{3}v_j\,, \end{aligned} \tag{B.3}$$

where $\phi_1 := (3\psi - 1 - 3\psi' h)$ as defined in Eq. (B.1). Since $\|\partial_{\mathcal{D}}(v^i\mathsf{k}_{ij}\mathcal{D})\|^2 = \sum_j \left(v^i\partial_{\mathcal{D}}(\mathsf{k}_{ij}\mathcal{D})\right)^2$, summing up the squares leads to

$$\begin{aligned} \|\partial_{\mathcal{D}}(v^i\mathsf{k}_{ij}\mathcal{D})\|^2 &= \frac{1}{4}\phi_1^2\sum_j\left(v^i\hat{n}_i\hat{n}_j - \frac{1}{3}v_j\right)^2 + \frac{1}{3}\phi_1 v^j\left(v^i\hat{n}_i\hat{n}_j - \frac{1}{3}v_j\right) + \frac{1}{9}v^j v_j \\ &= \left(\frac{\phi_1^2}{12} + \frac{\phi_1}{3}\right)(v^i\hat{n}_i)^2 + \left(\frac{\phi_1^2}{36} - \frac{\phi_1}{9} + \frac{1}{9}\right)v^i v_i\,. \end{aligned} \tag{B.4}$$

It follows from Lemma 12 (a) that $\phi_1(h) \in [-4, 0]$ for $h \in [0, 1]$. Therefore, $\left(\frac{\phi_1^2}{12} + \frac{\phi_1}{3}\right) = \frac{\phi_1}{12}\left(\phi_1 + 4\right) \geq 0$ and we have

$$\|\partial_{\mathcal{D}}(v^i\mathsf{k}_{ij}\mathcal{D})\|^2 \leq \left(\left(\frac{\phi_1^2}{12} + \frac{\phi_1}{3}\right) + \left(\frac{\phi_1^2}{36} - \frac{\phi_1}{9} + \frac{1}{9}\right)\right)v^i v_i = \frac{1}{9}\left(\phi_1 + 1\right)^2 v^i v_i\,. \tag{B.5}$$

Since $\phi_1 \in [-4, 0]$, $\frac{1}{9}\left(\phi_1 + 1\right)^2 \leq 1$, which concludes the proof. □

B.3. Proof of Lemma 9

Proof of Lemma 9. Using the definition of the closure terms k_{ij} in Eq. (10), we have from chain rule that

$$v^i \partial_{\mathcal{I}_k}(\mathsf{k}_{ij}\mathcal{D}) = \frac{1}{2}\psi'\left[3s\hat{n}_j - v_j\right]\hat{n}_k + \frac{(3\psi - 1)}{2h}\left[v_k\hat{n}_j + s\delta_{jk} - 2s\hat{n}_j\hat{n}_k\right]. \tag{B.6}$$

To show $\|\nabla_{\mathcal{I}}(v^i\mathsf{k}_{ij}\mathcal{D})\| \leq 2v$, we prove in the following that

$$\|\nabla_{\mathcal{I}}(v^i\mathsf{k}_{ij}\mathcal{D})\boldsymbol{y}\| \leq 2vy\,, \quad \forall \boldsymbol{y} = \left(y^1, y^2, y^3\right)^{\intercal}, \tag{B.7}$$

where $y := \sqrt{y^i y_i}$. Let $\phi_2 := (3\psi - 1)h^{-1}$ as defined in Eq. (B.1). Then,

$$v^i \partial_{I_k}(\mathsf{k}_{ij}\mathcal{D})y^k = \frac{1}{2}\psi'\left[3\, s\, \hat{n}_j - v_j\right](y^k \hat{n}_k) + \frac{1}{2}\phi_2\left[\hat{n}_j (y^k v_k) + s y_j - 2\, s\, \hat{n}_j (y^k \hat{n}_k)\right]. \tag{B.8}$$

Summing up the squares leads to

$$\begin{aligned}\|\nabla_{\mathcal{I}}(v^i \mathsf{k}_{ij}\mathcal{D})\boldsymbol{y}\|^2 = \sum_j \left(v^i \partial_{I_k}(\mathsf{k}_{ij}\mathcal{D})y^k\right)^2 &= \frac{1}{4}\phi_2^2 s^2 y^2 + \frac{1}{4}\phi_2^2 (y^k v_k)^2 + \frac{1}{4}(\psi')^2 v^2 (y^k \hat{n}_k)^2 \\ &+ \frac{1}{4}\left[3(\psi')^2 - 2\psi'\phi_2\right] s^2 (y^k \hat{n}_k)^2 + \frac{1}{2}\left[\psi'\phi_2 - \phi_2^2\right] s (y^k v_k)(y^k \hat{n}_k).\end{aligned} \tag{B.9}$$

Since $\phi_2^2 - \psi'\phi_2 \geq 0$ (Lemma 12 (b)), we apply the inequality $-2ab \leq a^2 + b^2$ and obtain

$$\frac{1}{2}\left[\psi'\phi_2 - \phi_2^2\right] s (y^k v_k)(y^k \hat{n}_k) \leq \frac{1}{4}\left[\phi_2^2 - \psi'\phi_2\right](y^k v_k)^2 + \frac{1}{4}\left[\phi_2^2 - \psi'\phi_2\right] s^2 (y^k \hat{n}_k)^2. \tag{B.10}$$

Therefore,

$$\begin{aligned}\|\nabla_{\mathcal{I}}(v^i \mathsf{k}_{ij}\mathcal{D})\boldsymbol{y}\|^2 \leq \frac{1}{4}\phi_2^2 s^2 y^2 &+ \frac{1}{4}\left[2\phi_2^2 - \psi'\phi_2\right](y^k v_k)^2 + \frac{1}{4}(\psi')^2 v^2 (y^k \hat{n}_k)^2 \\ &+ \frac{1}{4}\left[3(\psi')^2 - 3\psi'\phi_2 + \phi_2^2\right] s^2 (y^k \hat{n}_k)^2.\end{aligned} \tag{B.11}$$

Further, since $\phi_2^2 \geq 0$, $(\psi')^2 \geq 0$, $2\phi_2^2 - \psi'\phi_2 \geq 0$ (Lemma 12 (b)), and $3(\psi')^2 - 3\psi'\phi_2 + \phi_2^2 \geq 0$ (Lemma 12 (c)), we can take the upper bounds $s^2 \leq v^2$, $(y^k v_k)^2 \leq (vy)^2$, and $(y^k \hat{n}_k)^2 \leq y^2$ to obtain

$$\|\nabla_{\mathcal{I}}(v^i \mathsf{k}_{ij}\mathcal{D})\boldsymbol{y}\|^2 \leq \left[\phi_2^2 - \psi'\phi_2 + (\psi')^2\right](vy)^2. \tag{B.12}$$

It follows from Lemma 12 (d) that $\partial_h(\phi_2^2 - \psi'\phi_2 + (\psi')^2) > 0$, which implies $\max_{h\in[0,1]}\left[\phi_2^2 - \psi'\phi_2 + (\psi')^2\right] = \phi_2^2(1) - \psi'(1)\phi_2(1) + (\psi'(1))^2 = 4$. Thus,

$$\|\nabla_{\mathcal{I}}(v^i \mathsf{k}_{ij}\mathcal{D})\boldsymbol{y}\|^2 \leq 4(vy)^2, \quad \forall \boldsymbol{y} = \left(y^1, y^2, y^3\right)^{\mathsf{T}}, \tag{B.13}$$

which proves the claim. □

References

[1] R.B. Lowrie, D. Mihalas, J. Morel, Comoving-frame radiation transport for nonrelativistic fluid velocities, J. Quant. Spectrosc. Radiat. Transf. 69 (2001) 291–304.
[2] D. Mihalas, B.W. Mihalas, Foundations of Radiation Hydrodynamics, Dover, New York, 1999.
[3] M. Rampp, H.T. Janka, Radiation hydrodynamics with neutrinos. Variable Eddington factor method for core-collapse supernova simulations, Astron. Astrophys. 396 (2002) 361–392, https://doi.org/10.1051/0004-6361:20021398.
[4] O. Just, M. Obergaulinger, H.-T. Janka, A new multidimensional, energy-dependent two-moment transport code for neutrino-hydrodynamics, Mon. Not. R. Astron. Soc. 453 (2015) 3386–3413.
[5] M.A. Skinner, J.C. Dolence, A. Burrows, D. Radice, D. Vartanyan, FORNAX: a flexible code for multiphysics astrophysical simulations, Astrophys. J. Suppl. Ser. 241 (2019) 7.
[6] S.W. Bruenn, J.M. Blondin, W.R. Hix, E.J. Lentz, O.E.B. Messer, A. Mezzacappa, E. Endeve, J.A. Harris, P. Marronetti, R.D. Budiardja, M.A. Chertkow, C.-T. Lee, CHIMERA: a massively parallel code for core-collapse supernova simulations, Astrophys. J. Suppl. Ser. 248 (1) (2020) 11, https://doi.org/10.3847/1538-4365/ab7aff.
[7] A. Mezzacappa, E. Endeve, O.E.B. Messer, S.W. Bruenn, Physical, numerical, and computational challenges of modeling neutrino transport in core-collapse supernovae, Living Rev. Comput. Astrophys. 6 (1) (2020) 4, https://doi.org/10.1007/s41115-020-00010-8.
[8] O. Just, A. Bauswein, R. Ardevol Pulpillo, S. Goriely, H.T. Janka, Comprehensive nucleosynthesis analysis for ejecta of compact binary mergers, Mon. Not. R. Astron. Soc. 448 (1) (2015) 541–567, https://doi.org/10.1093/mnras/stv009.
[9] F. Foucart, Neutrino transport in general relativistic neutron star merger simulations, Living Rev. Comput. Astrophys. 9 (1) (2023) 1, https://doi.org/10.1007/s41115-023-00016-y.
[10] S. Chapman, T. Cowling, The Mathematical Theory of Non-uniform Gases, Cambridge Mathematical Library, Cambridge University Press, 1970.
[11] J.I. Castor, Radiative transfer in spherically symmetric flows, Astrophys. J. 178 (1972) 779–792, https://doi.org/10.1086/151834.
[12] J.R. Buchler, Radiation hydrodynamics in the fluid frame, J. Quant. Spectrosc. Radiat. Transf. 22 (1979) 293–300, https://doi.org/10.1016/0022-4073(79)90119-5.
[13] J.R. Buchler, Radiation transfer in the fluid frame, J. Quant. Spectrosc. Radiat. Transf. 30 (1983) 395–407, https://doi.org/10.1016/0022-4073(83)90102-4.
[14] D. Mihalas, R.I. Klein, On the solution of the time-dependent inertial-frame equation of radiative transfer in moving media to O(v/c), J. Comput. Phys. 46 (1982) 97–137, https://doi.org/10.1016/0021-9991(82)90007-9.
[15] H. Nagakura, K. Sumiyoshi, S. Yamada, Three-dimensional Boltzmann hydro code for core collapse in massive stars. I. Special relativistic treatments, Astrophys. J. Suppl. Ser. 214 (2) (2014) 16, https://doi.org/10.1088/0067-0049/214/2/16.
[16] Y.-F. Jiang, Multigroup radiation magnetohydrodynamics based on discrete ordinates including Compton scattering, Astrophys. J. Suppl. Ser. 263 (1) (2022) 4, https://doi.org/10.3847/1538-4365/ac9231.
[17] A. Burrows, S. Reddy, T.A. Thompson, Neutrino opacities in nuclear matter, Nucl. Phys. A 777 (2006) 356–394, https://doi.org/10.1016/j.nuclphysa.2004.06.012.
[18] H.-T. Janka, K. Langanke, A. Marek, G. Martínez-Pinedo, B. Müller, Theory of core-collapse supernovae, Phys. Rep. 442 (1) (2007) 38–74, https://doi.org/10.1016/j.physrep.2007.02.002.

[19] H.-T. Janka, Explosion mechanisms of core-collapse supernovae, Annu. Rev. Nucl. Part. Sci. 62 (1) (2012) 407–451, https://doi.org/10.1146/annurev-nucl-102711-094901.
[20] T. Fischer, G. Guo, K. Langanke, G. Martinez-Pinedo, Y.-Z. Qian, M.-R. Wu, Neutrinos and nucleosynthesis of elements, arXiv e-prints arXiv:2308.03962, 2023.
[21] G.N. Minerbo, Maximum entropy Eddington factors, J. Quant. Spectrosc. Radiat. Transf. 20 (1978) 541–545.
[22] J. Cernohorsky, S.A. Bludman, Maximum entropy distribution and closure for Bose-Einstein and Fermi-Dirac radiation transport, Astrophys. J. 433 (1) (1994) 450–455.
[23] M. Shibata, K. Kiuchi, Y. Sekiguchi, Y. Suwa, Truncated moment formalism for radiation hydrodynamics in numerical relativity, Prog. Theor. Phys. 125 (2011) 1255–1287.
[24] C.Y. Cardall, E. Endeve, A. Mezzacappa, Conservative 3+1 general relativistic variable Eddington tensor radiation transport equations, Phys. Rev. D 87 (2013) 103004.
[25] N. Vaytet, E. Audit, B. Dubroca, F. Delahaye, A numerical model for multigroup radiation hydrodynamics, J. Quant. Spectrosc. Radiat. Transf. 112 (2011) 1323–1335.
[26] B. Cockburn, C.-W. Shu, Runge-Kutta discontinuous Galerkin methods for convection-dominated problems, J. Sci. Comput. 16 (2001) 173–261.
[27] E.W. Larsen, J.E. Morel, Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes II, J. Comput. Phys. 83 (1989) 212–236.
[28] M.L. Adams, Discontinuous finite element transport solutions in thick diffusive problems, Nucl. Sci. Eng. 137 (3) (2001) 298–333.
[29] J.-L. Guermond, G. Kanschat, Asymptotic analysis of upwind discontinuous Galerkin approximation of the radiative transport equation in the diffusive limit, SIAM J. Numer. Anal. 48 (2010) 53–78.
[30] E. Audit, P. Charrier, J.P. Chièze, B. Dubroca, A radiation-hydrodynamics scheme valid from the transport to the diffusion limit, arXiv:astro-ph/0206281, 2002.
[31] B. Ayuso, J.A. Carrillo, C.-W. Shu, Discontinuous Galerkin methods for the one-dimensional Vlasov–Poisson system, Kinet. Relat. Models 4 (4) (2011) 955–989.
[32] Y. Cheng, I.M. Gamba, P.J. Morrison, Study of conservation and recurrence of Runge–Kutta discontinuous Galerkin schemes for Vlasov–Poisson systems, J. Sci. Comput. 56 (2013) 319–349.
[33] U. Ascher, S. Ruuth, R. Spiteri, Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations, Appl. Numer. Math. 25 (1997) 151–167.
[34] L. Pareschi, G. Russo, Implicit-explicit Runge-Kutta schemes and application to hyperbolic systems with relaxation, J. Sci. Comput. 25 (2005) 129–155.
[35] R. Chu, E. Endeve, C. Hauck, A. Mezzacappa, Realizability-preserving dg-imex method for the two-moment model of fermion transport, J. Comput. Phys. 389 (2019) 62–93.
[36] M.P. Laiu, E. Endeve, R. Chu, J.A. Harris, O.E.B. Messer, A DG-IMEX method for two-moment neutrino transport: nonlinear solvers for neutrino-matter coupling, Astrophys. J. Suppl. Ser. 253 (2) (2021) 52, https://doi.org/10.3847/1538-4365/abe2a8.
[37] C.D. Levermore, Moment closure hierarchies for kinetic theories, J. Stat. Phys. 83 (1996) 1021–1065.
[38] G.W. Alldredge, M. Frank, C.D. Hauck, A regularized entropy-based moment method for kinetic equations, SIAM J. Appl. Math. 79 (5) (2019) 1627–1653.
[39] E. Olbrant, C.D. Hauck, M. Frank, A realizability-preserving discontinuous Galerkin method for the m1 model of radiative transfer, J. Comput. Phys. 231 (17) (2012) 5612–5639.
[40] C.D. Hauck, High-order entropy-based closures for linear transport in slab geometry, Commun. Math. Sci. 9 (1) (2011) 187–205.
[41] L.F. Richardson, Ix. the approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam, Philos. Trans. R. Soc. Lond., Ser. A, Contain. Pap. Math. Phys. Character 210 (459–470) (1911) 307–357.
[42] Y. Saad, Iterative Methods for Sparse Linear Systems, SIAM, 2003.
[43] M. Liebendörfer, O.E.B. Messer, A. Mezzacappa, S.W. Bruenn, C.Y. Cardall, F.K. Thielemann, A finite difference representation of neutrino radiation hydrodynamics in spherically symmetric general relativistic spacetime, Astrophys. J. Suppl. Ser. 150 (1) (2004) 263–316, https://doi.org/10.1086/380191.
[44] B. Müller, H.-T. Janka, H. Dimmelmeier, A new multi-dimensional general relativistic neutrino hydrodynamic code for core-collapse supernovae. I. Method and code tests in spherical symmetry, Astrophys. J. Suppl. Ser. 189 (1) (2010) 104–133, https://doi.org/10.1088/0067-0049/189/1/104.
[45] R.W. Lindquist, Relativistic transport theory, Ann. Phys. 37 (1966) 487–518.
[46] A. Munier, R. Weaver, Radiation transfer in the fluid frame: a covariant formulation. Part I: radiation hydrodynamics, Comput. Phys. Rep. 3 (3) (1986) 127–164, https://doi.org/10.1016/0167-7977(86)90007-9.
[47] A. Munier, R. Weaver, Radiation transfer in the fluid frame: a covariant formulation. Part II: the radiation transfer equation, Comput. Phys. Rep. 3 (3) (1986) 165–208, https://doi.org/10.1016/0167-7977(86)90008-0.
[48] E. Endeve, C.Y. Cardall, A. Mezzacappa, Conservative moment equations for neutrino radiation transport with limited relativity, ArXiv e-prints arXiv:1212.4064, Dec. 2012.
[49] C.D. Levermore, Relating Eddington factors to flux limiters, J. Quant. Spectrosc. Radiat. Transf. 31 (1984) 149–160.
[50] S.C. Noble, C.F. Gammie, J.C. McKinney, L. Del Zanna, Primitive variable solvers for conservative general relativistic magnetohydrodynamics, Astrophys. J. 641 (2006) 626–637.
[51] R.E. Curto, L.A. Fialkow, Recursiveness, positivity, and truncated moment problems, Houst. J. Math. 17 (4) (1991).
[52] J.S. Hesthaven, T. Warburton, Nodal Discontinuous Galerkin Methods: Algorithms, Analysis and Applications, Springer, 2008.
[53] M. Abramowitz, I.A. Stegun, R.H. Romer, Handbook of mathematical functions with formulas, graphs, and mathematical tables, Am. J. Phys. 56 (10) (1988) 958, https://doi.org/10.1119/1.15378.
[54] E. Hairer, S. Nørsett, G. Wanner, Solving Ordinary Differential Equations I: Nonstiff Problems, Springer Series in Computational Mathematics, Springer Berlin Heidelberg, 1993.
[55] X. Zhang, C.-W. Shu, On maximum-principle-satisfying high order schemes for scalar conservation laws, J. Comput. Phys. 229 (2010) 3091–3120.
[56] X. Zhang, C.-W. Shu, On positivity preserving high order discontinuous Galerkin schemes for compressible Euler equations on rectangular meshes, J. Comput. Phys. 229 (2010) 8918–8934.
[57] A. Dubey, K. Weide, J. O'Neal, A. Dhruv, S. Couch, J.A. Harris, T. Klosterman, R. Jain, J. Rudi, B. Messer, M. Pajkos, J. Carlson, R. Chu, M. Wahib, S. Chawdhary, P.M. Ricker, D. Lee, K. Antypas, K.M. Riley, C. Daley, M. Ganapathy, F.X. Timmes, D.M. Townsley, M. Vanella, J. Bachan, P.M. Rich, S. Kumar, E. Endeve, W.R. Hix, A. Mezzacappa, T. Papatheodore, Flash-X: a multiphysics simulation software instrument, SoftwareX 19 (2022) 101168, https://doi.org/10.1016/j.softx.2022.101168.
[58] M.P. Laiu, J.A. Harris, R. Chu, E. Endeve, thornado-transport: Anderson- and GPU-accelerated nonlinear solvers for neutrino-matter coupling, J. Phys. Conf. Ser. 1623 (2020) 012013, https://doi.org/10.1088/1742-6596/1623/1/012013.
[59] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, D. Sorensen, LAPACK Users' Guide, 3rd edition, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1999.
[60] S. Tomov, J. Dongarra, M. Baboulin, Towards dense linear algebra for hybrid GPU accelerated manycore systems, Parallel Comput. 36 (5–6) (2010) 232–240, https://doi.org/10.1016/j.parco.2009.12.005.
[61] C.-W. Shu, S. Osher, Efficient implementation of essentially non-oscillatory shock-capturing schemes, J. Comput. Phys. 77 (1988) 439–471.
[62] D.G. Anderson, Iterative procedures for nonlinear integral equations, J. ACM 12 (4) (1965) 547–560, https://doi.org/10.1145/321296.321305, http://doi.acm.org/10.1145/321296.321305.
[63] H. Walker, P. Ni, Anderson acceleration for fixed-point iterations, SIAM J. Numer. Anal. 49 (4) (2011) 1715–1735, https://doi.org/10.1137/10078356X.

[64] D. Cavaglieri, T. Bewley, Low-storage implicit/explicit Runge–Kutta schemes for the simulation of stiff high-dimensional ODE systems, J. Comput. Phys. 286 (2015) 172–193.
[65] C. Cardall, A. Mezzacappa, Conservative formulations of general relativistic kinetic theory, Phys. Rev. D 68 (2) (2003) 023006, https://doi.org/10.1103/PhysRevD.68.023006.
[66] C.Y. Cardall, E.J. Lentz, A. Mezzacappa, Conservative special relativistic radiative transfer for multidimensional astrophysical simulations: motivation and elaboration, Phys. Rev. D 72 (4) (2005) 043007, https://doi.org/10.1103/PhysRevD.72.043007.
[67] C.Y. Cardall, E. Endeve, A. Mezzacappa, Conservative 3+1 general relativistic Boltzmann equation, Phys. Rev. D 88 (2013) 023011.