Preserving Nonnegativity in Discontinuous Galerkin Approximations to Scalar Transport via Truncation and Mass Aware Rescaling (TMAR)

DEVIN LIGHT[a] AND DALE DURRAN

University of Washington, Seattle, Washington

(Manuscript received 10 June 2016, in final form 9 August 2016)

ABSTRACT

An accurate nonnegativity preserving limiter is presented for use with discontinuous Galerkin (DG) discretizations of scalar advection equations. The nonnegativity of the tracer field is preserved through the application of a mass conservative limiter that truncates negatives within each element and linearly rescales the resulting DG polynomials to preserve element-mean mass. As a preliminary step, the DG fluxes through each side of the element are limited in a manner similar to flux-corrected transport to ensure that the element-mean mass remains nonnegative during each individual stage of the time integration. In this paper, it is proven that such a truncation and mass aware rescaling (TMAR) does not change the order of accuracy of the underlying unlimited DG approximation. Numerical tests with two-dimensional deforming flows confirm that the method remains accurate and efficient while preserving nonnegativity. In comparison to some popular previous approaches, TMAR limiting is particularly well suited to approximations that use high-degree polynomial expansions (quartics or higher) to capture features that are only moderately well resolved.

1. Introduction

Discontinuous Galerkin (DG) finite element methods are an increasingly popular means of producing numerical approximations to systems of hyperbolic conservation laws. Methods from this family are attractive because they are high-order accurate, geometrically flexible, *h–p* adaptive, compactly defined, and scale well on distributed memory systems (Giraldo et al. 2002). In this paper we will consider DG approximations to the one- and two-dimensional transport of an inert scalar tracer advected by a flow with velocity $\mathbf{u}(\mathbf{x}, t)$ through a spatial domain Ω having suitable boundary conditions. Letting $\psi(\mathbf{x}, t)$ denote the tracer concentration and $\rho(\mathbf{x}, t)$ the density, the tracer evolution satisfies

$$\frac{\partial}{\partial t}(\rho\psi) + \nabla\cdot(\rho\mathbf{u}\psi) = 0, \quad (\mathbf{x}, t) \in \Omega \times \mathbb{R}^+$$
$$\psi(\mathbf{x}, t = 0) = \psi_0(\mathbf{x}). \tag{1}$$

Using the dry mass continuity equation, the preceding may be written in the advective form:

$$\frac{\partial\psi}{\partial t} + \mathbf{u}\cdot\nabla\psi = 0. \tag{2}$$

Analytic solutions to (2) satisfy the boundedness condition that if $m \le \psi_0 \le M$ for all $\mathbf{x} \in \Omega$, then $m \le \psi(\mathbf{x}, t) \le M$ for all $t \ge 0$ and all $\mathbf{x} \in \Omega$. When the flow is nondivergent (i.e., $\nabla\cdot\mathbf{u} = 0$), solutions to (1) satisfy the same boundedness condition. However, even in divergent flows, solutions to (1) will nevertheless still satisfy the nonnegativity condition that if $0 \le \psi_0$ for all $\mathbf{x} \in \Omega$, then $0 \le \psi(\mathbf{x}, t)$ or all $t \ge 0$ and all $\mathbf{x} \in \Omega$. In the following we focus on the case of nondivergent flow with constant density $\rho = 1$ to keep the mathematical notation associated with solutions to (1) concise, but our results are easily extended to divergent flows and nonuniform density fields.

Beyond emulating a physical quality of the analytic solution, maintaining the nonnegativity of the numerical solution can be of great importance to stability, particularly when (1) is generalized to include nonlinear source and sink terms such as chemical reactions or cloud microphysical processes. In practice, the tendencies from these sources and sinks are often integrated separately from the advective tendencies through techniques like

[a] Current affiliation: Department of Applied Mathematics, University of Washington, Seattle, Washington.

Corresponding author address: Dale Durran, Dept. of Atmospheric Sciences, University of Washington, Box 351640, Seattle, WA 98125.
E-mail: drdee@uw.edu

DOI: 10.1175/MWR-D-16-0220.1

© 2016 American Meteorological Society

operator splitting. An example of this type of a problem in a geophysical context is the reactive transport system considered in Lauritzen et al. (2015). If spurious negative species concentrations are generated in the transport split step, they can quickly destabilize the solution by inducing reactions in the chemistry split step that would otherwise be impossible (Durran 2010). When a high-order method such as a DG scheme is used to simulate the transport of data, which contains poorly resolved steep gradients, Gibbs-like oscillations can generate spurious negatives. It is, therefore, often necessary to augment the standard DG discretization by adding a limiter to preserve nonnegativity when simulating tracer transport in the presence of nonlinear sources and sinks.

Proposed limiters have taken a variety of approaches such as adding artificial viscosity (Hartmann and Houston 2002; Persson and Peraire 2006), extending classical TVB limiters (Cockburn and Shu 1989; Cockburn et al. 1989), weighted nonoscillatory (WENO) DG limiters (Qiu and Shu 2004, 2005a, 2005b), the solution of a quadratic minimization problem (Guba et al. 2014), and a posteriori limiting (Dumbser et al. 2014). For a brief review of these methods, see Dumbser et al. (2014). Another widely used method for keeping DG approximations to conservation laws nonnegative is to adapt the bounds preserving limiter proposed in Zhang and Shu (2010, 2011), hereafter the ZS limiter, for use as a nonnegativity preserving limiter (Rossmanith and Seal 2011; Qiu and Shu 2011; Guo et al. 2014). The ZS limiter ensures nonnegativity by introducing a conservative linear rescaling originally proposed in Liu and Osher (1996) that, when combined with a suitably limited time step, prevents the element-integrated tracer mass from becoming negative. The ZS limiter is attractive for several reasons: it preserves high-order accuracy, is locally defined, and is straightforward to implement alongside existing methods (Zhang and Shu 2010).

We propose an alternative limiter that enjoys many of the same benefits as the ZS limiter but can perform better for higher-degree polynomial approximations with similar computational effort. Like the ZS approach, nonnegativity is preserved in two stages. In the first stage a flux-corrected transport (FCT) adjustment is made to the numerical fluxes just before the integration step to ensure that the element-integrated tracer mass remains nonnegative after the step. After the step, the local DG polynomial representation is corrected to remove any negatives that might have developed within the element using a nonlinear adjustment that truncates negatives values to zero while rescaling the concentrations at the other nodes to conserve mass.

The remainder of the paper is structured as follows. In section 2 we describe the basic DG framework and the ZS limiter. Section 3 presents the proposed nonnegativity preserving truncation and mass aware rescaling (TMAR) limiter. Section 4 examines the empirical performance of the TMAR limiter on several one-dimensional and two-dimensional test problems. Section 5 investigates the computational expense of implementing the TMAR limiter, and section 6 contains our conclusions.

2. The groundwork

a. Basic DG formulation

The proposed method is based on the standard Runge–Kutta DG (RKDG) formulation presented in Hesthaven and Warburton (2008), Durran (2010), and Ramachandran et al. (2011). We divide the computational domain Ω into nonoverlapping elements s_i and approximate the solution $\psi(\mathbf{x}, t)$ within element s_i as an expansion of basis polynomials $\varphi_k(\mathbf{x})$ defined locally over each element,

$$\phi_{s_i}(\mathbf{x},t) = \sum_k a_{i,k}(t)\varphi_k(\mathbf{x}) \quad \text{for} \quad \mathbf{x} \in s_i, \tag{3}$$

where the summation is taken over the total number of basis polynomials. For one-dimensional problems, the most common choices for the basis polynomials are Legendre polynomials (leading to modal methods) or Lagrange polynomials (leading to nodal methods). Lagrange polynomials of degree N are chosen to interpolate the $N+1$ Gauss–Legendre–Lobatto (GLL) quadrature nodes x_k, which have been mapped to s_i.

The DG approximation to (1) is obtained by multiplying the differential equation by the test function $\varphi_k(\mathbf{x})$ and formally integrating by parts over s_i to get

$$\frac{d}{dt}\int_{s_i}\phi_{s_i}(\mathbf{x},t)\varphi_k(\mathbf{x})\,d\mathbf{x} = \int_{s_i}\phi_{s_i}(\mathbf{x},t)\mathbf{u}(\mathbf{x},t)\cdot\nabla\varphi_k(\mathbf{x})\,d\mathbf{x} - \int_{\partial s_i}\phi_{s_i}(s,t)\mathbf{u}(s,t)\cdot\mathbf{n}\varphi_k(s)\,ds \quad \forall\varphi_k, \tag{4}$$

where $\mathbf{n}$ is the outward-facing unit normal on the boundary of s_i. The flux term $F(\phi_{s_i}) = \phi_{s_i}(s, t)\mathbf{u}(s, t)$ in the boundary integral in (4) is not uniquely defined because ϕ can be discontinuous across element interfaces. To specify a unique approximation this term is replaced with a numerical flux function $\hat{F}(\cdot, \cdot)$. A simple choice for transport equations is the upwind flux:

$$\hat{F}(\phi_h^-, \phi_h^+) = \begin{cases} \mathbf{u}\phi_h^- & \text{if} \quad \mathbf{u}\cdot\mathbf{n} \geq 0 \\ \mathbf{u}\phi_h^+ & \text{if} \quad \mathbf{u}\cdot\mathbf{n} < 0 \end{cases}, \tag{5}$$

where ϕ_h^- refers to the local (or interior) solution at the interface and ϕ_h^+ refers to the neighbor (or exterior) solution. Although upwind fluxes result in a first-order scheme in standard finite-volume methods, they do not adversely affect the spectral convergence of DG schemes.

Substituting the expansion (3) into (4) leads to the matrix equation:

$$\mathbf{M}\frac{d\mathbf{a}_i}{dt} = \mathbf{G}, \tag{6}$$

where $\mathbf{a}_i$ is the vector of expansion coefficients on element s_i, $\mathbf{M}$ is the mass matrix with entries:

$$\mathbf{M}_{j,k} = \int_{s_i} \varphi_j(\mathbf{x})\varphi_k(\mathbf{x})\, d\mathbf{x}, \tag{7}$$

and $\mathbf{G}$ is the vector:

$$\mathbf{G}_k = \int_{s_i} \phi_{s_i}(\mathbf{x},t)\mathbf{u}(\mathbf{x},t)\cdot\nabla\varphi_k(\mathbf{x})\, d\mathbf{x} - \int_{\partial s_i} \hat{F}(\phi_{s_i}^-, \phi_{s_i}^+)\cdot\mathbf{n}\varphi_k(s)\, ds. \tag{8}$$

The integrals in (7) and (8) are approximated using numerical quadrature. Nodal methods typically perform the quadrature on the GLL nodes, while modal methods use more accurate Gaussian quadrature. As a consequence of the orthogonality of Legendre polynomials, $\mathbf{M}$ is diagonal for modal methods, which allows (6) to be integrated using explicit time-stepping schemes. On the other hand, if the mass-matrix integral in (7) for a nodal method were to be evaluated exactly, $\mathbf{M}$ would be dense (Durran 2010, p. 344), coupling the time derivatives for all expansions coefficients within each element and leading to implicit algebraic equations for the $\mathbf{a}_i$ at each new time step. However, if the nodal mass matrix is instead approximated using the GLL integration rule implied by the underlying interpolation grid, the resulting mass matrix will be diagonal. This technique, sometimes known as mass lumping (Karniadakis and Sherwin 2005, p. 57), is used to obtain our nodal solutions. Regardless of quadrature technique, the semidiscrete system in (6) is integrated using the three-stage, third-order strong stability preserving Runge–Kutta (SSPRK) method in Gottlieb et al. (2009) for the numerical simulations performed in this paper.

It is computationally expensive, particularly in multidimensional problems, to ensure the solution is nonnegative at all $\mathbf{x}$ within element s_i, and it is unnecessary. In practice all that is required is that the solution be nonnegative over the set of subelement data that would be used in more general problems to compute (via operator splitting) interactions between tracer species that require nonnegativity. For nodal methods the natural choice for the subelement data is simply the numerical solution at the tensor-product GLL nodes, which as a consequence of Lagrange interpolation, are the values of the expansion coefficients themselves. For modal methods in d-dimensional space, we specify the subelement data to be the averages over a uniform Cartesian subgrid denoted by $s_{i,k}$, $k = 1, \ldots, (N+1)^d$ such that $s_i = \cup_k s_{i,k}$. Letting $|s_{i,k}|$ denote the length, area or volume of $s_{i,k}$, the average mass over each subgrid element can be evaluated as

$$\phi_{i,k} = \frac{1}{|s_{i,k}|}\int_{s_{i,k}} \phi_{s_i}(\mathbf{x},t)\, d\mathbf{x}. \tag{9}$$

The set of modal expansion coefficients $a_{i,k}$ in element s_i can be mapped to the set of $\phi_{i,k}$ by a projection operator P such that $\phi_i = P\mathbf{a}_i$. After the subgrid-element averages are adjusted for nonnegativity, a reconstruction operator R is applied to map these modified averages back to the polynomial coefficients at the beginning of the next time step. Because the subelement data have the same number of degrees of freedom as the original polynomial approximation, the matrix representation of P will be nonsingular and R will be its inverse P^{-1}.

b. ZS nonnegativity preservation

As the first half of a two-part formulation, the ZS limiter preserves element-mean nonnegativity through a restriction on the length of the time step (Zhang and Shu 2010). The largest time step for which element-mean nonnegativity is guaranteed is obtained by preserving nonnegativity at the set of $\hat{L}$ GLL nodes, where $\hat{L}$ is the smallest integer for which an $\hat{L}$-point GLL quadrature is exact for polynomials of degree N (i.e., $2\hat{L} - 3 \geq N$). Then if the quadrature weights w_k are expressed assuming the coordinate in each DG element has been rescaled to the interval $[-1, 1]$, the largest Courant number $\mu_{\max}$ for which a one-dimensional or Strang split multidimensional ZS method is guaranteed to preserve element-mean nonnegativity satisfies the following:

$$\mu_{\max} = \max_{x\in\Omega}|u|\frac{\Delta t}{\Delta x} \leq \min_k \frac{w_k}{2}. \tag{10}$$

Zhang and Shu (2010) also showed that two-dimensional schemes must obey the more restrictive condition:

$$\mu_x + \mu_y \leq \min_k \frac{w_k}{2}, \tag{11}$$

where

$$\mu_x = \max_{(x,y)\in\Omega} |u| \frac{\Delta t}{\Delta x}, \quad \mu_y = \max_{(x,y)\in\Omega} |v| \frac{\Delta t}{\Delta y}.$$

If the mesh is isotropic and $\max|u| = \max|v|$, this condition will correspond to a maximum time step that is half as large as in the corresponding one-dimensional problem.

After the integration step, the values at some of the GLL nodes may have become negative. These negatives must be eliminated before the next forward step, and the ZS limiter removes them through a linear rescaling. If ϕ is the local DG approximation to ψ over the element s_i with element-mean $\overline{\phi}$, the rescaled polynomial ϕ^* is computed as

$$\phi^* = \theta(\phi - \overline{\phi}) + \overline{\phi}, \quad \theta = \min\left\{\frac{|0 - \overline{\phi}|}{|m - \overline{\phi}|}, 1\right\}, \quad (12)$$

where

$$m = \min_{\mathbf{x}^*\in\Gamma} \phi(\mathbf{x}^*), \quad (13)$$

and in the most straightforward implementation, Γ is simply the set of $\hat{L}$ GLL nodes in element s_i.

Zhang and Shu (2011) noted that a more efficient implementation is possible, which in one-dimensional problems requires just three points: the two points at the edges of the element and a third internal point $\hat{x}^*$ whose value $\phi(\hat{x}^*)$ can be written in terms of the values of ϕ at the edges. This gain in efficiency becomes important in two or more dimensions, when it can significantly reduce the number of points at which $\phi(\hat{\mathbf{x}}^*)$ needs to be evaluated. For example, in two-dimensional problems with Cartesian elements, Γ will consist of points along the edges of the element plus an additional internal point $\hat{\mathbf{x}}^*$ whose value $\phi(\hat{\mathbf{x}}^*)$ may be determined from integrals of the aforementioned edge values. The minimum number of nodes along each edge required to integrate a one-dimensional polynomial of degree N exactly are L Gaussian quadrature points such that $2L - 1 \geq N$. Although it might seem faster to use the $\hat{L}$ GLL nodes along each boundary at which the solution is already known, in two dimensions it is more efficient to use the minimum number of Gaussian quadrature points because that set will have the largest minimum weight w_k and, therefore, allow the largest time step. Use of this larger time step offsets the extra computational effort required to evaluate the solution at the Gaussian quadrature points along each boundary.

Using the minimum number of nodes to preserve nonnegativity, as above, yields the largest minimum w_k and permits the largest possible time step as per (10) or (11), but it can allow negative values to develop on tensor product mesh, which contains $N + 1$ nodes along each coordinate. Therefore, an additional application of (12) is required in which Γ is the set of all nodes on the tensor product mesh. In our three-stage third-order SSPRK time integration, negatives are eliminated by applying (12) at the minimal number of nodes prior to each Runge–Kutta stage, and after the last stage (prior to the evaluation of any hypothetical coupling between scalars that requires nonnegativity), (12) is applied one additional time to eliminate negatives at the tensor product nodes.

3. TMAR nonnegativity preservation

a. One-dimensional formulation

In the context of simple forward time differencing, the basic DG algorithm is modified in two ways:

1) The numerical fluxes at element boundaries are adjusted prior to each forward step to ensure that the mean tracer concentration in each element remains nonnegative after the step.
2) After each time step, the solution inside the element is conservatively modified to remove any negative tracer concentrations in the discrete subelement data.

Let us consider the first adjustment. If $\phi(x, t)$ is an approximate solution generated by a DG method, then a scheme for a forward-in-time update of each element mean, or equivalently a numerical approximation to (4) when $\varphi_{k=1}$, can be written as follows:

$$\overline{\phi}_{s_i}^{n+1} = \overline{\phi}_{s_i}^{n} - \frac{\Delta t}{\Delta x}[\hat{F}(\phi^-_{i+(1/2)}, \phi^+_{i+(1/2)})^n - \hat{F}(\phi^-_{i-(1/2)}, \phi^+_{i-(1/2)})^n]. \quad (14)$$

To ensure that $\overline{\phi}_{s_i}^{n+1} \geq 0$, the standard upstream fluxes in (14) are replaced with modified fluxes $F^*_{i\pm(1/2)}$ determined as described in Smolarkiewicz (1989). This approach is a special case of the flux-corrected transport algorithms that were originally developed for finite-volume methods but have also been employed for element-mean nonnegativity in finite-element methods (Restelli et al. 2006; Ullrich and Norman 2014). The details of this algorithm can be found in Durran (2010) and Smolarkiewicz (1989), and a detailed implementation of this method in two dimensions is presented in section 3b. For higher-order SSPRK time stepping, the FCT adjustment is applied to the fluxes during each forward step in the integration, and since the SSPRK methods are convex

combinations of forward Euler steps, the full multistage update will also satisfy the element-mean nonnegativity. One important benefit of this FCT algorithm is that it does not impose an additional limitation on the length of the time step.

The second step of the proposed limiter applies a nonlinear truncation and mass aware rescaling (TMAR) to the $N+1$ subelement values in which negatives are truncated to zero and the remaining nonnegative values are rescaled to conserve mass. Such a rescaling will always be possible because the mean value of the approximate solution produced by the FCT limited forward step $\overline{\phi}_{s_i}$ is nonnegative. The truncation produces an intermediate approximation $\phi^+_{i,k}$ given by

$$\phi^+_{i,k} = \begin{cases} \phi_{i,k} & \text{if} \quad \phi_{i,k} \geq 0 \\ 0 & \text{if} \quad \phi_{i,k} < 0 \end{cases}.$$

The element-mean mass after truncation $\overline{\phi}^+_{s_i}$ is used to compute a rescaling ratio:

$$r_i = \frac{\overline{\phi}_{s_i}}{\overline{\phi}^+_{s_i}}. \tag{15}$$

Since $\overline{\phi}_{s_i} \geq 0$ and the truncation adds mass to the element, $0 \leq r_i \leq 1$. Finally, the original subelement values are replaced by

$$\phi^*_{i,k} = \begin{cases} r_i \phi_{i,k} & \text{if} \quad \phi_{i,k} \geq 0 \\ 0 & \text{if} \quad \phi_{i,k} < 0 \end{cases} \tag{16}$$

whose element-mean satisfies

$$\overline{\phi}^*_{s_i} = r_i \overline{\phi}^+_{s_i} = \overline{\phi}_{s_i}, \tag{17}$$

guaranteeing conservation. In our three-stage third-order SSPRK time integration, element-mean negatives are avoided by applying the FCT flux limiter during each Runge–Kutta stage. After the last stage (prior to any hypothetical chemistry), the TMAR limiter is applied to eliminate negatives at the tensor product nodes, or in the case of modal DG, in any of the equal subelement volumes.

Figure 1 illustrates the difference between the linear rescaling used in the ZS limiter (shown in red) and the TMAR adjustment described above (shown in green) when applied to a sample fifth-degree nodal polynomial $\phi(x)$ (shown in blue) having negative values at three of the GLL nodes: two at the element edges and one near the center. After applying the linear ZS rescaling, the two largest magnitude negative nodal values have been scaled to zero while the smaller magnitude negative in the center is pushed into positive values. Moving this node into positive values requires an unnecessarily large amount of mass be redistributed from the other positive nodes to maintain conservation. In contrast, the TMAR modification, which simply sets the value at this node to zero, produces less damping at the positive nodes.

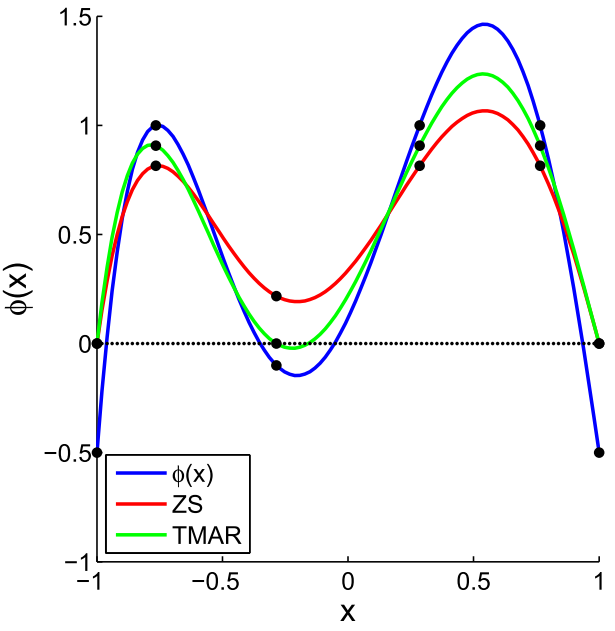

FIG. 1. Fifth-degree polynomial $\phi(x)$ with negative values at three GLL nodes (blue) and the polynomial as modified by linear the ZS rescaling in (12) (red) and by TMAR (green). The GLL nodal values are indicated by black dots.

Neither the TMAR nor the ZS-adjusted solutions need be nonnegative at every point in the element. Nonnegativity is ensured only for the set of subelement data that could potentially be used in a subsequent time-split calculation of chemical reactions or cloud-microphysical tendencies. In the case shown in Fig. 1, the subelement data are the nodal values, and the TMAR-adjusted result remains slightly negative over a small region.

We conclude this section with a theorem guaranteeing that TMAR limiting does not degrade the high-order accuracy of the underlying DG method; the proof is given in the appendix.

Theorem 1. Let $\phi_{s_i}(x)$ be the unmodified Mth-order DG approximation to $\psi(x)$ in s_i at some arbitrary time, subject to the constraint that $\overline{\phi}_{s_i} \geq 0$. Then the TMAR limited solution $\phi^*_{s_i}(x)$ is also an Mth-order approximation to $\psi(x)$.

b. *Two-dimensional formulation*

The TMAR limiter described in section 3 can be readily extended to multidimensional problems. Let $\overline{\phi}_{i,j}$

denote the average value of the degree N local DG approximating polynomial within a rectangular element $s_{i,j}$. A DG scheme for updating $\overline{\phi}_{i,j}$ in a forward Euler step from time t^n will satisfy the following:

$$\overline{\phi}_{i,j}^{n+1} = \overline{\phi}_{i,j}^{n} - \frac{\Delta t}{\Delta x \Delta y}\left\{\int_{y_{j-(1/2)}}^{y_{j+(1/2)}} [F_{i+(1/2)}^{n}(y) - F_{i-(1/2)}^{n}(y)]\,dy + \int_{x_{i-(1/2)}}^{x_{i+(1/2)}} [G_{j+(1/2)}^{n}(x) - G_{j-(1/2)}^{n}(x)]\,dx\right\}, \tag{18}$$

where Δx and Δy are the length of the elements along the x and y coordinates, respectively; and $F_{i\pm(1/2)}^{n}(y)$ and $G_{j\pm(1/2)}^{n}(x)$ are the numerical flux functions at time t^n through the interfaces perpendicular to the x and y coordinates, respectively. Equation (18) can be rewritten more compactly in terms of the mean fluxes $\tilde{F}_{i\pm(1/2)}$ and $\tilde{G}_{j\pm(1/2)}$ through each interface as

$$\overline{\phi}_{i,j}^{n+1} = \overline{\phi}_{i,j}^{n} - \frac{\Delta t}{\Delta x \Delta y}[\Delta y(\tilde{F}_{i+(1/2)}^{n} - \tilde{F}_{i-(1/2)}^{n}) + \Delta x(\tilde{G}_{j+(1/2)}^{n} - \tilde{G}_{j-(1/2)}^{n})]. \tag{19}$$

Notice that (19) is in the same form as a finite-volume update to $\overline{\phi}_{i,j}^{n}$ using the mean fluxes. With this in mind, we apply the standard multidimensional FCT algorithm presented in Zalesak (1979) to (19) to determine corrected mean fluxes $\tilde{F}_{i\pm(1/2)}^{*}$ and $\tilde{G}_{j\pm(1/2)}^{*}$, which will not drive $\overline{\phi}_{i,j}^{n}$ negative. For completeness, this approach is summarized below:

1) Let $Q_{i,j}$ be the maximum outward flux sustainable over a single time step without forcing $\overline{\phi}_{i,j}^{n+1}$ negative:

$$Q_{i,j} = \frac{\overline{\phi}_{i,j}^{n}\Delta x \Delta y}{\Delta t}.$$

2) Let $P_{i,j}$ be the net mean flux out of element $s_{i,j}$, given by

$$P_{i,j} = \Delta y[\max(0, \tilde{F}_{i+(1/2)}^{n}) - \min(0, \tilde{F}_{i-(1/2)}^{n})] + \Delta x[\max(0, \tilde{G}_{j+(1/2)}^{n}) - \min(0, \tilde{G}_{j-(1/2)}^{n})].$$

3) Determine the ratio by which the mean fluxes will be corrected to ensure that a negative concentration will not be generated:

$$R_{i,j} = \min\left(1, \frac{Q_{i,j}}{P_{i,j} + \varepsilon}\right),$$

where ε is a small parameter (nominally 10^{-10} times a typical magnitude for ψ) that is added to avoid division by zero.

4) Evaluate the corrected mean fluxes such that

$$\tilde{F}_{i+(1/2)}^{*} = \begin{cases} R_{i,j}\tilde{F}_{i+(1/2)}^{n} & \text{if} \quad \tilde{F}_{i+(1/2)}^{n} \geq 0 \\ R_{i+1,j}\tilde{F}_{i+(1/2)}^{n} & \text{if} \quad \tilde{F}_{i+(1/2)}^{n} < 0 \end{cases},$$

$$\tilde{G}_{j+(1/2)}^{*} = \begin{cases} R_{i,j}\tilde{G}_{j+(1/2)}^{n} & \text{if} \quad \tilde{G}_{j+(1/2)}^{n} \geq 0 \\ R_{i,j+1}\tilde{G}_{j+(1/2)}^{n} & \text{if} \quad \tilde{G}_{j+(1/2)}^{n} < 0 \end{cases}.$$

This approach yields a modification to the mean fluxes that will keep $\overline{\phi}_{i,j}$ nonnegative. However, in practice it is the pointwise fluxes, evaluated at quadrature locations around the boundary of the element that are required for numerical evaluation of the integrals in (18). Therefore, it is necessary to map the modification to the mean fluxes into an equivalent modification of the pointwise fluxes. Let ξ_k and w_k denote the one-dimensional GLL quadrature points and weights along the element boundary centered at $(x_{i+(1/2)}, y_j)$. The mean flux through this boundary is the linear combination of the pointwise fluxes:

$$\tilde{F}_{i+(1/2)} = \frac{1}{2}\sum_{k=0}^{N} w_k F_{i+(1/2)}(\xi_k). \tag{20}$$

We adopt the simple approach of applying the FCT multiplicative correction factor for the mean flux to each pointwise flux. In other words, if $\tilde{F}_{i+(1/2)}^{*} = c\tilde{F}_{i+(1/2)}$ for some correction factor $0 \leq c \leq 1$, then the modified nodal fluxes $F_{i+(1/2)}^{*}(\xi_k)$ that will be used in the forward step are given by $F_{i+(1/2)}^{*}(\xi_k) = cF_{i+(1/2)}(\xi_k)$.

These FCT flux corrections are applied in each stage of the SSPRK integration so that the element-mean concentrations will be nonnegative after the last stage. Immediately after this last stage (before any hypothetical chemistry step), the TMAR limiter is used to truncate negatives and rescale the positive concentrations in the discrete subelement data. The rescaling ratio is computed in a manner similar to (15),

$$r_{i,j} = \frac{\overline{\phi}_{i,j}}{\overline{\phi}_{i,j}^{+}}, \tag{21}$$

where $\overline{\phi}_{i,j}^{+}$ is the mean value of the truncated approximation over the two-dimensional subelement data. The limited subgrid-element averages $\phi_{i,j,k}^{*}$ for $k = 1, \ldots, (N+1)^2$ are then given by

$$\phi_{i,j,k}^{*} = \begin{cases} r_{i,j}\phi_{i,j,k} & \text{if} \quad \phi_{i,j,k} > 0 \\ 0 & \text{if} \quad \phi_{i,j,k} < 0 \end{cases}. \tag{22}$$

A proof similar to that for Theorem A in the appendix shows that this two-dimensional limiter does not reduce

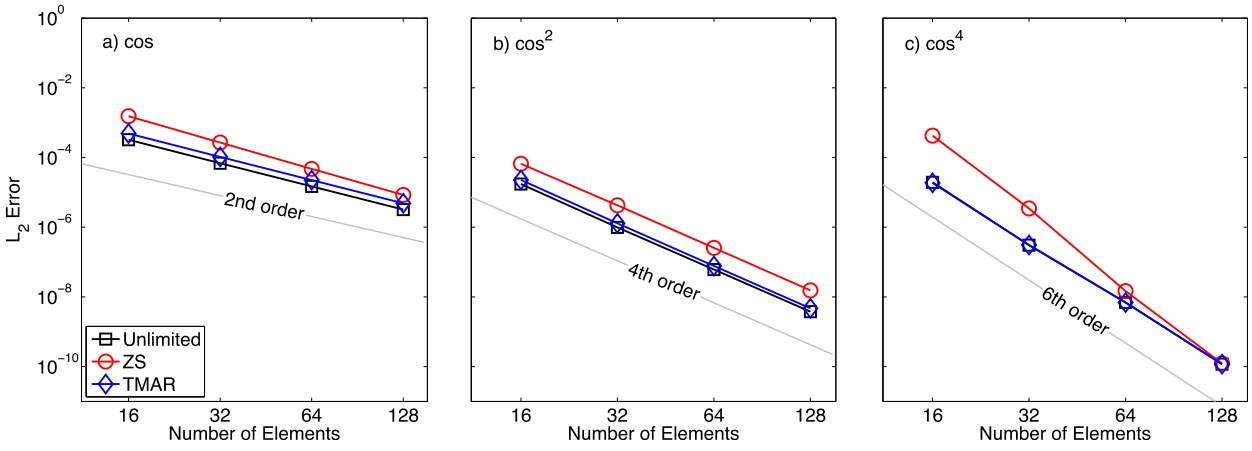

FIG. 2. Convergence results under h refinement: log–log plot of the L_2 error as a function of the number of elements for (a) C^1, (b) C^3 and, (c) C^7 cosine bell tests with initial conditions defined in (23). Each test uses a nodal basis with $N = 5$ and a time step of $\Delta t = 0.5\Delta x^2$.

the asymptotic convergence rate of the underlying DG approximation.

4. Numerical tests

a. One-dimensional tests

One of the most important benefits of implementing DG methods is that they allow the flexibility of refining the approximate solution by either adding additional elements (h refinement) or using local polynomials of higher degree (p refinement). Therefore, we will consider the influence of each limiter on the h- and p-convergence rates for smooth initial data. Let $\tau(x) = 4|x - 1/4|$ and define an initial tracer density as a member of the family of cosine bells:

$$\psi_{0,q}(x) = \begin{cases} \left[\dfrac{1 + \cos(\pi\tau)}{2}\right]^q & \text{if} \quad \tau \le 1 \\ 0 & \text{otherwise} \end{cases}, \qquad (23)$$

where $q = 1$, 2, or 4. Because $\psi_{0,q}$ has $2q - 1$ continuous derivatives ($\psi_{0,q}$ is C^{2q-1}), larger values of q permit greater convergence rates as the degree p of the DG polynomial truncation is increased. Although $\psi_{0,q}$ is nonnegative, for all values of q considered here, the initial profile contains gradients steep enough to generate negative concentrations in the unlimited numerical solution. In these tests, ψ is advected by a constant wind speed $u = 1$ around the periodic domain $\Omega = [0, 1]$.

Figure 2 illustrates the impact of the limiting methods on the h convergence of the L_2 error. The three methods shown use a nodal basis of fifth-degree polynomials with a time step chosen for the SSPRK3 integration so that $\Delta t \propto (\Delta x)^2$, which guarantees that spatial convergence rates are observed. For the C^1 and C^3 tests the observable convergence rates (i.e., the slopes in Fig. 2) are constrained by the smoothness of the analytic solution, which limits the observed order of accuracy to be roughly second and fourth order, respectively. However, in the third panel the optimal sixth-order accuracy of the method is observed in the unlimited-method results. While both limited solutions are slightly less accurate than the unmodified solution, their convergence rates nevertheless closely match the unlimited convergence rate for all three initial conditions considered. This result is consistent with the proofs that both methods preserve the original convergence rate of the unlimited method under h refinement.

Figure 3 examines the impact of limiting on p-convergence rates for the same initial conditions considered in Fig. 2. To measure the impact of p refinement we define the effective spacing $\Delta x_e = \Delta x/N$ to be the average grid spacing between the GLL nodes. A time step $\Delta t \propto (\Delta x)^{(N+1)/3}$ is chosen so that the spatial convergence rates are observed, and the local polynomial degree is refined from 4 to 9 using a fixed mesh of 32 elements. As in the h-refinement tests, the convergence rates are influenced by smoothness of the exact solutions. The unlimited and the TMAR solutions both produce very similar errors that decrease as the polynomial degree increases (i.e., as Δx_e decreases), attaining roughly fourth-order convergence for the C^1 test, eighth order in the C^3 test, and approximately twenty-second order in the C^7 test. On the other hand, the errors in the ZS solutions are not reduced as the polynomial-order increases in the C^1 and C^3 tests. The ZS solutions do show some improvement under p refinement in the C^7 test, although this may be primarily due to the reduction in

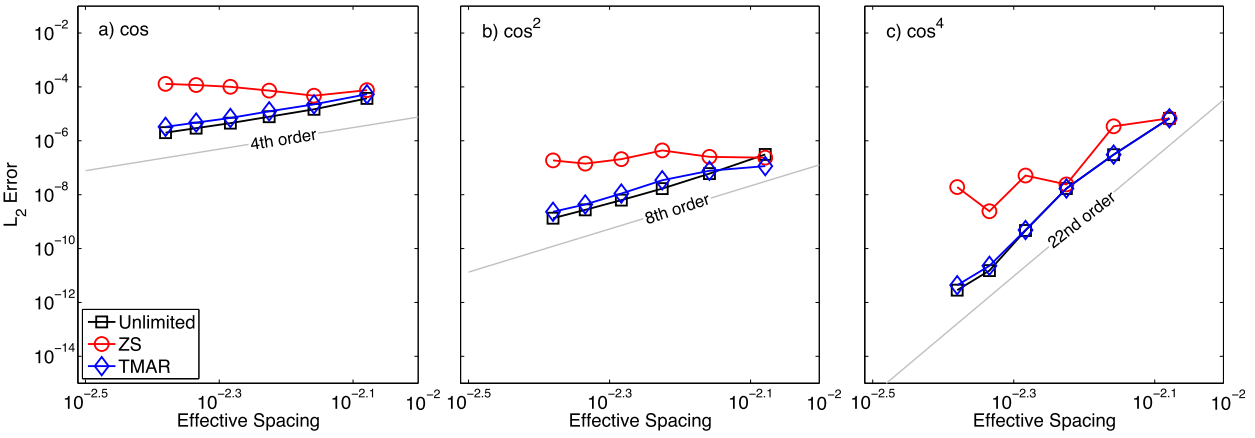

FIG. 3. Convergence results under p refinement: log–log plot of the L_2 error as a function of the effective spacing Δx_e for (a) C^1, (b) C^3, and (c) C^7 cosine bell tests. Each test uses a nodal basis with a 32 element mesh and a time step of $\Delta t = 0.5\Delta x^{(N+1)/3}$.

the magnitudes of the undershoots generated in the unlimited solution in the C^7 case, which tends to reduce the differences between the unlimited and both limited solutions. In summary, the TMAR limiter does a good job of maintaining the original accuracy of the unlimited solution under p refinement.

b. *Two-dimensional tests*

We now examine the performance of the TMAR limiter applied both to fully two-dimensional DG methods and to dimensionally split DG schemes, which advance the solution through pairs of one-dimensional integrations (Strang 1968). Table 1 lists numerical values for the maximum permissible one-dimensional Courant numbers for polynomial degrees between 2 and 5 when using SSPRK3 time stepping. The maximum permissible time step taken in the unlimited and TMAR methods is determined based on the maximum possible Courant–Friedrichs–Lewy (CFL) number for stability. An unsplit unlimited method or TMAR-limited method has a maximum stable Δt that is smaller by a factor of $2^{-1/2}$ than the values listed in Table 1. For ZS-limited methods, the time step condition given by (10) for split methods or (11) for unsplit methods must be satisfied to guarantee element-mean nonnegativity. Adopting a strategy similar to that in many practical applications, our solutions are obtained using a time step that brings the maximum local Courant number throughout the integration to approximately 95% of its limiting value.

In multidimensional problems, complex velocity fields can stretch and deform initially smooth data into thin filaments creating poorly resolved concentration gradients regardless of how well the initial conditions are resolved. This behavior occurs in our test case, in which a sheared swirling flow deforms an initially circular tracer distribution into a narrow coil before reversing and returning the tracer to its original shape. This test was originally described in LeVeque (1996). The velocity field is periodic over a time interval $0 \le t \le T = 5$ and is defined by the streamfunction:

$$\Psi(x, y, t) = \frac{1}{\pi}\sin(\pi x)^2 \sin(\pi y)^2 \cos\left(\frac{\pi t}{T}\right), \qquad (24)$$

and the relations:

$$u(x, y, t) = \frac{\partial \Psi}{\partial y}, \quad v(x, y, t) = -\frac{\partial \Psi}{\partial x}. \qquad (25)$$

The initial data $\psi_0(x, y)$ for this test are a C^3 cosine bell centered at the point $(x_0, y_0) = (1/4, 1/4)$ given by (23), with $q = 2$ and $\tau(x)$ replaced with $\tau(x, y)$ defined by

$$\tau(x, y) = \frac{1}{r_0}[(x - x_0)^2 + (y - y_0)^2]^{1/2}, \qquad (26)$$

where the initial radius is $r_0 = 1/4$. Tests involving the same flow and initial condition were considered in Ullrich and Norman (2014). Figure 4a shows the exact solution at $t = 0$ and T, while Fig. 4b shows a reference

TABLE 1. Maximum permissible Courant numbers $\mu_{\max}$ for DG methods of polynomial degree 2–5 schemes with SSPRK3 time stepping. The ZS values are from Zhang and Shu (2010); modal and nodal values are from Ullrich (2014).

	$\mu_{\max}$		
Degree	Modal	Nodal	ZS
2	0.210	0.450	0.167
3	0.130	0.255	0.167
4	0.090	0.168	0.083
5	0.067	0.120	0.083

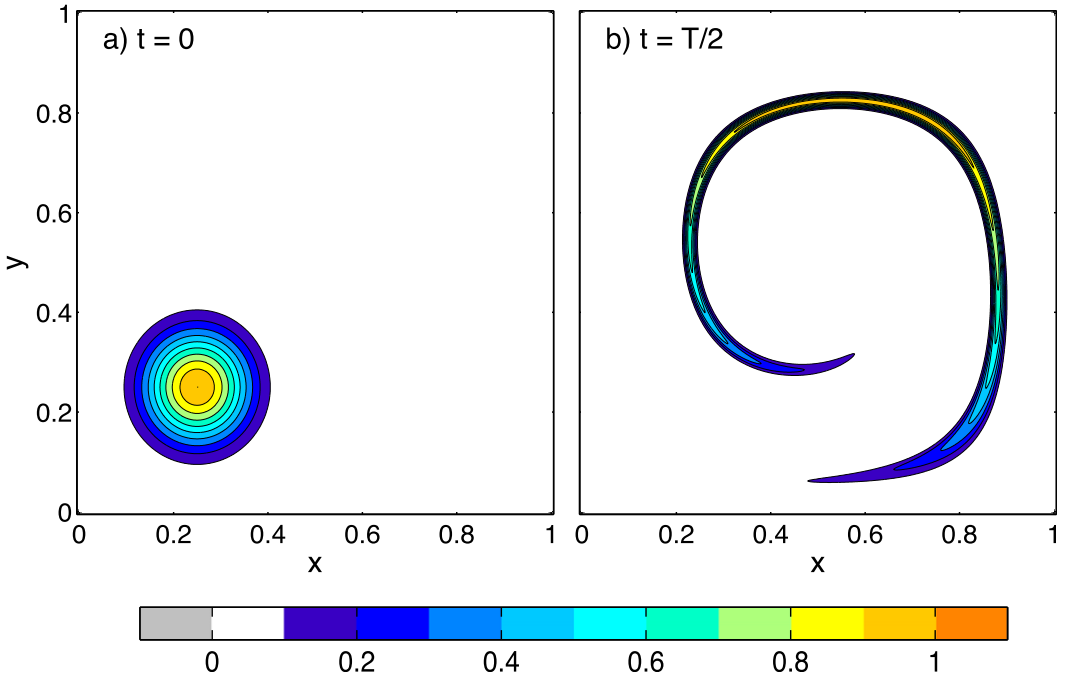

FIG. 4. Exact solution for the C^3 cosine bell tracer concentration field in the reversing deformation flow in (25) at times (a) 0 and T, and (b) converged solution for time $T/2$. Contours are at intervals of 0.1.

solution computed using very high time and space resolution[1] at the time of maximum deformation $t = T/2$.

Figure 5 shows numerical results at $t = T$ for nine different combinations of limiter and DG approximation strategies. Each DG approximation uses a 24×24 element grid with $N = 4$, for a total of 120 degrees of freedom (DOFs) along each coordinate. The results in the top row were obtained using a fully two-dimensional nodal DG implementation, while those in the bottom two rows were obtained using dimensional splitting with nodal (second row) and modal (third row) basis functions. The columns show the unlimited (left), ZS-limited (middle), and TMAR-limited (right) solutions. The maximum and minimum values of ϕ are listed in each panel, as well as the error measured from the analytic solution in the L^2 and L^∞ grid norms (labeled as E_2 and E_∞, respectively).

All three of the unlimited approximations do a fair job of maintaining the original amplitude of the exact solution, although as expected, they also generate spurious negative concentrations with magnitudes ranging up to 7% of the initial amplitude of the bell. These negatives are completely removed in each of the limited solutions. Relative to the unlimited solutions, the TMAR limited nodal solutions (Figs. 5c,f) see a 5%–7% decrease in their global maximum and a slight increase in both error norms. TMAR limiting has an even smaller impact on the accuracy of the modal solution, reducing its maximum amplitude by less than 1%. The superiority of both the unlimited and TMAR-limited modal solutions over their nodal equivalents arises from their use of the exact mass matrix (7). While the TMAR-limited solutions appear qualitatively similar to their unlimited counterparts with the negatives removed, all of the ZS limited solutions are significantly degraded. The ZS limiter produces substantial distortion of the originally symmetric tracer field, a 12%–25% reduction in the global maximum, and a large increase to both error norms.

The performance of the ZS limiter can be improved considerably by doubling the number of elements along each coordinate to 48, as shown in the middle column of Fig. 6. Nevertheless, as also shown in Fig. 6, the TMAR-limited solutions remain superior at this higher resolution; they look nearly identical to the unlimited solutions with the negatives removed and suffer very little degradation in the maximum amplitude. In particular, the TMAR-limited modal solution (Fig. 6i) is essentially identical to the nonnegative part of its unlimited counterpart (Fig. 6g).

We now consider the influence of the limiters on solutions to the same swirling-flow problem under p refinement with the total DOF held constant. Figure 7

[1] The reference solution was computed using $N = 5, \Delta x = 0.00625$, and $\Delta t = 5.182 \times 10^{-4}$.

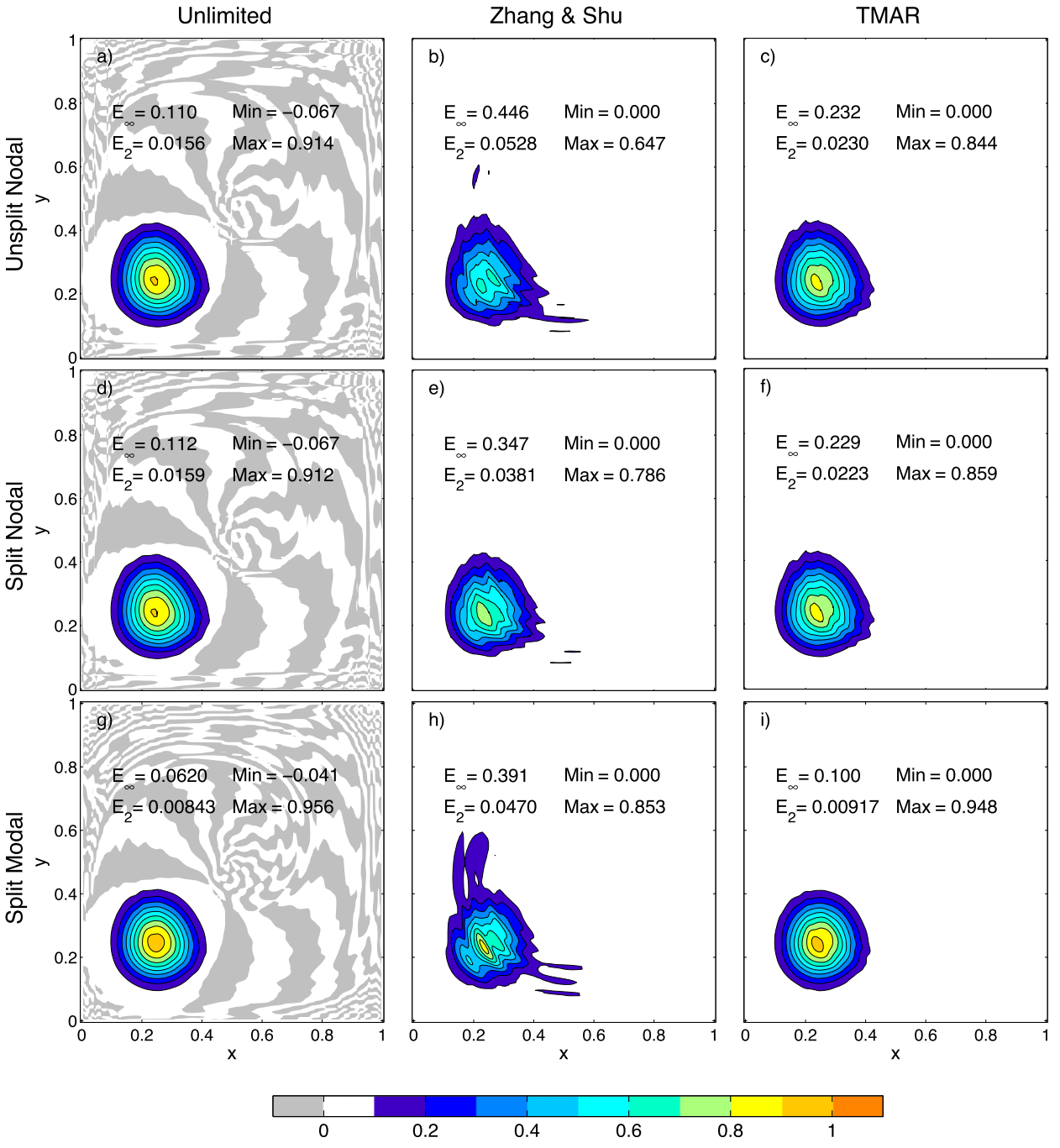

FIG. 5. Comparison of tracer concentration fields at $t = T$ using (left) no limiting, (middle) ZS limiting, or (right) TMAR limiting. (a)–(c) Unsplit nodal solutions, (d)–(f) split nodal solutions, and (g)–(i) split modal solutions. The domain is partitioned into a 24×24 element grid. Error norms and the domain maximum and minimum concentrations are noted in each plot. Contours are plotted every 0.1, and regions of negative concentration are highlighted in light gray.

shows results obtained using the unsplit nodal scheme as the polynomial degree is increased while the number of elements is reduced to keep the total DOF constant. The solutions in the top row were computed using a 30×30 element grid with local polynomial order $N = 3$. Thus, along each coordinate there are 4 DOF per element and 120 DOF across the full domain. The second row shows solutions computed using a 24×24 grid with $N = 4$, while those in the last row were generated on a 20×20 element grid with $N = 5$.

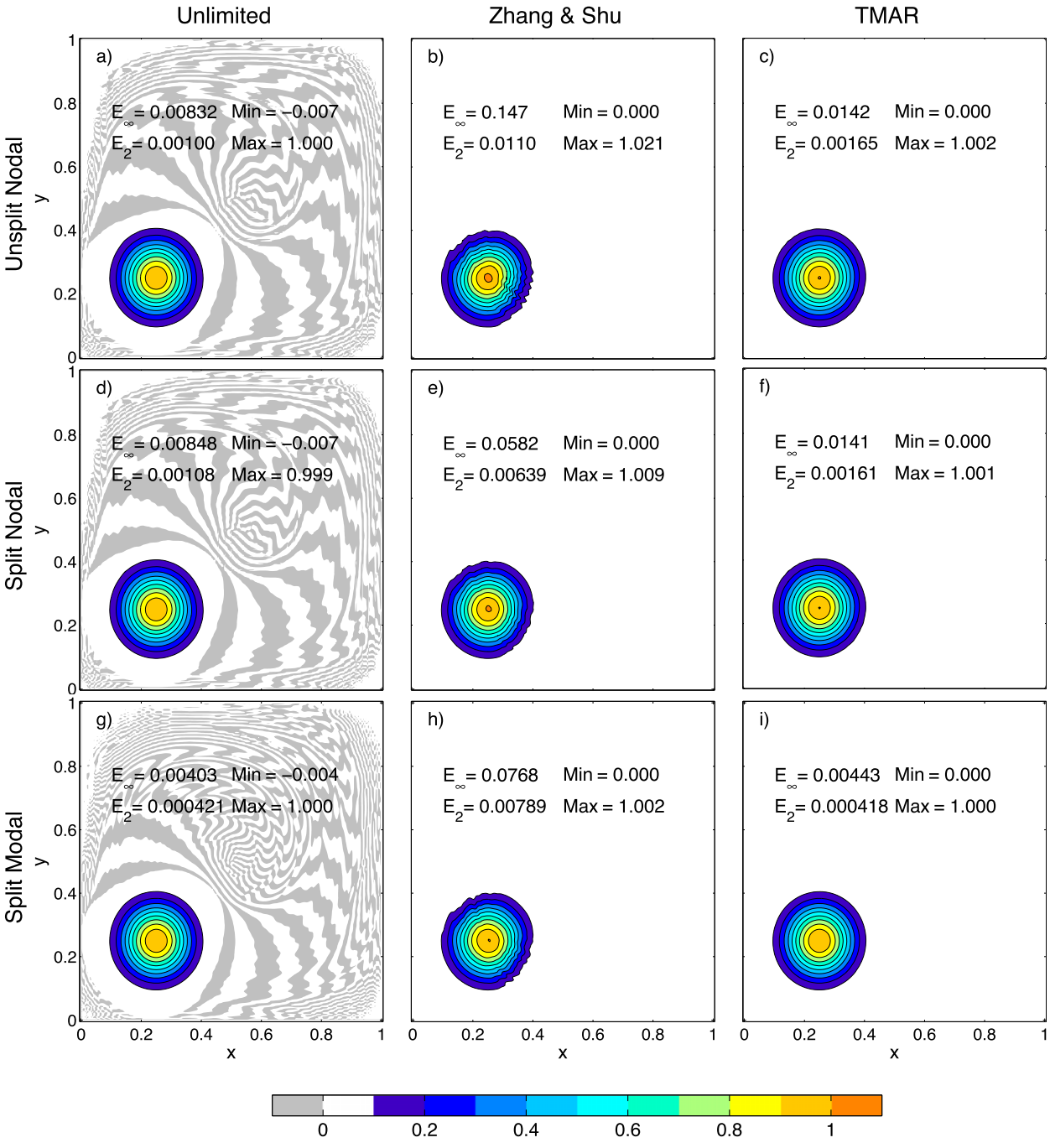

FIG. 6. As in Fig. 5, but using a grid of 48×48 elements.

In the unlimited case (left column of Fig. 7), using higher-degree polynomials reduces the error in both norms and the magnitudes of the negative undershoots. Higher degree also helps maintain maximum tracer concentration and the circular symmetry of the solution at $t = T$. Qualitatively similar improvements are seen as N is increased in the TMAR-limited solution (right column of Fig. 7), although as would be expected from Fig. 5, the error norms and the reduction in the maximum concentration in the TMAR-limited solutions exceed those obtained without limiting. In contrast, the quality of the ZS-limited solutions (middle column of Fig. 7) degrades as N is increased, a behavior consistent with that shown for our one-dimensional tests in Fig. 3.

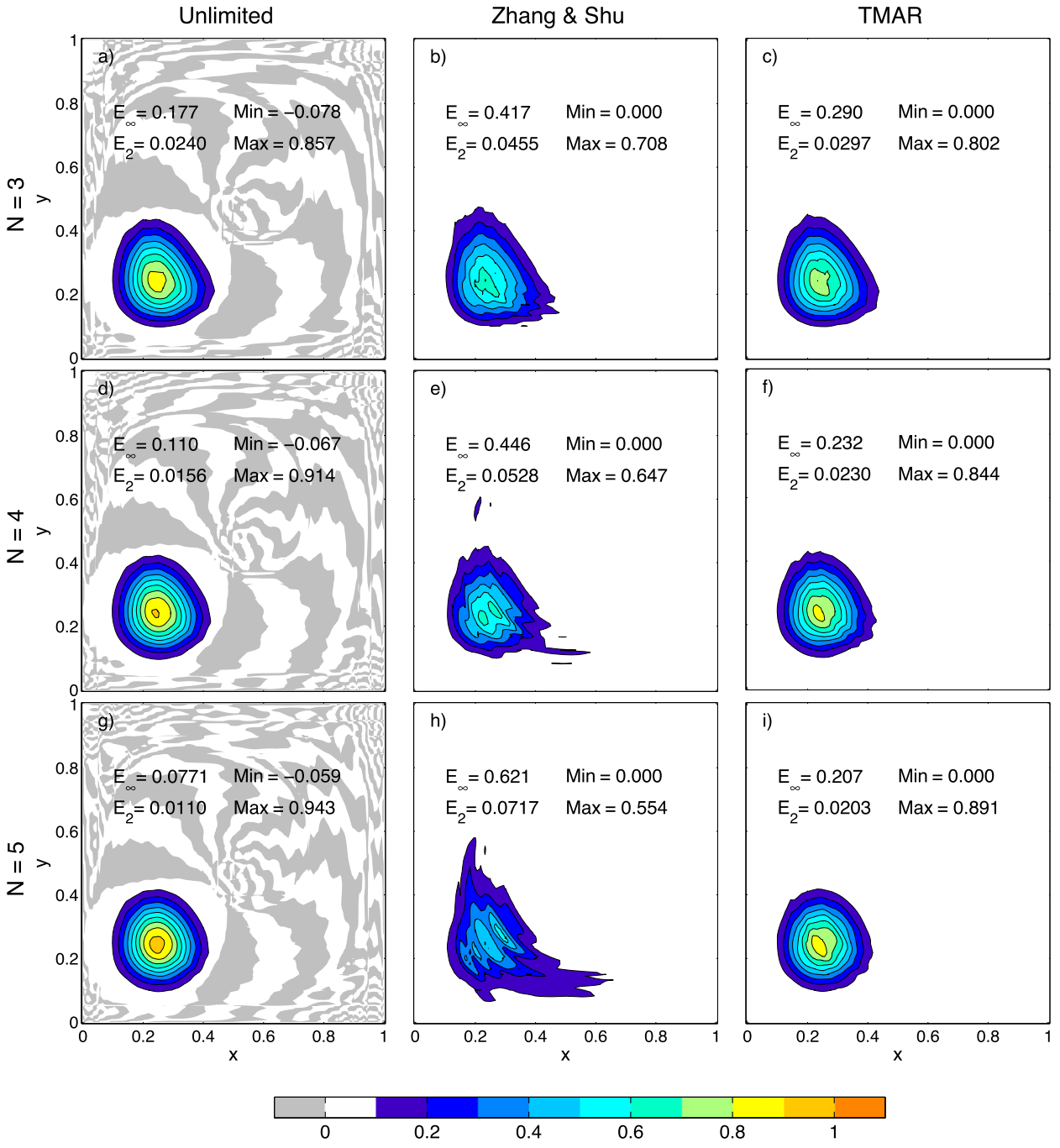

FIG. 7. Impact of polynomial refinement on tracer concentration fields at $t = T$. Each panel uses a total of 120 DOF along each coordinate. The limiting method is labeled at the top of each column and the degree of polynomial truncation is labeled to the left of each row. Error norms and the domain maximum and minimum concentrations are noted in each plot. Contours are plotted every 0.1, and negative regions are highlighted in light gray.

The difference in performance between the TMAR and ZS limiters in these tests can be largely explained by comparing their impact on the values at the nodes as illustrated in Fig. 1. Both limiters involve a linear rescaling applied to nonnegative nodal values, but that rescaling, as well as the modifications made to the negative nodal values, are different. When the ZS limiter is active, the local polynomial is modified so that the largest negative undershoot is scaled to zero while smaller undershoots take positive values. In contrast, when the TMAR limiter

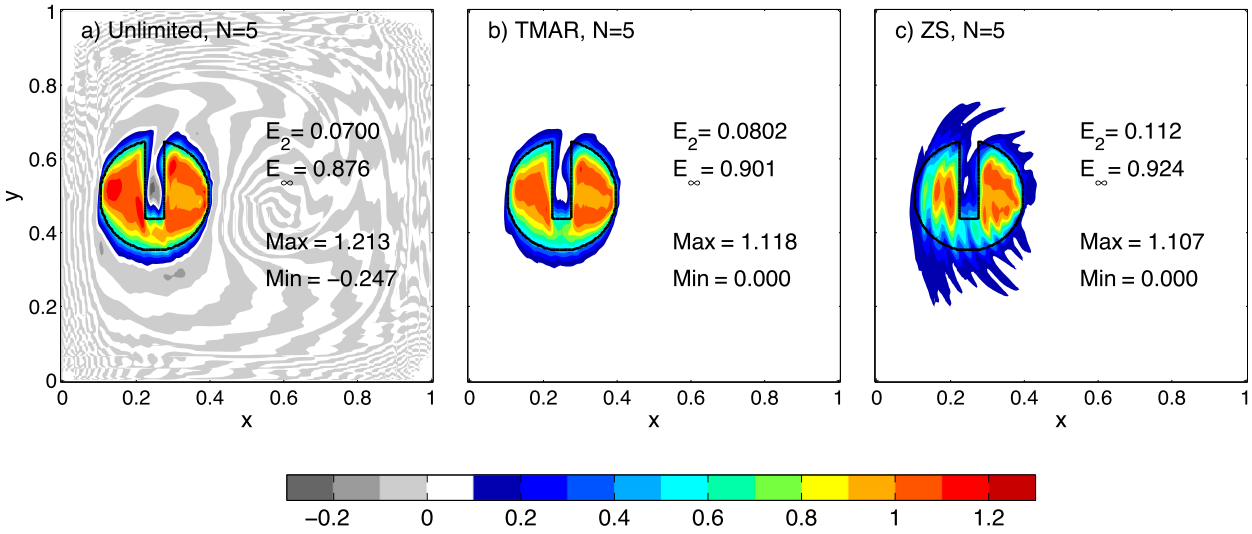

FIG. 8. Comparison of (a) unlimited, (b) TMAR-limited, and (c) ZS-limited solutions tracer concentration for the slotted cylinder at $t = T$. Each panel uses an unsplit nodal basis with a 32×32 element mesh. Error norms and the domain maximum and minimum concentrations are noted in each plot. Contours are plotted every 0.1, and negative regions are highlighted in light gray. The exact solution is outlined by the heavy black line.

is active, all negative undershoots are truncated to zero, giving a smaller modification for all but the minimum nodal value. By minimizing the magnitude of the adjustment at the nodes with negative values, the TMAR limiter minimizes the amount by which the positive part of the solution must be damped to conserve mass. Also in contrast to the ZS limiter, if the undershoot is at a node with a small weighted contribution to the element-averaged mass, the TMAR adjustment at that node has a correspondingly small effect on the total adjustment required for mass conservation. The mass-weighted adjustment used in the TMAR approach can be particularly important when using basis functions of high polynomial degree, where most of the nodes are clustered near the edges of each element and make relatively small contributions to the mean mass.

As a final test, we examine the behavior of these limiters in a case with large discontinuities by replacing the initial tracer field used in the previous tests with a unit-amplitude slotted cylinder of radius 0.15 centered at $(x_0, y_0) = (0.25, 0.5)$. The slot, in which the tracer density is zero, includes all points within the cylinder for which $|x - x_0| < 0.025$ and $y > y_0 + 0.0625$. Figure 8 compares unsplit nodal solutions at time T obtained using no limiting, TMAR limiting, and ZS limiting with $N = 5$ and 32 elements (for a total of 192 DOF) along each axis. Because of the discontinuous initial data, the maximum magnitude of overshoots and undershoots in the unlimited solution now exceeds 20% of the initial height of the cylinder. The TMAR solution removes the negatives and reduces the overshoot to about 12% without significantly increasing the E_2 and E_∞ errors. The worst solution with the largest error norms is generated using the ZS limiter, which produces short-wavelength noise that significantly distorts the solution.

5. Computational efficiency

We now turn to comparing the computational efficiency of the schemes used to obtain the solutions in the preceding two-dimensional deformation tests. The efficiency of a given scheme is a function of the maximum time step allowed by that scheme, the work required per time step, and the accuracy of the result at a given spatial and temporal resolution. The work per time step is both machine dependent and influenced by the efficiency of the code written for its implementation. All of our results are obtained using the same computing resource, and we have endeavored to impose a similar degree of optimization in the FORTRAN codes used to implement these methods. Despite these efforts, the following results are subject to the caveat that they are still somewhat machine and implementation specific.

Both the ZS and TMAR limiters require calculations in connection with each substep of the SSPRK scheme as well as a final adjustment to eliminate negatives after the full SSPRK update. The work per time step required in the final update, via (12) for the ZS method or (22) for the TMAR scheme, is similar. Somewhat more work is, however, required during each individual substep by the FCT flux correction for the TMAR method than for the most efficient ZS rescaling algorithm that uses the Gauss

TABLE 2. Work per time step and total time required to integrate deformation test for ZS and TMAR limiters applied to unsplit DG using $N = 4$ and an 192×192 element grid. Entries are normalized by the values for the unlimited method.

Method	CPU time per step	Total time
Unlimited	1.00	1.00
ZS	1.22	3.68
TMAR	1.34	1.34

points along each element boundary to implicitly evaluate the solution at one additional interior point. Table 2 compares the average CPU time spent for a single time step for the ZS and TMAR methods in the C^3-cosine-bell deformation-flow test using $N = 4$ and 192×192 elements. The values in Table 2 are normalized by the time required for a single step of the unlimited scheme.[2] On a per-time-step basis, the ZS limiter requires 22% more computation time than the unlimited scheme, while the TMAR method requires 34% more time.

These results for the work per time step do not take into account important differences in the maximum permissible time step for each method. The time step for the ZS scheme must satisfy the bound (11) in order to guarantee nonnegativity (roughly $\mu_{\max} = 0.04$ in this case),[3] while the TMAR limiter can use the maximum time step for which the unlimited method is stable (roughly $\mu_{\max} = 0.12$). Thus, as shown in the last column of Table 2, the ZS method takes about 3 times as long to complete the integration as does the TMAR-limited scheme. This comparison could, nevertheless, be less favorable for the TMAR scheme in a massively parallel implementation because the FCT modification of the fluxes needs to be communicated among the various elements, whereas the unmodified ZS fluxes does not require such extra communication. If the modest additional parallel communication required by the FCT limiter proves to be too inefficient, the FCT step in our limiter can be avoided by limiting the time step to the same value used in the standard ZS scheme.[4]

[2] Our benchmark unlimited integrations were also performed using the three-stage third-order SSPRK scheme, allowing an easy assessment of the computational overhead introduced by the nonnegativity preserving algorithms. Nevertheless, if limiting were not required, the SSPRK integrator could be replaced by a variety of classical three-stage third-order Runge–Kutta schemes for which the maximum stable time step increases by a factor of 1.73.

[3] In a series of empirical tests, we found small negatives developed when the ZS limiter was used in this test case with time steps for which $\mu_{\max}$ exceeded about 0.065.

[4] When element-mean nonnegative mass is ensured by limiting the time step, rather than the fluxes, TMAR limiting must be performed after each stage of an SSPRK integration (like the ZS rescaling).

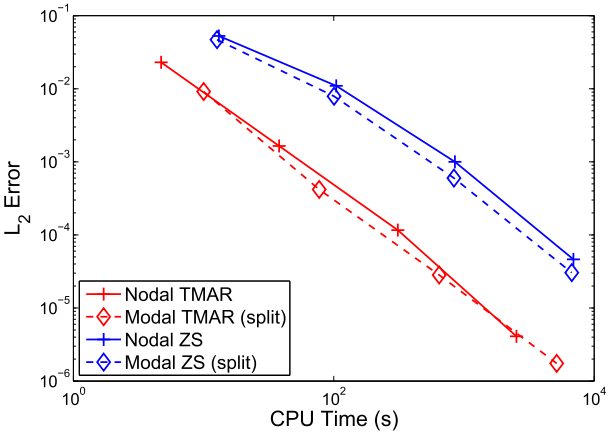

FIG. 9. The L_2 norm of error as a function of computational time spent to integrate the C^3 cosine bell deformation test using unsplit nodal and split modal methods in combination with TMAR or ZS limiting. Data points are shown for these simulations, which use 24, 48, 96, or 192 elements along each coordinate.

Another key measure of efficiency is the time required to obtain a solution of a desired accuracy. Figure 9 shows the L_2 error plotted as a function of the CPU time required to obtain unsplit-nodal and split-modal solutions to the C^3-cosine-bell deformation-flow test at $t = T$. All computations were performed with fourth-degree polynomials and 24, 48, 96, or 192 elements along each coordinate axis. (The concentration fields for the 24- and 48-element cases were plotted in Figs. 5 and 6, respectively.) At any particular CPU time, the TMAR-limited solution clearly gives a more accurate result, with an L_2 error roughly an order of magnitude smaller than that generated by the ZS-limited solution. Note also is that for both limiters, the split modal methods are more accurate than the corresponding unsplit nodal results.

6. Conclusions

We have introduced a nonnegativity preserving limiter for discontinuous Galerkin approximations to multidimensional scalar transport problems with nonnegative initial data. The proposed TMAR limiter truncates negative nodal values while simultaneously applying a mass aware rescaling to the remaining positive nodal values. This approach requires the element-mean concentration field to be positive prior to the TMAR adjustment, and this was ensured by an FCT-style correction to the standard DG fluxes computed at the element boundary in each forward substep of the SSPRK time integration. The limiter is also suitable for modal DG implementations in which the nonnegativity of the solution is maintained over subelement volumes of uniform size.

TMAR limiting maintains the geometric flexibility, compact formulation, and *h–p* adaptivity of the original DG formulation. It also maintains the order of accuracy of the underlying unlimited scheme. Our tests show that it can perform better than the well-known limiter due to Zhang and Shu (2010) in two-dimensional problems in which localized initial tracer fields deform in a swirling flow. TMAR limiting differs from the approach in Zhang and Shu (2010) because it does not overcorrect some negatives into positive values and because spurious undershoots at nodes near element boundaries, which have relatively less influence on the element-mean mass, exert only a correspondingly minor influence on the rescaling. The advantage of the TMAR limiter is particularly pronounced when using DG polynomials of modestly high degree (fifth or greater) and relatively coarse spatial resolution. In our tests, the TMAR limiter was also able to obtain solutions with L_2 errors an order of magnitude smaller than those computed over the same CPU time using the ZS limiter.

TMAR limiting may be easily extended to spectral element methods in which the basic functions are tensor products of Lagrange polynomials interpolating the GLL nodes. As shown by Guba et al. (2014), a restriction on the time step similar to that used with the ZS limiter will ensure that the element-mean mass remains nonnegative after each forward step of an SSPRK integration. TMAR limiting may then be performed after each forward step to preserve nonnegativity within each element. The performance of the TMAR limiter in spectral-element discretizations on the cubed sphere, and the extent to which the TMAR approach preserves tracer correlations, will be examined in a future publication.

Acknowledgments. The authors greatly benefited from discussions with Frank Giraldo, Peter Blossey, and comments from two anonymous reviewers. This research was supported by National Science Foundation Grant DMS-1216576.

APPENDIX

Proof of Theorem 1

Theorem 1. Let $\phi_{s_i}(x)$ be the unmodified Mth-order DG approximation to $\psi(x)$ in s_i at some arbitrary time, subject to the constraint that $\overline{\phi}_{s_i} \geq 0$. Then the TMAR limited solution $\phi^*_{s_i}(x)$ is also an Mth-order approximation to $\psi(x)$.

Proof. Because $\phi_{s_i}(x)$ is an Mth-order approximation to $\psi(x)$ it follows that

$$\max_{x\in s_i} |\psi(x) - \phi_{s_i}(x)| = \mathcal{O}(\Delta x^M), \tag{A1}$$

so it suffices to show that the TMAR modification is small in the sense that

$$\max_{x\in s_i} |\phi^*_{s_i}(x) - \phi_{s_i}(x)| = \mathcal{O}(\Delta x^M). \tag{A2}$$

Since the coefficients in the Lagrange polynomial representation of $\phi^*_{s_i}(x)$ are given by the modified nodal values in (16):

$$\max_{x\in s_i} |\phi^*_{s_i}(x) - \phi_{s_i}(x)| = \max_{x\in s_i} \left| \sum_{k=0}^{N} [\phi^*_{s_i}(x_k) - \phi_{s_i}(x_k)]\varphi_k(x) \right|, \tag{A3}$$

$$\leq C_N \sum_{k=0}^{N} |\phi^*_{s_i}(x_k) - \phi_{s_i}(x_k)|, \tag{A4}$$

where C_N is a constant that depends only on N. Thus we need to show that $\sum_{k=0}^{N} |\phi^*_{s_i}(x_k) - \phi_{s_i}(x_k)| = \mathcal{O}(\Delta x^M)$.

There are two cases to consider. First suppose that $\phi_{s_i}(x_k) \leq 0$, then $\phi^*_{s_i}(x_k) = 0$, so $\phi_{s_i}(x_k) \leq \phi^*_{s_i}(x_k) \leq \psi(x_k)$. Since $\phi_{s_i}(x)$ is an approximation to ψ with error $\mathcal{O}(\Delta x^M)$, it follows that

$$|\phi^*_{s_i}(x_k) - \phi_{s_i}(x_k)| \leq |\psi(x_k) - \phi_{s_i}(x_k)| = \mathcal{O}(\Delta x^M). \tag{A5}$$

On the other hand suppose that $\phi_{s_i}(x_k) > 0$, then $\phi^*_{s_i}(x_k) = r_i \phi_{s_i}(x_k)$ where r is given in (15). Noting that $r_i \leq 1$:

$$|\phi^*_{s_i}(x_k) - \phi_{s_i}(x_k)| = (1 - r_i)\phi_{s_i}(x_k). \tag{A6}$$

Defining

$$\phi^+_{s_i}(x_l) = \begin{cases} \phi_{s_i}(x_l) & \text{if} \quad \phi_{s_i}(x_l) \geq 0 \\ 0 & \text{if} \quad \phi_{s_i}(x_l) < 0 \end{cases} \quad \text{and}$$

$$\phi^-_{s_i}(x_l) = \begin{cases} 0 & \text{if} \quad \phi_{s_i}(x_l) \geq 0 \\ \phi_{s_i}(x_l) & \text{if} \quad \phi_{s_i}(x_l) < 0 \end{cases}, \tag{A7}$$

and using (15), the coefficient $(1 - r_i)$ may be rewritten as

$$1 - r_i = \frac{\sum_{l=0}^{N} w_l |\phi^-_{s_i}(x_l)|}{\sum_{l=0}^{N} w_l |\phi^+_{s_i}(x_l)|}. \tag{A8}$$

Substituting (A8) into (A6) and using the inequality $\sum_{l=0}^{N} w_l |\phi^+_{s_i}(x_l)| \geq w_k \phi_{s_i}(x_k)$ to bound the denominator in (A8) from below gives

$$|\phi^*_{s_i}(x_k)-\phi_{s_i}(x_k)| \le \frac{1}{w_k}\sum_{l=0}^{N} w_l|\phi^-_{s_i}(x_l)|. \quad \text{(A9)}$$

From (A5) it follows that $|\phi^-_{s_i}(x_l)| = \mathcal{O}(\Delta x^M)$ for all $l = 0, \ldots, N$. Thus,

$$|\phi^*_{s_i}(x_k)-\phi_{s_i}(x_k)| \le D_N \Delta x^M = \mathcal{O}(\Delta x^M), \quad \text{(A10)}$$

where D_N is a constant which depends only on N. Therefore, $|\phi^*_{s_i}(x_k)-\phi_{s_i}(x_k)| = \mathcal{O}(\Delta x^M)$ for all k, and the TMAR limiting maintains Mth-order accuracy with respect to h refinement.

REFERENCES

Cockburn, B., and C.-W. Shu, 1989: TVB Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws. II. General framework. *Math. Comput.*, **52** (186), 411–435.

——, S.-Y. Lin, and C.-W. Shu, 1989: TVB Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws III: One-dimensional systems. *J. Comput. Phys.*, **84**, 90–113, doi:10.1016/0021-9991(89)90183-6.

Dumbser, M., O. Zanotti, R. Loubère, and S. Diot, 2014: A posteriori subcell limiting of the discontinuous Galerkin finite element method for hyperbolic conservation laws. *J. Comput. Phys.*, **278**, 47–75, doi:10.1016/j.jcp.2014.08.009.

Durran, D. R., 2010: *Numerical Methods for Fluid Dynamics: With Applications to Geophysics.* 2nd ed. Texts in Applied Mathematics Series, Vol. 32, Springer-Verlag, 516 pp., doi:10.1007/978-1-4419-6412-0.

Giraldo, F. X., J. S. Hesthaven, and T. Warburton, 2002: Nodal high-order discontinuous Galerkin methods for the spherical shallow water equations. *J. Comput. Phys.*, **181**, 499–525, doi:10.1006/jcph.2002.7139.

Gottlieb, S., D. I. Ketcheson, and C.-W. Shu, 2009: High order strong stability preserving time discretizations. *J. Sci. Comput.*, **38**, 251–289, doi:10.1007/s10915-008-9239-z.

Guba, O., M. Taylor, and A. St-Cyr, 2014: Optimization-based limiters for the spectral element method. *J. Comput. Phys.*, **267**, 176–195, doi:10.1016/j.jcp.2014.02.029.

Guo, W., R. D. Nair, and J.-M. Qiu, 2014: A conservative semi-Lagrangian discontinuous Galerkin scheme on the cubed sphere. *Mon. Wea. Rev.*, **142**, 457–475, doi:10.1175/MWR-D-13-00048.1.

Hartmann, R., and P. Houston, 2002: Adaptive discontinuous Galerkin finite element methods for the compressible Euler equations. *J. Comput. Phys.*, **183**, 508–532, doi:10.1006/jcph.2002.7206.

Hesthaven, J. S., and T. Warburton, 2008: *Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications.* Texts in Applied Mathematics, Vol. 54, Springer, 494 pp., doi:10.1007/978-0-387-72067-8.

Karniadakis, G., and S. Sherwin, 2005: *Spectral/hp Element Methods for Computational Fluid Dynamics*. Oxford University Press, 656 pp.

Lauritzen, P., A. Conley, J.-F. Lamarque, F. Vitt, and M. Taylor, 2015: The terminator ''toy'' chemistry test: A simple tool to assess errors in transport schemes. *Geosci. Model Dev.*, **8**, 1299–1313, doi:10.5194/gmd-8-1299-2015.

LeVeque, R., 1996: High-resolution conservative algorithms for advection in incompressible flow. *SIAM J. Numer. Anal.*, **33**, 627–665, doi:10.1137/0733033.

Liu, X.-D., and S. Osher, 1996: Nonoscillatory high order accurate self-similar maximum principle satisfying shock capturing schemes I. *SIAM J. Numer. Anal.*, **33**, 760–779, doi:10.1137/0733038.

Persson, P.-O., and J. Peraire, 2006: Sub-cell shock capturing for discontinuous Galerkin methods. *44th AIAA Aerospace Sciences Meeting and Exhibit*, Reno, NV, Amer. Institute of Aeronautics and Astronautics, AIAA 2006-112, doi:10.2514/6.2006-112.

Qiu, J., and C.-W. Shu, 2004: Hermite WENO schemes and their application as limiters for Runge–Kutta discontinuous Galerkin method: One-dimensional case. *J. Comput. Phys.*, **193**, 115–135, doi:10.1016/j.jcp.2003.07.026.

——, and ——, 2005a: Hermite WENO schemes and their application as limiters for Runge–Kutta discontinuous Galerkin method II: Two dimensional case. *Comput. Fluids*, **34**, 642–663, doi:10.1016/j.compfluid.2004.05.005.

——, and ——, 2005b: Runge–Kutta discontinuous Galerkin method using WENO limiters. *SIAM J. Sci. Comput.*, **26**, 907–929, doi:10.1137/S1064827503425298.

Qiu, J.-M., and C.-W. Shu, 2011: Positivity preserving semi-Lagrangian discontinuous Galerkin formulation: Theoretical analysis and application to the Vlasov–Poisson system. *J. Comput. Phys.*, **230**, 8386–8409, doi:10.1016/j.jcp.2011.07.018.

Ramachandran, D., M. Levy, and P. Lauritzen, 2011: Emerging numerical methods for atmospheric modeling. *Numerical Techniques for Global Atmospheric Models*, P. Lauritzen et al., Eds., Lecture Notes in Computational Science and Engineering, Vol. 80, Springer, 251–311, doi:10.1007/978-3-642-11640-7_9.

Restelli, M., L. Bonaventura, and R. Sacco, 2006: A semi-Lagrangian discontinuous Galerkin method for scalar advection by incompressible flows. *J. Comput. Phys.*, **216**, 195–215, doi:10.1016/j.jcp.2005.11.030.

Rossmanith, J., and D. Seal, 2011: A positivity-preserving high-order semi-Lagrangian discontinuous Galerkin scheme for the Vlasov–Poisson equations. *J. Comput. Phys.*, **230**, 6203–6232, doi:10.1016/j.jcp.2011.04.018.

Smolarkiewicz, P. K., 1989: Comments on ''A positive definite advection scheme obtained by nonlinear renormalization of the advective fluxes.'' *Mon. Wea. Rev.*, **117**, 2626–2632, doi:10.1175/1520-0493(1989)117<2626:COPDAS>2.0.CO;2.

Strang, G., 1968: On the construction and comparison of difference schemes. *SIAM J. Numer. Anal.*, **5**, 506–517, doi:10.1137/0705041.

Ullrich, P., 2014: Understanding the treatment of waves in atmospheric models. Part 1: The shortest resolved waves of the 1D linearized shallow water equations. *Quart. J. Roy. Meteor. Soc.*, **140**, 1426–1440, doi:10.1002/qj.2226.

——, and M. Norman, 2014: The flux-form semi-Lagrangian spectral element (FF-SLSE) method for tracer transport. *Quart. J. Roy. Meteor. Soc.*, **140**, 1069–1085, doi:10.1002/qj.2184.

Zalesak, S. T., 1979: Fully multidimensional flux-corrected transport algorithms for fluids. *J. Comput. Phys.*, **31**, 335–362, doi:10.1016/0021-9991(79)90051-2.

Zhang, X., and C.-W. Shu, 2010: On maximum-principle-satisfying high order schemes for scalar conservation laws. *J. Comput. Phys.*, **229**, 3091–3120, doi:10.1016/j.jcp.2009.12.030.

——, and ——, 2011: Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: Survey and new developments. *Proc. Roy. Soc.*, **467A**, 2752–2776, doi:10.1098/rspa.2011.0153.