# An explicit high-order single-stage single-step positivity-preserving finite difference WENO method for the compressible Euler equations

**David C. Seal** · **Qi Tang** · **Zhengfu Xu** ·
**Andrew J. Christlieb**

**Abstract** In this work we construct a high-order, single-stage, single-step positivity-preserving method for the compressible Euler equations. Space is discretized with the finite difference weighted essentially non-oscillatory (WENO) method. Time is discretized through a Lax-Wendroff procedure that is constructed from the Picard integral formulation (PIF) of the partial differential equation. The method can be viewed as a modified flux approach, where a linear combination of a low- and high-order flux defines the numerical flux used for a single-step update. The coefficients of the linear combination are constructed by solving a simple optimization problem at each time step. The high-order flux itself is constructed through the use of Taylor series and the Cauchy-Kowalewski procedure that incorporates higher-order terms. Numerical results in one- and two-dimensions are presented.

David C. Seal
Department of Mathematics
U.S. Naval Academy
121 Blake Road
Annapolis, MD 21402 USA
Tel.: +1(410) 293-6784
E-mail: seal@usna.edu

Qi Tang
Department of Mathematical Sciences
Rensselaer Polytechnic Institute
Troy, NY 12180, USA

Zhengfu Xu
Department of Mathematical Sciences
Michigan Technological University
Houghton, MI 49931, USA

Andrew J. Christlieb
Department of Mathematics and Department of Electrical and Computer Engineering
Michigan State University
East Lansing, MI 48824, USA

# 1 Introduction

The objective of this work is to define a single-stage, single-step finite difference WENO method that is provably positivity-preserving for the compressible Euler equations. These equations describe the evolution of density $\rho$, momentum $\rho\mathbf{u}$ and energy $\mathscr{E}$ of an ideal gas through

$$\begin{pmatrix} \rho \\ \rho\mathbf{u} \\ \mathscr{E} \end{pmatrix}_{,t} + \nabla_{\mathbf{x}} \cdot \begin{pmatrix} \rho\mathbf{u} \\ \rho\mathbf{u} \otimes \mathbf{u} + p\mathbf{I} \\ (\mathscr{E}+p)\mathbf{u} \end{pmatrix} = 0. \tag{1}$$

The energy $\mathscr{E}$ is related to the primitive variables $\rho$, $\mathbf{u}$ and pressure $p$ by the equation of state that we take to be $\mathscr{E} = \frac{p}{\gamma-1} + \frac{1}{2}\rho\|\mathbf{u}\|^2$. The ratio of specific heats is the constant $\gamma$. Numerical difficulties for solving this system include the following:

– Low (1st and 2nd) order methods generally suffer from an inordinate amount of numerical diffusion. However, they are oftentimes more robust, and in some cases they have provable convergence to the correct entropy solution. Historically, 2nd-order schemes [1–4] have been called "high-resolution" methods when compared to their 1st-order counterparts [5–8].
– High-order methods [9–11] provide greater accuracy and resolution for much less overall computational effort. However, they are oftentimes less robust, and do not necessarily have provable convergence to the correct entropy solution.

In this work, we define a high-order conservative finite difference method based upon the Picard integral formulation of the PDE. We make a further modification to the fluxes and define a numerical scheme that obtains the following properties:

– High-order accuracy in space ($5^{th}$-order) and time ($3^{rd}$-order). Our method can be extended to arbitrary order in space or time.
– A robust scheme that stems from provable positivity preservation of the pressure and density. Numerical results indicate that high-order accuracy is retained with our positivity-preserving limiter turned on.

Our scheme is the first single-stage, single-step numerical method that simultaneously attains high-order accuracy, with *provable* positivity preservation. When compared to other positivity-preserving schemes, our method has the following advantages:

– In order to retain positivity, we only solve one simple optimization problem per time step. Unlike positivity-"preserving" methods that use Runge-Kutta discretizations [12, 13], positivity of our solution is guaranteed during the entire simulation because we do not have internal stages where the solution can go negative.
– Compared to other positivity-preserving schemes [14, 15], the addition of the positivity preserving limiter introduces none of the additional time step restrictions that are often introduced in order to retain positivity.

In addition, our method is amenable to adaptive mesh refinement (AMR) technology. At present, we aim to lay the necessary foundation that would be required to do so. An in depth investigation of this property is the subject of a future study.

## 1.1 An overview of the proposed method

The Euler equations define a system of hyperbolic conservation laws. In 1D, such an equation is given by

$$q_{,t} + f(q)_{,x} = 0, \tag{2}$$

where $q(t,x) : \mathbb{R}^+ \times \mathbb{R} \to \mathbb{R}^m$ is the unknown vector of $m$ conserved quantities and $f : \mathbb{R}^m \to \mathbb{R}^m$ is a prescribed flux function. The conserved variables for the 1D Euler equations are $q = \left( \rho, \rho u^1, \mathscr{E} \right)^T$. A typical finite difference solver for (2) discretizes space with a uniform grid of $m_x$ equidistant points in $\Omega = [a,b]$,

$$x_i = a + \left( i - \frac{1}{2} \right) \Delta x, \qquad \Delta x = \frac{b-a}{m_x}, \qquad i \in \{1, \dots, m_x\}, \tag{3}$$

and looks for a pointwise approximation $q_i^n \approx q(t^n, x_i)$ solution to hold at discrete time levels $t^n$. In a conservative finite difference WENO method, the update of the unknowns is typically defined by

$$q_i^{n+1} = q_i^n - \frac{\Delta t}{\Delta x} \left( F_{i+\frac{1}{2}}^n - F_{i-\frac{1}{2}}^n \right), \tag{4}$$

where the numerical flux $F_{i\pm\frac{1}{2}}^n$ is constructed from a linear combination of the WENO reconstruction procedure applied to stage values from a Runge-Kutta solver.

In this work, we propose the following procedure:

1. Construct a high-order approximation to the *time-averaged fluxes* [9, 16]

$$F_i^n := \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} f(q(t,x_i)) \, dt \tag{5}$$

   at each grid point $x_i$. In this work, we consider the Taylor discretization of Eqn. (5) for conservative finite difference methods.

2. Construct a **high-order in space and time** numerical flux $\hat{F}_{i-\frac{1}{2}}^n$ based upon applying the WENO reconstruction procedure to the time-averaged fluxes [16].

3. Replace the flux constructed in Step 2 with

$$\tilde{F}_{i-\frac{1}{2}}^n := \theta_{i-\frac{1}{2}}^n (\hat{F}_{i-\frac{1}{2}}^n - \hat{f}_{i-\frac{1}{2}}^n) + \hat{f}_{i-\frac{1}{2}}^n, \tag{6}$$

   where $\theta_{i-\frac{1}{2}}^n \in [0,1]$ is found by solving a single optimization problem, and $\hat{f}_{i-\frac{1}{2}}^n$ is a low-order flux that guarantees positivity of the solution. This procedure is described in §4.

4. Insert the result of Step 3 into Eqn. (4), and update the solution.

Steps 1 and 2 have already been proposed in [16], where high-order accuracy is obtained through a flux modification that incorporates the high-order temporal discretization. A review of this procedure is presented in §3. Step 3 can be thought of as a further flux modification, where an automatic switch adjusts between the high-order non positivity-preserving scheme, and a low-order, positivity-preserving scheme. The original idea is attributed to Harten and Zwas [17], but has since been extended to high-order WENO schemes [12]. The details of this procedure are presented in §4 and §5.

## 2 Background

The compressible Euler equations have been an object of study ever since the infancy of numerical methods [1, 5–8]. In recent years, high-order methods have attracted considerable interest because of their ability to obtain higher accuracy on certain problems with an equivalent computational cost of a low-order method. Among many choices of high-order

schemes are the classical essentially non-oscillatory (ENO) method [9], the extensions to finite difference (FD) and finite volume (FV) WENO methods [11, 18, 19], and the discontinuous Galerkin (DG) method [20]. These methods all seek to simultaneously obtain two properties: retain high-order accuracy in smooth regions, and capture shocks without introducing spurious oscillations near discontinuities of the solution. One added difficulty with high-order schemes is the necessity of defining and selecting a high-order time integrator. Runge-Kutta methods applied to the method of lines (MOL) approach is the most widely used discretization for high-order schemes. These methods all treat space and time as separate entities.

Over the past decade, there has been a rejuvenation of interest in high-order single-stage, single-step methods for hyperbolic conservation laws, including the compressible Euler equations. All of these methods are typically based upon a Taylor temporal discretization that uses the Cauchy-Kowalewski procedure to exchange temporal for spatial derivatives. Lax and Wendroff performed this very procedure in 1960 [1], and this technique has since been called the Lax-Wendroff procedure within the numerical analysis community. Methods for defining a second and higher-order single-step version of Godunov's method were investigated in the 1980s [21–23]. The original high-order ENO method of Harten et. al [9] used Taylor series for their temporal discretization, although most of the attention they have received is for their emphasis on the spatial discretization. In 2001, the preliminary definitions for the so-called Arbitrary DERivative (ADER) methods [24–26] were put in place. Additionally, various FD WENO methods with Lax-Wendroff time discretizations have been constructed and tested on the Euler equations [16, 27, 28]. Recent ADER methods have been defined by Balsara and collaborators for hydrodynamics and magnetohydrodynamics [29, 30], and have later been extended to an adaptive mesh refinement (AMR) setting [31]. Other recent work in single-stage, single-step methods for Euler equations includes Lax-Wendroff time stepping coupled with DG [32–34], and high-order Lagrangian schemes [35]. The present work is an extension of the Taylor discretization of the Picard integral formulation that uses finite differences for its spatial discretization [16], which falls into this single-stage, single-step class of methods.

Defining high-order numerical schemes that retain positivity of the solution for hydrodynamics (or magnetohydrodynamics) simulations is genuinely a non-trivial task. This has been an ongoing subject of study even for low and the so-called "high-resolution" schemes [36–43]. All methods that are second or higher order share the same disadvantage that without care, they may violate a natural weak stability condition that the density and pressure need to keep positive, which is necessary to ensure physical meaningfulness of the solution and hyperbolicity of the mathematical problem. Of some of the early positivity works, Perthame and Shu propose a general reconstruction approach to obtain a high-order positivity-preserving finite volume schemes from a low-order scheme [39]. In addition, they prove that the explicit Lax-Friedrichs scheme is positivity-preserving with a CFL number up to 0.5. Later on, a more general result extends the positivity-preserving property to CFL numbers up to 1 for both explicit and implicit Lax-Friedrichs methods [42]. With those building blocks, a positivity-preserving limiter is proposed for DG schemes [44–47] and FD and FV WENO schemes [14, 15] for the Euler equations. In [48], a flux cut-off limiter is also applied to FD WENO schemes to retain positivity. In addition to gas dynamics, plasma physics is another area where retaining positivity of numerical solutions is critical, and therefore has seen recent attention in the literature [49, 50]. For example, collision operators for Vlasov equations require a positive distribution function in order to avoid creating artificial singularities.

Our method is based upon a parameterized flux limiter that can be dated back to at least 1972 with the work of Harten and Zwas [17]. There, the authors propose a second-order shock-capturing *self adjusting hybrid scheme* through a simple linear combination of low- and high-order fluxes that is identical to Eqn. (6). The original idea is to combine a "high-order" flux with a first-order flux such that it has better accuracy in smooth region and produces a smooth profile around shock regions. A similar approach called *flux-corrected transport* is proposed by Boris and Book [51–54], where the purpose of limiting the high-order flux is to control overshoots and undershoots around shock regions. Sod performs an extensive review of these and other classical finite difference methods in a classical paper [2]. Xu and Lian recently extend this work to WENO methods that maintain the maximum-principle-preserving property for scalar hyperbolic conservation laws with the so-called "parametrized flux limiter" [55, 56]. Later on, these limiters are applied to FD WENO schemes on rectangular meshes [57] and FV WENO schemes on triangular meshes [13] for the Euler equations. These limiters are also applied to magnetohydrodynamics with a constrained transport framework [58]. The basic idea for all of these methods is the same: modify the high-order non positivity-preserving numerical flux by a linear combination of a low- and high-order flux in order to retain positivity of the solution. The modification is carefully designed so that the high-order accuracy remains.

The purpose of this work is to define a single-stage, single-step finite difference WENO method that is provably positivity-preserving for the compressible Euler equations. Of the various finite difference schemes constructed from the Picard integral formulation [16], we begin with the Taylor discretization, and then apply recently developed flux limiters [13, 57] in order to retain positivity of the solution. One advantage of the chosen limiter is that positivity is preserved without introducing additional time step restrictions, however, our primary contribution is that the present method is the first scheme to simultaneously be high-order, single-stage, single-step and have provable positivity preservation.

The outline of this paper is as follows. In §3, we briefly review the high-order finite difference WENO method that is based upon the Picard integral formulation of the PDE with a Taylor temporal discretization [16]. In §4 and §5, we present the positivity-preserving limiter for PIF-WENO schemes applied to the compressible Euler system in single and multiple dimensions. Numerical examples of the positivity-preserving PIF-WENO scheme applied to the problems with low density and low pressure is provided in §6. Finally, conclusions and future work are given in §7.

## 3 A single-stage single-step finite difference WENO method

The numerical method that is the subject of this work is based upon the Taylor discretization of the Picard integral formulation of Euler's equations, which is one of the many methods developed in [16]. Our focus is on the finite difference WENO method based on a Taylor discretization of the time-averaged fluxes because it easily lends itself to the positivity-preserving limiters that are presented in §4. In this section, we review the minimal details presented in [16] that are necessary to reproduce the present work. In addition, this section serves to set the notation that is used in upcoming sections.

In two dimensions, a hyperbolic conservation law is defined by a flux function with two components,

$$q_t + f(q)_{,x} + g(q)_{,y} = 0, \tag{7}$$

where $q(t,x,y) : \mathbb{R}^+ \times \mathbb{R}^2 \to \mathbb{R}^m$ is the vector of conserved variables, and $f,g : \mathbb{R}^m \to \mathbb{R}^m$ are the two components of the flux function. The Euler equations are an example of a set of equations from this class of problems.

Formal integration of (7) in time over $t \in [t^n, t^{n+1}]$ defines the 2D *Picard integral formulation* [16] as

$$q^{n+1} = q^n - \Delta t \left(F^n(x,y)\right)_{,x} - \Delta t \left(G^n(x,y)\right)_{,y}, \tag{8a}$$

where the *time-averaged flux* is defined as

$$F^n(x,y) := \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} f(q(t,x,y)) \, dt, \quad G^n(x,y) := \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} g(q(t,x,y)) \, dt. \tag{8b}$$

The basic idea of the Picard integral formulation of WENO (PIF-WENO) [16] is to first approximate the time-averaged fluxes (5) at each grid point using some temporal discretization, and then approximate spatial derivatives in (8a) by applying WENO reconstruction to the resulting time-averaged fluxes. In this work, we approximate Eqn. (8a) with the finite difference WENO method, and we use a third-order Taylor discretization for (8b). We remark that the positivity-preserving limiter proposed in §4 can be generally applied to any form of the Picard integral formulation, including Runge-Kutta time discretizations.

Given a domain $\Omega = [a_x, b_x] \times [a_y, b_y]$, a finite difference approximation seeks pointwise approximations $q_{i,j}^n \approx q(t^n, x_i, y_j)$ to hold at each

$$x_i = a_x + \left(i - \frac{1}{2}\right)\Delta x, \qquad \Delta x = \frac{b_x - a_x}{m_x}, \qquad i \in \{1, \dots, m_x\}, \tag{9a}$$

$$y_j = a_y + \left(j - \frac{1}{2}\right)\Delta y, \qquad \Delta y = \frac{b_y - a_y}{m_y}, \qquad j \in \{1, \dots, m_y\}, \tag{9b}$$

for discrete values of $t = t^n$. The 2D PIF-WENO scheme [16] solves Eqn. (7) with a conservative form

$$q_{i,j}^{n+1} = q_{i,j}^n - \lambda_x \left(\hat{F}_{i+\frac{1}{2},j}^n - \hat{F}_{i-\frac{1}{2},j}^n\right) - \lambda_y \left(\hat{G}_{i,j+\frac{1}{2}}^n - \hat{G}_{i,j-\frac{1}{2}}^n\right), \tag{10}$$

where $\lambda_x = \Delta t/\Delta x$, $\lambda_y = \Delta t/\Delta y$, and $\hat{F}_{i\pm\frac{1}{2},j}^n$ and $\hat{G}_{i,j\pm\frac{1}{2}}^n$ are high-order fluxes obtained by applying the classical WENO reconstruction to the time-averaged fluxes in place of a typical "frozen-in-time" approximation to the fluxes. This requires a total of two steps: construct a time-averaged flux, followed by performing a WENO reconstruction to the resulting modified fluxes.

We first define numerical time-averaged fluxes at each grid point $(x_i, y_j)$ through Taylor expansions. After taking temporal derivatives of $f$ and $g$, we integrate the resulting Taylor polynomials over $[t^n, t^{n+1}]$ to yield

$$F_T^n(x,y) := f(q(t^n,x,y)) + \frac{\Delta t}{2!}\frac{df}{dt}(q(t^n,x,y)) + \frac{\Delta t^2}{3!}\frac{d^2 f}{dt^2}(q(t^n,x,y)) \tag{11a}$$

$$G_T^n(x,y) := g(q(t^n,x,y)) + \frac{\Delta t}{2!}\frac{dg}{dt}(q(t^n,x,y)) + \frac{\Delta t^2}{3!}\frac{d^2 g}{dt^2}(q(t^n,x,y)). \tag{11b}$$

The temporal derivatives that appear can be found via the Cauchy-Kowalewski procedure. For example, the first two time derivatives of the first component of the flux function are given by

$$\frac{df}{dt} = -\frac{\partial f}{\partial q} \cdot (f_{,x} + g_{,y}), \tag{12a}$$

and

$$\frac{d^2 f}{dt^2} = \frac{\partial^2 f}{\partial q^2} \cdot \left(f_{,x} + g_{,y}, f_{,x} + g_{,y}\right) + \frac{\partial f}{\partial q} \cdot \left(-f_{,x} - g_{,y}\right)_{,t}. \tag{12b}$$

The last time derivative can be further simplified to

$$-\left(f_{,x} + g_{,y}\right)_{,t} = \frac{\partial^2 f}{\partial q^2} \cdot \left(q_{,x}, f_{,x} + g_{,y}\right) + \frac{\partial f}{\partial q} \cdot \left(f_{,xx} + g_{,xy}\right) + \tag{13}$$
$$\frac{\partial^2 g}{\partial q^2} \cdot \left(q_{,y}, f_{,x} + g_{,y}\right) + \frac{\partial g}{\partial q} \cdot \left(f_{,xy} + g_{,yy}\right).$$

Temporal derivatives of $g$ have a similar structure, and can be found in [16].

We approximate each $\partial_x, \partial_{xx}$ and $\partial_y, \partial_{yy}$ in (12a) and (12b) by applying the 5-point finite difference formulae

$$u_{i,x} := \frac{1}{12\Delta x} \left(u_{i-2} - 8u_{i-1} + 8u_{i+1} - u_{i+2}\right) = u_{,x}(x_i) + \mathscr{O}\left(\Delta x^4\right) \tag{14a}$$

$$u_{i,xx} := \frac{1}{12\Delta x^2} \left(-u_{i-2} + 16u_{i-1} - 30u_i + 16u_{i+1} - u_{i+2}\right) = u_{,xx}(x_i) + \mathscr{O}\left(\Delta x^4\right) \tag{14b}$$

in each direction. In order to retain a compact stencil, we compute the cross derivatives $\partial_{xy}$ with a second-order approximation

$$u_{ij,xy} := \frac{1}{4\Delta x \Delta y} \left(u_{i+1,j+1} - u_{i-1,j+1} - u_{i+1,j-1} + u_{i-1,j-1}\right), \tag{15}$$

which is sufficient to retain third-order accuracy in time. After defining these higher derivatives, we define numerical fluxes by $F_{i,j}^n := F_T^n(x_i, y_j)$ and $G_{i,j}^n := G_T^n(x_i, y_j)$. We then apply WENO reconstruction in a dimension by dimension fashion to each component of the flux to construct interface values $F_{i\pm\frac{1}{2},j}^n$ and $G_{i,j\pm\frac{1}{2}}^n$. The complete description of this process can be found in [16].

**Remark 1** *The Picard integral formulation sets up a discretization for the fluxes, and not the conserved variables.*

The significance of this remark is that further flux modifications can be incorporated into the Picard integral formulation. Previous finite difference WENO methods with Lax-Wendroff type time discretizations (e.g. [27, 28]) rely on Taylor expansions of the *conserved variables*, and not the fluxes; the Taylor discretization of the Picard integral formulation computes Taylor expansions of the fluxes, and not the conserved variables. In [27], conservation of mass comes from the fact that higher derivatives of the conserved variables are computed with a central stencil. In our scheme, we directly discretize the fluxes, and are automatically mass conservative because we insert the result into the WENO reconstruction procedure. Because ours is an operation on the fluxes, we have the opportunity to consider further flux modifications. In this work, we further modify the fluxes to obtain a provably positivity-preserving method for Euler equations, which we now describe.

## 4 The positivity-preserving method: the 1D case

We begin with the 1D formulation of the proposed positivity-preserving scheme. Recall that the update for the vector of conserved variables is given by Eqn. (4) for the 1D conservation law defined in (2). We consider a numerical flux $\hat{F}^n_{i-\frac{1}{2}}$ that is high order accurate in time (and space) that is constructed from the 1D Taylor discretization of the Picard integral formulation (PIF) of FD-WENO [16], and we consider a low-order flux $\hat{f}^n_{i-\frac{1}{2}}$ that is constructed from the Lax-Friedrichs scheme (that is provably positivity-preserving [39]). Both fluxes are constructed by looking at the solution $q^n$ at time level $t^n$.

We propose modifying the high-order flux by

$$\tilde{F}^n_{i-\frac{1}{2}} := \theta^n_{i-\frac{1}{2}}(\hat{F}^n_{i-\frac{1}{2}} - \hat{f}^n_{i-\frac{1}{2}}) + \hat{f}^n_{i-\frac{1}{2}}, \tag{16}$$

where a simple optimization problem is solved for the *limiting parameter* $\theta^n_{i-\frac{1}{2}} \in [0,1]$ at each time step.

We observe that if $\theta^n_{i-\frac{1}{2}} = 0$, the scheme reduces to the first-order Lax-Friedrichs scheme, which is positivity preserving, and therefore it is always possible to find a value that retains positive density and pressure. If $\theta^n_{i-\frac{1}{2}} = 1.0$, the scheme reduces to the high-order scheme, but does not guarantee positivity of the numerical solution. In order to retain high-order accuracy, we would like to choose $\theta^n_{i-\frac{1}{2}}$ as close to 1.0 as possible without violating positivity of the density and pressure.

The positivity-preserving Algorithm we outline in this section follows a two step procedure: i) guarantee positivity of the density, and then ii) guarantee positivity of the pressure. The details of this procedure are spelled out in the following subsections.

### 4.1 Step 1: Maintain positivity of the density

This discussion focuses on the first component of the modified flux

$$\tilde{f}^{n,\rho}_{i+\frac{1}{2}} := \theta^n_{i+\frac{1}{2}}(\hat{F}^{n,\rho}_{i+\frac{1}{2}} - \hat{f}^{n,\rho}_{i+\frac{1}{2}}) + \hat{f}^{n,\rho}_{i+\frac{1}{2}}, \tag{17}$$

where $\hat{f}^{n,\rho}$ is the first component of the low-order flux $\hat{f}^n$, and $\hat{F}^{n,\rho}$ is the first component of the high-order flux $\hat{F}^n$. In this step, we must assume that the density $\rho^n_i > 0$ is positive at the current time. We further define the low-order update for the density as

$$\hat{\rho}^{n+1}_i := \rho^n_i - \lambda\left(\hat{f}^{n,\rho}_{i+\frac{1}{2}} - \hat{f}^{n,\rho}_{i-\frac{1}{2}}\right),$$

and define a numerical lower bound of the high-order updated density $\rho^{n+1}$ as $\varepsilon^{n+1}_\rho := \min(\min_i(\hat{\rho}^{n+1}_i), \varepsilon_0)$. The use of $\varepsilon_0 > 0$ guarantees finite wave speeds, because the sound speed $c := \sqrt{\gamma p/\rho}$ goes to infinity as $\rho \to 0$. In our simulations, we take $\varepsilon_0 = 10^{-13}$, which is consistent with recent high-order positivity-preserving work [44]. Thanks to the positivity of the low-order flux [39], we observe that $\varepsilon^{n+1}_\rho > 0$.

After the low- and high-order fluxes have been computed, the update for the density at a single grid point $x_i$ only depends on two values of $\theta^n_{i\pm\frac{1}{2}}$ through

$$\rho^{n+1}_i\left(\theta^n_{i-\frac{1}{2}}, \theta^n_{i+\frac{1}{2}}\right) = \rho^n_i - \lambda\left(\tilde{f}^{n,\rho}_{i+\frac{1}{2}} - \tilde{f}^{n,\rho}_{i-\frac{1}{2}}\right), \quad \lambda = \frac{\Delta t}{\Delta x}, \quad i \in \{1, 2, \ldots, m_x\}.$$

This, and each of the conserved variables are *linear functions* with respect to the variable $\left( \theta^n_{i-\frac{1}{2}}, \theta^n_{i+\frac{1}{2}} \right) \in [0,1]^2$. To preserve the positivity of the density $\rho^{n+1}$, we want to guarantee $\rho^{n+1}_i \geq \varepsilon^{n+1}_\rho$. Therefore, we seek bounds $\Lambda^\rho_{\pm\frac{1}{2},I_i}$ such that whenever

$$\left( \theta^n_{i-\frac{1}{2}}, \theta^n_{i+\frac{1}{2}} \right) \in \left[ 0, \Lambda^\rho_{-\frac{1}{2},I_i} \right] \times \left[ 0, \Lambda^\rho_{+\frac{1}{2},I_i} \right] \subseteq [0,1]^2,$$

we have

$$\rho^{n+1}_i \left( \theta^n_{i-\frac{1}{2}}, \theta^n_{i+\frac{1}{2}} \right) = \rho^n_i - \lambda \left( \tilde{f}^{n,\rho}_{i+\frac{1}{2}} - \tilde{f}^{n,\rho}_{i-\frac{1}{2}} \right) \geq \varepsilon^{n+1}_\rho. \tag{18}$$

The purpose of defining such a set is that in Step 2 of §4.2, we will further limit the fluxes to maintain positivity of the pressure.

We insert the definition of $\hat{\rho}^{n+1}_i$ into Eqn. (18) to see

$$\hat{\rho}^{n+1}_i - \lambda \left[ \theta^n_{i+\frac{1}{2}} \left( \hat{F}^{n,\rho}_{i+\frac{1}{2}} - \hat{f}^{n,\rho}_{i+\frac{1}{2}} \right) - \theta^n_{i-\frac{1}{2}} \left( \hat{F}^{n,\rho}_{i-\frac{1}{2}} - \hat{f}^{n,\rho}_{i-\frac{1}{2}} \right) \right] \geq \varepsilon^{n+1}_\rho, \tag{19}$$

which is equivalent to

$$\theta^n_{i-\frac{1}{2}} \Delta f_{i-\frac{1}{2}} - \theta^n_{i+\frac{1}{2}} \Delta f_{i+\frac{1}{2}} \geq \varepsilon^{n+1}_\rho - \hat{\rho}^{n+1}_i, \tag{20}$$

where $\Delta f_{i-\frac{1}{2}} := \lambda(\hat{F}^{n,\rho}_{i-\frac{1}{2}} - \hat{f}^{n,\rho}_{i-\frac{1}{2}})$ is a measure of the deviation of the high- and low-order fluxes. Note that the right hand side satisfies $\varepsilon^{n+1}_\rho - \hat{\rho}^{n+1}_i \leq 0$, and therefore there is at least one solution that can be found for Eqn. (20) (namely $\theta^n_{i-\frac{1}{2}} = \theta^n_{i+\frac{1}{2}} = 0$).

We determine bounds on $\Lambda^\rho_{\pm\frac{1}{2},I_i}$ through a case-by-case discussion based on the signs of $\Delta f_{i-\frac{1}{2}}$ and $\Delta f_{i+\frac{1}{2}}$. This analysis has already been performed for single [55] and multidimensional [56] scalar problems. There are a total of four cases of Eqn. (20) to consider:

- Case 1. If $\Delta f_{i-\frac{1}{2}} \geq 0$ and $\Delta f_{i+\frac{1}{2}} \leq 0$, then we set

$$\left( \Lambda^\rho_{-\frac{1}{2},I_i}, \Lambda^\rho_{+\frac{1}{2},I_i} \right) := (1,1).$$

- Case 2. If $\Delta f_{i-\frac{1}{2}} \geq 0$ and $\Delta f_{i+\frac{1}{2}} > 0$, then we define

$$\left( \Lambda^\rho_{-\frac{1}{2},I_i}, \Lambda^\rho_{+\frac{1}{2},I_i} \right) := \left( 1, \min \left( 1, \frac{\varepsilon^{n+1}_\rho - \hat{\rho}^{n+1}_i}{-\Delta f_{i+\frac{1}{2}}} \right) \right).$$

- Case 3. If $\Delta f_{i-\frac{1}{2}} < 0$ and $\Delta f_{i+\frac{1}{2}} \leq 0$, then we set

$$\left( \Lambda^\rho_{-\frac{1}{2},I_i}, \Lambda^\rho_{+\frac{1}{2},I_i} \right) := \left( \min \left( 1, \frac{\varepsilon^{n+1}_\rho - \hat{\rho}^{n+1}_i}{\Delta f_{i-\frac{1}{2}}} \right), 1 \right).$$

- Case 4. If $\Delta f_{i-\frac{1}{2}} < 0$ and $\Delta f_{i+\frac{1}{2}} > 0$, there are two sub-cases to consider.

– <u>Case 4a.</u> If the inequality (20) is satisfied with $(\theta_{i-\frac{1}{2}}^n, \theta_{i+\frac{1}{2}}^n) = (1,1)$ then we set

$$\left(\Lambda_{-\frac{1}{2}, I_i}^\rho, \Lambda_{+\frac{1}{2}, I_i}^\rho\right) := (1,1).$$

– <u>Case 4b.</u> Otherwise, we choose

$$\left(\Lambda_{-\frac{1}{2}, I_i}^\rho, \Lambda_{+\frac{1}{2}, I_i}^\rho\right) := \left(\frac{\varepsilon_\rho^{n+1} - \hat{\rho}_i^{n+1}}{\Delta f_{i-\frac{1}{2}} - \Delta f_{i+\frac{1}{2}}}, \frac{\varepsilon_\rho^{n+1} - \hat{\rho}_i^{n+1}}{\Delta f_{i-\frac{1}{2}} - \Delta f_{i+\frac{1}{2}}}\right).$$

After considering each of the above cases at each grid element $x_i$, we define the following set

$$S_i^\rho := \left[0, \Lambda_{-\frac{1}{2}, I_i}^\rho\right] \times \left[0, \Lambda_{+\frac{1}{2}, I_i}^\rho\right]. \tag{21}$$

The obtained set has the property that $\rho_i^{n+1}(\theta_{i-\frac{1}{2}}^n, \theta_{i+\frac{1}{2}}^n) \geq \varepsilon_\rho^{n+1}$ for any $(\theta_{i-\frac{1}{2}}^n, \theta_{i+\frac{1}{2}}^n) \in S_i^\rho$.

4.2 Step 2: Maintain positivity of the pressure

The second step focuses on the pressure $p_i^{n+1}$. We begin with the following Lemma, which has already been used in the past [12, 44].

**Lemma 1** *The pressure function satisfies*

$$p\left(q_i^n\left(\alpha\overrightarrow{\theta}^1 + (1-\alpha)\overrightarrow{\theta}^2\right)\right) \geq \alpha p\left(q_i^n\left(\overrightarrow{\theta}^1\right)\right) + (1-\alpha)p\left(q_i^n\left(\overrightarrow{\theta}^2\right)\right)$$

*for any $\alpha \in [0,1]$ and $\overrightarrow{\theta}^1, \overrightarrow{\theta}^2 \in S_i^\rho$.*

*Proof* Provided $\rho > 0$, the pressure function

$$p(q) := (\gamma - 1)\left(\mathscr{E} - \frac{\|\rho\mathbf{u}\|^2}{2\rho}\right)$$

is concave with respect to the conserved variables $q = (\rho, \rho\mathbf{u}, \mathscr{E})$ [12, 40, 44]. By definition of the limiting parameter, each of the conserved variables are *linear functions* of $\overrightarrow{\theta}$, which means

$$q_i^n\left(\alpha\overrightarrow{\theta}^1 + (1-\alpha)\overrightarrow{\theta}^2\right) = \alpha q_i^n\left(\overrightarrow{\theta}^1\right) + (1-\alpha)q_i^n\left(\overrightarrow{\theta}^2\right).$$

Together, and as a result of the construction in Step 1, these imply

$$p\left(q_i^n\left(\alpha\overrightarrow{\theta}^1 + (1-\alpha)\overrightarrow{\theta}^2\right)\right) = p\left(\alpha q_i^n\left(\overrightarrow{\theta}^1\right) + (1-\alpha)q_i^n\left(\overrightarrow{\theta}^2\right)\right),$$
$$\geq \alpha p\left(q_i^n\left(\overrightarrow{\theta}^1\right)\right) + (1-\alpha)p\left(q_i^n\left(\overrightarrow{\theta}^2\right)\right).$$

$\square$

We define $p_i\left(\overrightarrow{\theta}\right) := p\left(q_i^n\left(\overrightarrow{\theta}\right)\right)$ for any $\overrightarrow{\theta} \in [0,1]^2$ in order to simplify the notation for the ensuing discussion.

If we use $\hat{p}^{n+1}$ to denote the low-order pressure solved by the flux $\hat{f}^n$, we can similarly define a numerical lower bound for the pressure as $\varepsilon_p^{n+1} := \min(\min_i(\hat{p}_i), \varepsilon_0)$. Next, we consider the following subset

$$S_i^p := \left\{ (\theta_{i-\frac{1}{2}}^n, \theta_{i+\frac{1}{2}}^n) \in S_i^\rho : p_i\left(\theta_{i-\frac{1}{2}}^n, \theta_{i+\frac{1}{2}}^n\right) \geq \varepsilon_p^{n+1} \right\} \subseteq S_i^\rho, \tag{22}$$

and we observe that $S_i^p$ is convex, thanks to the result of Lemma 1. We do not attempt to find the entire boundary of $S_i^p$ because that would be computationally intractable. Instead, we define a single rectangle $R_i^{\rho,p}$ inside of $S_i^p$ that define bounds on the limiting parameters.

To do this, we consider finitely many points on the boundary of $S_i^p$. To begin, consider the four vertices of $S_i^\rho$ denoted by $A^{k_1,k_2} := (k_1\Lambda_{-\frac{1}{2},i}^\rho, k_2\Lambda_{+\frac{1}{2},i}^\rho)$, with $k_1, k_2$ being 0 or 1. For each $(k_1, k_2)$, we define $B^{k_1,k_2}$ based on two cases:

- Case 1. If $p_i(A^{k_1,k_2}) \geq \varepsilon_p^{n+1}$, we put $B^{k_1,k_2} := A^{k_1,k_2}$. The origin always falls into this case.
- Case 2. Otherwise, we solve the quadratic equation $p_i(rA^{k_1,k_2}) = \varepsilon_p^{n+1}$ for the unknown variable $r \in [0,1]$, and define $B^{k_1,k_2} := rA^{k_1,k_2}$.

After checking each vertex of $S_i^\rho$, we define

$$R_i^{\rho,p} := \left[0, \Lambda_{-\frac{1}{2},I_i}\right] \times \left[0, \Lambda_{+\frac{1}{2},I_i}\right] \subseteq S_i^p \subseteq S_i^\rho, \tag{23}$$

where

$$\Lambda_{-\frac{1}{2},I_i} := \min_{k_2=0,1}\left(B^{1,k_2}\right), \quad \Lambda_{+\frac{1}{2},I_i} := \min_{k_1=0,1}\left(B^{k_1,1}\right). \tag{24}$$

After performing this two step process at each grid cell $x_i$, the end result of this construction is the following theorem.

**Theorem 1** *The numerical flux in Eqn. (16) preserves positivity of the solution for any*

$$\theta_{i-\frac{1}{2}}^n \in \left[0, \min\left(\Lambda_{+\frac{1}{2},I_{i-1}}, \Lambda_{-\frac{1}{2},I_i}\right)\right]. \tag{25}$$

Although we could in principle choose any value in the interval defined in Eqn. (25) (e.g. $\theta_{i-\frac{1}{2}}^n = 0$), in order to retain high-order accuracy, we choose the largest possible value that we can prove retains positivity. That is, we define

$$\theta_{i-\frac{1}{2}}^n := \min\left(\Lambda_{+\frac{1}{2},I_{i-1}}, \Lambda_{-\frac{1}{2},I_i}\right) \tag{26}$$

at each cell interface.

This finishes the discussion for the 1D scheme. We reiterate that this entire process relies on flux modifications, which the Picard integral formulation was designed to accept, and is pointed out in Rmk. 1.

**Remark 2** *The positivity of the solution is guaranteed for the _entire simulation_.*

One consequence of being a single-stage, single-step method is that we do not have stages where the density or pressure can become negative, whereas multistage Runge-Kutta methods typically introduce either additional computational cost or artificial sound speeds in order to retain positivity. For example, in [14, 15, 44] the limiter is applied after each stage in the Runge-Kutta method. This introduces additional computational complexity (i.e. there are multiple applications of the limiter per time step) as well as further constraints on the time step selection because the limiter relies on positivity of the forward Euler method. The modifications made in [12, 13] are part of an effort to decrease the computational complexity by only applying the limiter once per time step. However, this happens at the expense of potentially introducing negative pressure and density. In order to compensate for this, in [12, 13] the authors indicate they artificially define the sound speed as $c = \sqrt{\gamma |p|/|\rho|}$ for each stage in the Runge-Kutta method. This is necessary to implement the characteristic decomposition required for the high-order WENO reconstruction, and although it does not affect the refinement study in [12, 13], this treatment may lead to a potential numerical instability for some extreme cases. A similar issue can be found for the ideal MHD equations [58].

## 5 The positivity-preserving method: the 2D case

In this section, we apply the positivity-preserving limiter to the two-dimensional case. Extensions to a general multi-D case follow from what is provided here.

Recall that our single-stage, single-step update is given by Eqn. (10). Similar to the 1D case, we use $\hat{f}^n_{i-\frac{1}{2},j}$ and $\hat{g}^n_{i,j-\frac{1}{2}}$ to denote the low-order positivity-preserving fluxes, and our numerical method uses modified fluxes through

$$\tilde{F}^n_{i-\frac{1}{2},j} := \theta^n_{i-\frac{1}{2},j}(\hat{F}^n_{i-\frac{1}{2},j} - \hat{f}^n_{i-\frac{1}{2},j}) + \hat{f}^n_{i-\frac{1}{2},j}, \tag{27a}$$

$$\tilde{G}^n_{i,j-\frac{1}{2}} := \theta^n_{i,j-\frac{1}{2}}(\hat{G}^n_{i,j-\frac{1}{2}} - \hat{g}^n_{i,j-\frac{1}{2}}) + \hat{g}^n_{i,j-\frac{1}{2}}. \tag{27b}$$

Identical to the single-dimensional case, the positivity-preserving limiting procedure consists of two steps. If we still use $\hat{\rho}^{n+1}$ and $\hat{p}^{n+1}$ to denote the low-order density and pressure solved by the flux $\hat{f}^n$ and $\hat{g}^n$, we can similarly define the 2D numerical lower bounds for density and pressure as $\varepsilon^{n+1}_\rho := \min\left(\min_{i,j}\left(\hat{\rho}^{n+1}_{i,j}\right), \varepsilon_0\right)$ and $\varepsilon^{n+1}_p := \min\left(\min_{i,j}\left(\hat{p}^{n+1}_{i,j}\right), \varepsilon_0\right)$.

### 5.1 Step 1: Maintain positivity of the density

Our fist step is to find local bounds $\Lambda^\rho_{L,I_{i,j}}$, $\Lambda^\rho_{R,I_{i,j}}$, $\Lambda^\rho_{U,I_{i,j}}$ and $\Lambda^\rho_{D,I_{i,j}}$, such that for any $(\theta^n_{i-\frac{1}{2},j}, \theta^n_{i+\frac{1}{2},j}, \theta^n_{i,j-\frac{1}{2}}, \theta^n_{i,j+\frac{1}{2}}) \in S^\rho_{i,j}$, we have

$$\rho^{n+1}_{i,j} = \rho^n_{i,j} - \lambda_x\left(\tilde{f}^{n,\rho}_{i+\frac{1}{2},j} - \tilde{f}^{n,\rho}_{i-\frac{1}{2},j}\right) - \lambda_y\left(\tilde{g}^{n,\rho}_{i,j+\frac{1}{2}} - \tilde{g}^{n,\rho}_{i,j-\frac{1}{2}}\right) \geq \varepsilon^{n+1}_\rho, \tag{28}$$

where

$$S^\rho_{i,j} := \left[0, \Lambda^\rho_{L,I_{i,j}}\right] \times \left[0, \Lambda^\rho_{R,I_{i,j}}\right] \times \left[0, \Lambda^\rho_{D,I_{i,j}}\right] \times \left[0, \Lambda^\rho_{U,I_{i,j}}\right]. \tag{29}$$

Again, we have used the notation $g^\rho$ to refer to the first component of the flux function, $g$. We define the low-order update as

$$\hat{\rho}_{i,j}^{n+1} := \rho_{i,j}^n - \lambda_x \left( \hat{f}_{i+\frac{1}{2},j}^{n,\rho} - \hat{f}_{i-\frac{1}{2},j}^{n,\rho} \right) - \lambda_y \left( \hat{g}_{i,j+\frac{1}{2}}^{n,\rho} - \hat{g}_{i,j-\frac{1}{2}}^{n,\rho} \right),$$

and observe that it satisfies $\hat{\rho}_{i,j}^{n+1} \geq \varepsilon_\rho^{n+1} > 0$ for all $i, j$, provided the density is positive at time $t^n$. Similar to Eqn. (20), we rewrite Eqn. (28) as

$$\theta_{i-\frac{1}{2},j}^n \Delta f_{i-\frac{1}{2},j} - \theta_{i+\frac{1}{2},j}^n \Delta f_{i+\frac{1}{2},j} + \theta_{i,j-\frac{1}{2}}^n \Delta g_{i,j-\frac{1}{2}} - \theta_{i,j+\frac{1}{2}}^n \Delta g_{i,j+\frac{1}{2}} \geq \varepsilon_\rho^{n+1} - \hat{\rho}_{i,j}^{n+1}, \quad (30)$$

where we have defined the deviation between the high- and low-order fluxes as

$$\begin{cases} \Delta f_{i-\frac{1}{2},j} := \lambda_x (\hat{F}_{i-\frac{1}{2},j}^{n,\rho} - \hat{f}_{i-\frac{1}{2},j}^{n,\rho}), \\ \Delta f_{i+\frac{1}{2},j} := \lambda_x (\hat{F}_{i+\frac{1}{2},j}^{n,\rho} - \hat{f}_{i+\frac{1}{2},j}^{n,\rho}), \\ \Delta g_{i,j-\frac{1}{2}} := \lambda_y (\hat{G}_{i,j-\frac{1}{2}}^{n,\rho} - \hat{g}_{i,j-\frac{1}{2}}^{n,\rho}), \\ \Delta g_{i,j+\frac{1}{2}} := \lambda_y (\hat{G}_{i,j-\frac{1}{2}}^{n,\rho} - \hat{g}_{i,j-\frac{1}{2}}^{n,\rho}). \end{cases} \quad (31)$$

Similar to the 1D case, we solve (30) based on the signs of $\Delta f_{i\pm\frac{1}{2},j}$ and $\Delta g_{i,j\pm\frac{1}{2}}$ at each node $(x_i, y_j)$. The basic idea requires a total of two steps:

1. Identify the negative values of each of the four numbers

$$\left\{ \Delta f_{i-\frac{1}{2},j}, -\Delta f_{i+\frac{1}{2},j}, \Delta g_{i,j-\frac{1}{2}}, -\Delta g_{i,j+\frac{1}{2}} \right\}. \quad (32)$$

2. Corresponding to the collective negative values, we define upper bounds on the limiting parameters by solving Eqn. (30) for each value of $\theta$ after neglecting any positive values found. For example, if $\Delta f_{i-\frac{1}{2},j}, -\Delta f_{i+\frac{1}{2},j} < 0$ and $\Delta g_{i,j-\frac{1}{2}}, -\Delta g_{i,j+\frac{1}{2}} \geq 0$, then we define

$$\begin{cases} \Lambda_{L,I_{i,j}}^\rho := \Lambda_{R,I_{i,j}}^\rho := \min \left( \frac{\varepsilon_\rho^{n+1} - \hat{\rho}_{i,j}^{n+1}}{\Delta f_{i-\frac{1}{2},j} - \Delta f_{i+\frac{1}{2},j}}, 1 \right), \\ \Lambda_{D,I_{i,j}}^\rho := \Lambda_{U,I_{i,j}}^\rho := 1. \end{cases} \quad (33)$$

Likewise, if $-\Delta f_{i+\frac{1}{2},j}, \Delta g_{i,j-\frac{1}{2}} < 0$ and $\Delta f_{i-\frac{1}{2},j}, -\Delta g_{i,j+\frac{1}{2}} \geq 0$, then we define

$$\begin{cases} \Lambda_{R,I_{i,j}}^\rho := \Lambda_{D,I_{i,j}}^\rho := \min \left( \frac{\varepsilon_\rho^{n+1} - \hat{\rho}_{i,j}^{n+1}}{-\Delta f_{i+\frac{1}{2},j} + \Delta g_{i,j-\frac{1}{2}}}, 1 \right), \\ \Lambda_{L,I_{i,j}}^\rho := \Lambda_{U,I_{i,j}}^\rho := 1. \end{cases} \quad (34)$$

There are a total of 16 cases. Each follow similarly, and are omitted for brevity.

5.2 Step 2: Maintain positivity of the pressure

Using the same construction from §4.2, we identify a rectangle $R_{i,j}^{\rho,p} \subseteq S_{i,j}^{\rho}$ where the pressure satisfies $p_{i,j}(\theta_{i-\frac{1}{2},j}^n, \theta_{i+\frac{1}{2},j}^n, \theta_{i,j-\frac{1}{2}}^n, \theta_{i,j+\frac{1}{2}}^n) \geq \varepsilon_p^{n+1}$. Again, we consider the vertices of the region that were computed in the first step. In 2D, we define them as

$$A^{k_1,k_2,k_3,k_4} := (k_1\Lambda_{L,I_{ij}}^{\rho}, k_2\Lambda_{R,I_{ij}}^{\rho}, k_3\Lambda_{D,I_{ij}}^{\rho}, k_4\Lambda_{U,I_{ij}}^{\rho}), \quad k_1,k_2,k_3,k_4 \in \{0,1\}.$$

We rescale each vertex in an identical manner to the 1D case presented in subsection 4.2. There are two cases:

–  Case 1. If $p_{i,j}(A^{k_1,k_2,k_3,k_4}) \geq \varepsilon_p^{n+1}$, we define the vertex $B^{k_1,k_2,k_3,k_4} := A^{k_1,k_2,k_3,k_4}$.
–  Case 2. We solve the quadratic equation $p_{i,j}(rA^{k_1,k_2,k_3,k_4}) = \varepsilon_p^{n+1}$ for the unknown $r \in [0,1]$ and put $B^{k_1,k_2,k_3,k_4} := rA^{k_1,k_2,k_3,k_4}$.

In the final step, we identify a rectangular box inside $S_{i,j}^p$ through

$$R_{i,j}^{\rho,p} := [0,\Lambda_{L,I_{i,j}}] \times [0,\Lambda_{R,I_{i,j}}] \times [0,\Lambda_{D,I_{i,j}}] \times [0,\Lambda_{U,I_{i,j}}], \tag{35}$$

where

$$
\begin{aligned}
\Lambda_{L,I_{i,j}} &:= \min_{k_{2,3,4} \in \{0,1\}} \left( B^{1,k_2,k_3,k_4} \right), \quad \Lambda_{R,I_{i,j}} := \min_{k_{1,3,4} \in \{0,1\}} \left( B^{k_1,1,k_3,k_4} \right), \\
\Lambda_{D,I_{i,j}} &:= \min_{k_{1,2,4} \in \{0,1\}} \left( B^{k_1,k_2,1,k_4} \right), \quad \Lambda_{U,I_{i,j}} := \min_{k_{1,2,3} \in \{0,1\}} \left( B^{k_1,k_2,k_3,1} \right).
\end{aligned}
\tag{36}
$$

After repeating this procedure for each node $(x_i, y_j)$, we finish by defining the scaling parameter as

$$\theta_{i-\frac{1}{2},j}^n := \min(\Lambda_{R,I_{i-1,j}}, \Lambda_{L,I_{i,j}}), \quad \theta_{i,j-\frac{1}{2}}^n := \min(\Lambda_{U,I_{i,j-1}}, \Lambda_{D,I_{i,j}}), \tag{37}$$

and insert the result into Eqn (27) to define our modified fluxes. This finishes the discussion for the 2D scheme, and a similar positivity-preserving theorem follows as in Thm. 25.


# 6 Numerical results

In this section, we perform numerical simulations with our proposed positivity-preserving method on 1D and 2D compressible Euler equations.

*6.0.1 Implementation details*

The parameters we use for our WENO reconstructions include a power parameter $p = 2$, and a regularization parameter $\varepsilon = 10^{-6}$, and a gas constant of $\gamma = 1.4$. In addition, we follow common practice and use a global (as opposed to a local) value for $\alpha$ in the Lax-Friedrichs flux splitting for all of our simulations. This introduces additional numerical dissipation that helps to prevent unphysical oscillations in this high-order scheme. Contrary to what typically happens with first-order finite volume schemes, the additional numerical dissipation introduced by this choice does not introduce an exorbitant amount of artificial diffusion. In every simulation save one, the CFL number is 0.35. All of our numerical results can be found in the open source software package FINESS [59].

**Table 1** Accuracy test of the 2D smooth vortex. We show the $L_1$-errors and $L_\infty$-errors at time $t = 0.01$ of the density. The solutions converge at fifth-order accuracy.

| Mesh | $L_1$-Error | Order | $L_\infty$-Error | Order |
|:---:|:---:|:---:|:---:|:---:|
| $80 \times 80$ | 2.970e-06 | - | 2.494e-04 | - |
| $160 \times 160$ | 1.627e-07 | 4.190 | 2.442e-05 | 3.353 |
| $320 \times 320$ | 7.384e-09 | 4.462 | 1.390e-06 | 4.135 |
| $640 \times 640$ | 2.428e-10 | 4.927 | 4.718e-08 | 4.881 |

## 6.1 Accuracy test

To test the accuracy of our method, we use the smooth vortex problem with low density and low pressure [12, 15]. Initially, we have a mean flow

$$(\rho, u^1, u^2, u^3, p) = (1, 1, 1, 0, 1), \tag{38}$$

with perturbations on the velocities $u^1$, $u^2$ and the temperature $T = p/\rho$, given by

$$(\delta u^1, \delta u^2) = \frac{\varepsilon}{2\pi} e^{0.5(1-r^2)}(-y, x), \quad \delta T = -\frac{(\gamma-1)\varepsilon^2}{8\gamma\pi^2} e^{1-r^2}, \quad r^2 := x^2 + y^2.$$

The initial pressure and density are determined by keeping the entropy $S = p/\rho^\gamma$ constant. The domain is $(x, y) \in [-5, 5] \times [-5, 5]$ with periodic boundary condition on all sides. The vortex strength $\varepsilon$ is set as 10.0828 such that the lowest density and lowest pressure in the center of the vortex are $7.8 \times 10^{-15}$ and $1.7 \times 10^{-20}$ respectively.

A convergence study is presented in Table 1. The $L_1$-errors and $L_\infty$-errors of the density are computed at a final time of $t = 0.01$. We observe the fifth-order accuracy of the proposed scheme, which is comparable with those demonstrated in [12, 15]. In [15], the authors took $\Delta t = \Delta x^{\frac{5}{3}}$ in order to make the spatial error dominate the numerical error. We find this treatment is not necessary to observe high-order spatial accuracy because of the short the final time. In our table, we only present the results with a constant CFL number of 0.35 that has been chosen for this, and all other simulations save one. Without the addition of the positivity limiter, we observe negative density and negative pressure with the Taylor formulation of the PIF-WENO scheme that appears in the center of vortex.

## 6.2 1D Sedov blast wave problem

The first 1D problem we considered is the 1D Sedov blast problem originally from the book by Sedov [60]. The problem describes an intense explosion in a gas where the disturbed air is separated from the undisturbed air by a shock wave. Initially, we deposit a quantity of energy $\mathscr{E} = 3200000$ into the center cell of the domain with the length of $\Delta x$, and the energy in every other cell is set to $10^{-12}$. The other quantities are initialized with a constant values of $\rho = 1$ and $u^1 = 0$. The numerical results are displayed in Fig. 1, where we see the shock wave is captured with the proposed limiter used. In Fig. 1, we use the exact solution given in Sedov's book [60] as the reference solution. Our results are in agreement with other recent work [12, 15, 44].

### 6.3 1D double rarefaction problem

The second 1D problem we considered is the double rarefaction problem. It is a Riemann problem with an initial condition of $\rho_L = \rho_R = 7$, $u_L^1 = -1$, $u_R^1 = 1$ and $p_L = p_R = 0.2$. The exact solution consists of two rarefaction waves traveling in opposite directions, which results in the creation of a vacuum in the center of the domain. Only with the proposed limiter are we are able to solve this low density and low pressure problem with the high-order finite difference WENO method. For this problem only, we find it is necessary to reduce the CFL number from 0.35 to 0.15 in order to avoid introducing spurious oscillations near the top of the rarefactions. The numerical results are presented in Fig. 2, where we use the same resolution of $\Delta x = 1/200$ as those in reference [12, 15, 44]. Our results are in agreement with the exact solution. Here, the exact solution is a highly resolved solution with $\Delta x = 1/1000$. Other Riemann problems have been investigated, and our method gives similar results as those found elsewhere in the literature (e.g. [38, 40]).

### 6.4 2D Sedov blast wave problem

We also consider a 2D version of the Sedov blast wave problem. In 2D case, the problem has an exact self-similar solution and we expect the numerical result has a similar structure. In the simulation, we only compute one quadrant of the whole domain, where we choose the computation domain to be $(x, y) \in [0, 1.1] \times [0, 1.1]$. Similar to the 1D case, we deposit a quantity of energy $\mathscr{E} = 0.244816$ into the lower left corner cell, and set the energy in every other cell to be $\mathscr{E} = 10^{-12}$. The other initial values are identical to the 1D case. We apply solid wall boundary conditions along the bottom ($x = 0$) and left ($y = 0$) boundaries so that the setup is equivalent to computing the whole domain $[-1.1, 1.1] \times [-1.1, 1.1]$ with $\mathscr{E} = 0.979264$. The density is presented in Fig. 3, from which we can see the result has a nice self-similar structure. Additionally, we observe that the density cut at $y = 0$ agrees with the exact solution.

### 6.5 2D Shock diffraction problem

The second 2D problem we consider is the shock diffraction problem. The computational domain is $[0, 1] \times [6, 11] \bigcup [1, 13] \times [0, 11]$. There is a shock wave of Mach number 5.09 initially located at $\{x = 0.5, 6 \leq y \leq 11\}$. As time evolves, the wave moves into undisturbed air with a density of $\rho = 1.4$ and pressure of $p = 1$. We use an inflow boundary condition at $\{x = 0, 6 \leq y \leq 11\}$, and an outflow boundary condition at $\{x = 13, 0 \leq y \leq 11\}$, $\{1 \leq x \leq 13, y = 0\}$ and $\{0 \leq x \leq 13, y = 11\}$. For the other parts of the boundary where $\{0 \leq x \leq 1, y = 6\}$ and $\{x = 1, 0 \leq y \leq 6\}$, solid wall boundary conditions are applied. As the shock passes the corner, negative density and negative pressure is observed without the addition of the positivity limiter to the Taylor PIF-WENO scheme. This issue is resolved with the proposed modifications to the older scheme. In Fig. 4, we present results for the density and pressure at time $t = 2.3$.

## 7 Conclusions and future work

In this paper we propose a high-order, single-stage, single-step, positivity-preserving method for the compressible Euler equations. The base scheme is the Taylor discretization of the
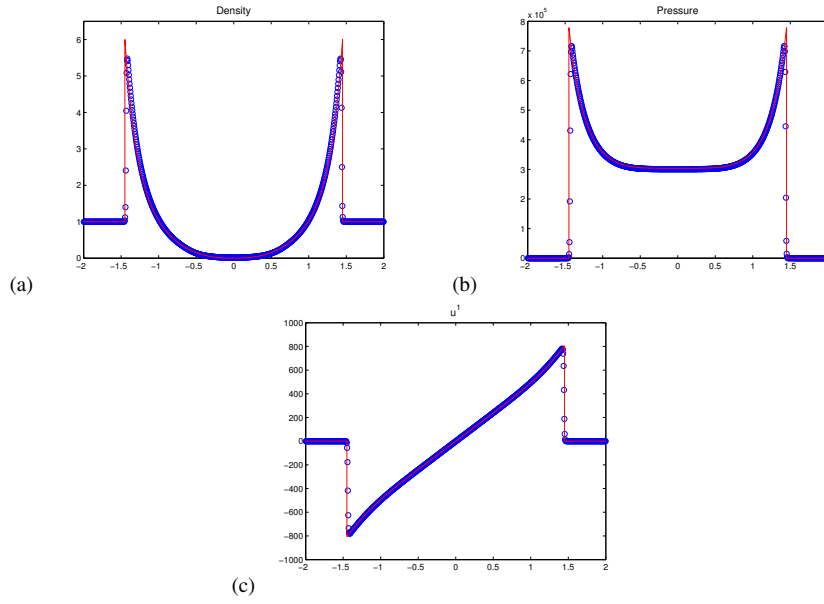
**Fig. 1** 1D Sedov blast wave problem. These panels show plots at time $t = 0.001$ of (a) the density, (b) the pressure and (c) the velocity $u^1$. The solid lines are the exact solutions. The solution was obtained on a mesh with $\Delta x = 1/200$ and a CFL of 0.35.
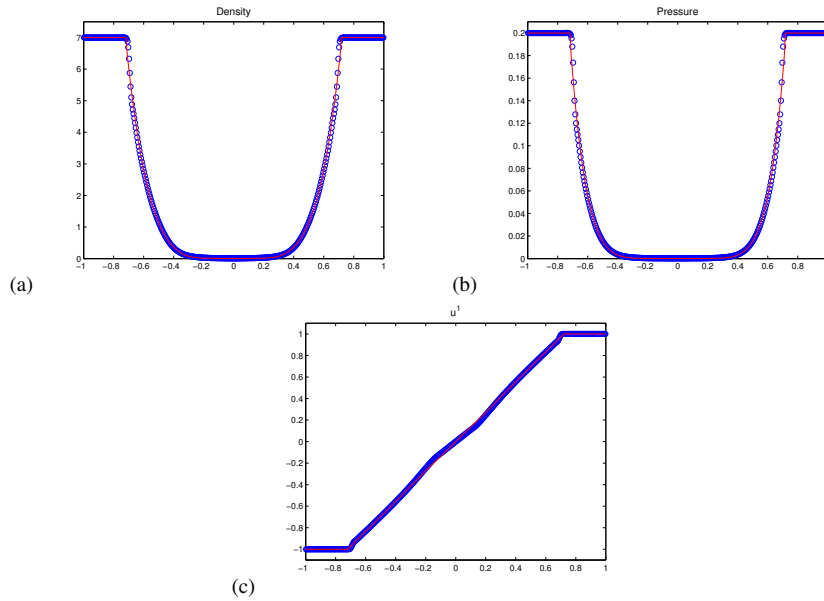


**Fig. 2** 1D double rarefaction problem. These panels show plots at time $t = 0.6$ of (a) the density, (b) the pressure and (c) $u^1$. The solid lines are the exact solutions. The solution is obtained on a mesh with $\Delta x = 1/200$ and a smaller CFL of 0.15 that help to reduce unphysical oscillations.
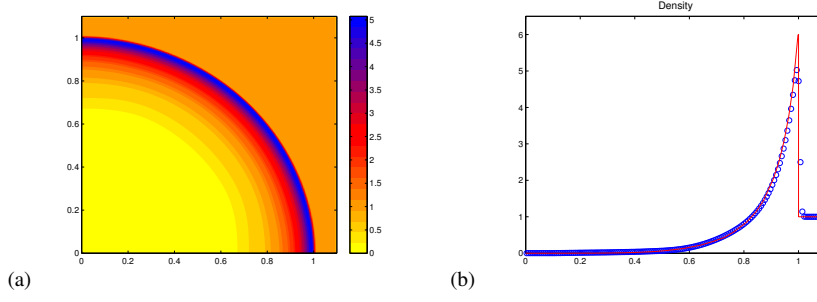
(a)                                                              (b)

**Fig. 3** 2D Sedov blast wave problem. These panels show plots at time $t = 1$ of (a) the density, and (b) a slice of the density along $y = 0$. The solid line in (b) is the exact solution. The solution is obtained on a $160 \times 160$ mesh and a CFL number of 0.35.
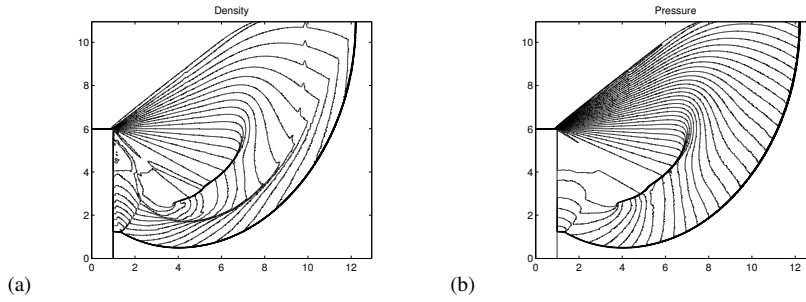


(a)                                                              (b)

**Fig. 4** 2D Shock diffraction problem. These panels show plots at time $t = 2.3$ of (a) the density, and (b) the pressure. A total of 20 equally spaced contour lines from $\rho = 0.0662$ to 7.07 are plotted in (a). We use 40 equally spaced contour lines from $p = 0.091$ to 37 are in (b). The solution is computed on a mesh with $\Delta x = \Delta y = 1/30$ and a CFL number of 0.35.

Picard integral formulation of the PDE, where a single finite difference WENO reconstruction is applied once per time step. A positivity-preserving limiter is introduced in such a way that the positivity of the solution is preserved for the *entire simulation*, which adds a degree of robustness to our scheme. In addition, we have no excessive CFL restriction in order to retain positivity, which makes our new method more efficient compared to recent positivity-preserving methods. We demonstrate the effectiveness and efficiency of the positivity-preserving schemes on one- and two-dimensional problems with low density and pressure. High-order accuracy is retained after the introduction of our positivity preserving limiter on a test problem that has near zero pressure and density. Future work includes the construction of positivity-preserving multiderivative methods [61], applying the proposed method to other systems such as magnetohydrodynamics, as well as incorporating our method into an AMR framework.

# References

1. P. Lax and B. Wendroff. Systems of conservation laws. *Comm. Pure Appl. Math.*, 13:217–237, 1960.
2. G. A. Sod. A survey of several finite difference methods for systems of nonlinear hyperbolic conservation laws. *J. Comput. Phys.*, 27(1):1–31, 1978.
3. A. Harten. The artificial compression method for computation of shocks and contact discontinuities. III. Self-adjusting hybrid schemes. *Math. Comp.*, 32(142):363–389, 1978.
4. P. L. Roe. Approximate Riemann solvers, parameter vectors, and difference schemes. *J. Comput. Phys.*, 43(2):357–372, 1981.
5. J. Von Neumann and R. D. Richtmyer. A method for the numerical calculation of hydrodynamic shocks. *J. Appl. Phys.*, 21:232–237, 1950.
6. R. Courant, E. Isaacson, and M. Rees. On the solution of nonlinear hyperbolic differential equations by finite differences. *Comm. Pure. Appl. Math.*, 5:243–255, 1952.
7. P. D. Lax. Weak solutions of nonlinear hyperbolic equations and their numerical computation. *Comm. Pure Appl. Math.*, 7:159–193, 1954.
8. S. K. Godunov. Difference method of computation of shock waves. *Uspehi Mat. Nauk (N.S.)*, 12(1(73)):176–177, 1957.
9. A. Harten, B. Engquist, S. Osher, and S. R. Chakravarthy. Uniformly high-order accurate essentially nonoscillatory schemes. III. *J. Comput. Phys.*, 71(2):231–303, 1987.
10. C.-W. Shu and S. Osher. Efficient implementation of essentially nonoscillatory shock-capturing schemes. *J. Comput. Phys.*, 77(2):439–471, 1988.
11. X.-D. Liu, S. Osher, and T. Chan. Weighted essentially non-oscillatory schemes. *J. Comput. Phys.*, 115(1):200–212, 1994.
12. T. Xiong, J.-M. Qiu, and Z. Xu. Parametrized positivity preserving flux limiters for the high order finite difference WENO scheme solving compressible Euler equations. http://arxiv.org/abs/1403.0594, 2014.
13. A. J. Christlieb, Y. Liu, Q. Tang, and Z. Xu. High order parametrized maximum-principle-preserving and positivity-preserving WENO schemes on unstructured meshes. *J. Comput. Phys.*, 281:334–351, 2015.
14. X. Zhang and C.-W. Shu. Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: survey and new developments. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 467(2134):2752–2776, 2011.
15. X. Zhang and C.-W. Shu. Positivity-preserving high order finite difference WENO schemes for compressible Euler equations. *J. Comput. Phys.*, 231(5):2245–2258, 2012.
16. A. J. Christlieb, Y. Güçlü, and D. C. Seal. The Picard integral formulation of weighted essentially nonoscillatory schemes. *SIAM J. Numer. Anal.*, 53(4):1833–1856, 2015.
17. A. Harten and G. Zwas. Self-adjusting hybrid schemes for shock computations. *J. Comput. Phys.*, 9:568–583, 1972.
18. G.-S. Jiang and C.-W. Shu. Efficient implementation of weighted ENO schemes. *J. Comput. Phys.*, 126(1):202–228, 1996.
19. C.-W. Shu. High order weighted essentially nonoscillatory schemes for convection dominated problems. *SIAM Rev.*, 51(1):82–126, 2009.
20. B. Cockburn, S. Y. Lin, and C.-W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. III. One-dimensional systems. *J. Comput. Phys.*, 84(1):90–113, 1989.
21. A. V. Rodionov. Methods of increasing the accuracy in Godunov's scheme. *USSR Computational Mathematics and Mathematical Physics*, 27(6):164 – 169, 1987.

22. J. B. Bell, P. Colella, and J. A. Trangenstein. Higher order Godunov methods for general systems of hyperbolic conservation laws. *J. Comput. Phys.*, 82(2):362–397, 1989.

23. I. S. Men'shov. Increasing the order of approximation of Godunov's scheme using solutions of the generalized Riemann problem. *USSR Computational Mathematics and Mathematical Physics*, 30(5):54 – 65, 1990.

24. E. F. Toro, R. C. Millington, and L. A. M. Nejad. Towards very high order godunov schemes. In *Godunov Methods*, pages 907–940. Springer, 2001.

25. E. F. Toro and V. A. Titarev. Solution of the generalized Riemann problem for advection-reaction equations. *R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci.*, 458(2018):271–281, 2002.

26. E. F. Toro and V. A. Titarev. ADER: Arbitrary high order Godunov approach. *J. Sci. Comput*, 17(1-4):609–618, 2002.

27. J. Qiu and C.-W. Shu. Finite difference WENO schemes with Lax-Wendroff-type time discretizations. *SIAM J. Sci. Comput.*, 24(6):2185–2198, 2003.

28. Y. Jiang, C.-W. Shu, and M. Zhang. An alternative formulation of finite difference weighted ENO schemes with Lax-Wendroff time discretization for conservation laws. *SIAM J. Sci. Comput.*, 35(2):A1137–A1160, 2013.

29. D. S. Balsara, T. Rumpf, M. Dumbser, and C.-D. Munz. Efficient, high accuracy ADER-WENO schemes for hydrodynamics and divergence-free magnetohydrodynamics. *J. Comput. Phys.*, 228(7):2480–2516, 2009.

30. D. S. Balsara, C. Meyer, M. Dumbser, H. Du, and Z. Xu. Efficient implementation of ADER schemes for Euler and magnetohydrodynamical flows on structured meshes—speed comparisons with Runge-Kutta methods. *J. Comput. Phys.*, 235:934–969, 2013.

31. M. Dumbser, Ol. Zanotti, A. Hidalgo, and D. S. Balsara. ADER-WENO finite volume schemes with space-time adaptive mesh refinement. *J. Comput. Phys.*, 248:257–286, 2013.

32. J. Qiu, M. Dumbser, and C.-W. Shu. The discontinuous Galerkin method with Lax-Wendroff type time discretizations. *Comput. Methods Appl. Mech. Eng.*, 194(42-44):4528–4543, 2005.

33. M. Dumbser and C.-D. Munz. Building blocks for arbitrary high order discontinuous Galerkin schemes. *J. Sci. Comput.*, 27(1-3):215–230, 2006.

34. G. Gassner, M. Dumbser, F. Hindenlang, and C.-D. Munz. Explicit one-step time discretizations for discontinuous Galerkin and finite volume schemes based on local predictors. *J. Comput. Phys.*, 230(11):4232–4247, 2011.

35. W. Liu, J. Cheng, and C.-W. Shu. High order conservative Lagrangian schemes with Lax-Wendroff type time discretization for the compressible Euler equations. *J. Comput. Phys.*, 228(23):8872–8891, 2009.

36. B. Einfeldt, C.-D. Munz, P. L. Roe, and B. Sjögreen. On Godunov-type methods near low densities. *J. Comput. Phys.*, 92(2):273–295, 1991.

37. J.-L. Estivalezes and P. Villedieu. A new second order positivity preserving kinetic schemes for the compressible euler equations. In S. M. Deshpande, S. S. Desai, and R. Narasimha, editors, *Fourteenth International Conference on Numerical Methods in Fluid Dynamics*, volume 453 of *Lecture Notes in Physics*, pages 96–100. Springer Berlin Heidelberg, 1995.

38. J.-L. Estivalezes and P. Villedieu. High-order positivity-preserving kinetic schemes for the compressible Euler equations. *SIAM J. Numer. Anal.*, 33(5):2050–2067, 1996.

39. B. Perthame and C.-W. Shu. On positivity preserving finite volume schemes for Euler equations. *Numer. Math.*, 73(1):119–130, 1996.

40. T. Tang and K. Xu. Gas-kinetic schemes for the compressible Euler equations: positivity-preserving analysis. *Z. Angew. Math. Phys.*, 50(2):258–281, 1999.

41. Bruno Dubroca. Solveur de Roe positivement conservatif. *C. R. Acad. Sci. Paris Sér. I Math.*, 329(9):827–832, 1999.

42. H.-Z. Tang and K. Xu. Positivity-preserving analysis of explicit and implicit Lax-Friedrichs schemes for compressible Euler equations. *J. Sci. Comput.*, 15(1):19–28, 2000.

43. Gérard Gallice. Positive and entropy stable Godunov-type schemes for gas dynamics and MHD equations in Lagrangian or Eulerian coordinates. *Numer. Math.*, 94(4):673–713, 2003.

44. X. Zhang and C.-W. Shu. On positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations on rectangular meshes. *J. Comput. Phys.*, 229(23):8918–8934, 2010.

45. X. Zhang and C.-W. Shu. Positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations with source terms. *J. Comput. Phys.*, 230(4):1238–1248, 2011.

46. X. Zhang, Yi. Xia, and C.-W. Shu. Maximum-principle-satisfying and positivity-preserving high order discontinuous Galerkin schemes for conservation laws on triangular meshes. *J. Sci. Comput.*, 50(1):29–62, 2012.

47. K. Kontzialis and J. A. Ekaterinaris. High order discontinuous Galerkin discretizations with a new limiting approach and positivity preservation for strong moving shocks. *Comput. & Fluids*, 71:98–112, 2013.

48. X. Y. Hu, N. A. Adams, and C.-W. Shu. Positivity-preserving method for high-order conservative schemes solving compressible Euler equations. *J. Comput. Phys.*, 242:169–180, 2013.

49. J. A. Rossmanith and D. C. Seal. A positivity-preserving high-order semi-Lagrangian discontinuous Galerkin scheme for the Vlasov-Poisson equations. *J. Comput. Phys.*, 230(16):6203–6232, 2011.

50. Y. Güçlü, A. J. Christlieb, and W. N. G. Hitchon. Arbitrarily high order convected scheme solution of the Vlasov-Poisson system. *J. Comput. Phys.*, 270:711–752, 2014.

51. J. P. Boris and D. L. Book. Flux-corrected transport. I. SHASTA, a fluid transport algorithm that works. *J. Comput. Phys.*, 11(1):38 – 69, 1973.

52. D. L. Book, J. P. Boris, and K. Hain. Flux-corrected transport II: Generalizations of the method. *J. Comput. Phys.*, 18(3):248 – 283, 1975.

53. J. P. Boris and D. L. Book. Flux-corrected transport. III. Minimal-error FCT algorithms. *J. Comput. Phys.*, 20(4):397 – 431, 1976.

54. D. L. Book, J. P. Boris, and S. T. Zalesak. Flux-corrected transport. In *Finite-difference techniques for vectorized fluid dynamics calculations*, Springer Ser. Comput. Phys., pages 29–55. Springer, New York, 1981.

55. Z. Xu. Parametrized maximum principle preserving flux limiters for high order schemes solving hyperbolic conservation laws: one-dimensional scalar problem. *Math. Comp.*, 83(289):2213–2238, 2014.

56. C. Liang and Z. Xu. Parametrized maximum principle preserving flux limiters for high order schemes solving multi-dimensional scalar hyperbolic conservation laws. *J. Sci. Comput.*, 58(1):41–60, 2014.

57. T. Xiong, J.-M. Qiu, and Z. Xu. A parametrized maximum principle preserving flux limiter for finite difference RK-WENO schemes with applications in incompressible flows. *J. Comput. Phys.*, 252:310–331, 2013.

58. A. J. Christlieb, Y. Liu, Q. Tang, and Z. Xu.  Positivity-preserving finite difference
    weighted eno schemes with constrained transport for ideal magnetohydrodynamic equa-
    tions. *SIAM J. Sci. Comput.*, 37(4):A1825–A1845, 2015.
59. D.C. Seal. FINESS software, 2015. Available from
    `https://bitbucket.org/dseal/finess`.
60. L. I. Sedov. *Similarity and dimensional methods in mechanics*. Academic Press, 1959.
61. David C. Seal, Yaman Güçlü, and Andrew J. Christlieb. High-order multiderivative time
    integrators for hyperbolic conservation laws. *J. Sci. Comput.*, 60(1):101–140, 2014.