Journal of Computational Physics 475 (2023) 111879

Contents lists available at ScienceDirect

Journal of Computational Physics

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

Global and local conservation of mass, momentum and kinetic energy in the simulation of compressible flow

Gennaro Coppola [a], Arthur E.P. Veldman [b,*]

[a] Dipartimento di Ingegneria Industriale, Università di Napoli "Federico II", Napoli, Italy
[b] Bernoulli Institute, University of Groningen, Groningen, the Netherlands

ARTICLE INFO

Article history:
Received 13 February 2022
Received in revised form 21 October 2022
Accepted 21 December 2022
Available online 29 December 2022

Keywords:
Compressible flow
Finite difference
Finite volume
Local and global conservation
Primary and secondary invariants

ABSTRACT

The spatial discretization of convective terms in compressible flow equations is studied from an abstract viewpoint, for finite-difference methods and finite-volume type formulations with cell-centered numerical fluxes. General conditions are sought for the local and global conservation of primary (mass and momentum) and secondary (kinetic energy) invariants on Cartesian meshes. The analysis, based on a matrix approach, shows that sharp criteria for global and local conservation can be obtained and that in many cases these two concepts are equivalent. Explicit numerical fluxes are derived in all finite-difference formulations for which global conservation is guaranteed, even for non-uniform Cartesian meshes. The treatment reveals also an intimate relation between conservative finite-difference formulations and cell-centered finite-volume type approaches. This analogy suggests the design of wider classes of finite-difference discretizations locally preserving primary and secondary invariants.

1. Introduction

The design of accurate and reliable numerical methods for the simulation of turbulent compressible flows is a challenging and active research topic [1,2]. Owing to the possible coexistence of several physical phenomena (e.g. shock waves, acoustic waves, turbulent fluctuations) the discretization procedure should satisfy different, and in some cases contrasting, requirements. Convergence toward a weak solution, essential for a correct shock capturing, needs a locally conservative spatial discretization of advective terms, whereas acoustic waves require non dissipative and non dispersive schemes, usually of high order, e.g. [3]. Moreover, the accumulation of aliasing errors due to the spatial discretization of advective terms poses strong challenges to numerical schemes in terms of (in)stability of long-time integrations for turbulent flows. The current paper focuses on the influence of the discretized advective terms on the latter issue.

In incompressible flows this last issue is usually addressed by designing the numerical discretization in such a way that discrete advective terms do not contribute to the induced balance of kinetic energy, which in the limit of vanishing viscosity, is an invariant of the equations for incompressible flow [4–7]. The reproduction of this property at a discrete level furnishes an important nonlinear stability criterion, as well as a more faithful representation of energy transfer across the different scales of the flow field. In the case of inviscid compressible flows, global kinetic energy is not strictly conserved because of pressure-work contributions, but advection still does not spoil its global balance. As mentioned above, in the

* Corresponding author.
E-mail address: a.e.p.veldman@rug.nl (A.E.P. Veldman).

https://doi.org/10.1016/j.jcp.2022.111879

paper we will focus on the discrete treatment of these advective terms, and leave the (complementary) treatment of the pressure gradient and other thermodynamic terms for later (for basic ideas we refer to [8–10]). The adoption of numerical discretization procedures that do not perturb the balance of secondary invariants has shown great benefits in terms of stability and reliability of the simulations also in the case of compressible flows [1,11–13].

In the finite-difference community this target is usually achieved by using a so-called split form, in which the advective term is written as a linear combination of different expressions obtained by applying the product rule of derivatives to the divergence of the flux. Although analytically equivalent, the different forms of the advective derivatives behave differently when directly discretized. Selected energy-preserving forms were constructed by analogy with the incompressible skew-symmetric form [11,14]. Only recently a quite complete characterization of all the possible energy preserving forms for finite-difference central schemes has been derived [15] and an even more general criterion for (global) kinetic-energy preserving discretizations has been identified by [16]. In all cases the theory shows preservation of kinetic energy *globally*, which means that the total amount of kinetic energy is preserved by the discretization of the advective terms. The problem of the correct reproduction of local kinetic energy fluxes is usually not mentioned in purely finite-difference formulations.

The equations of motion can be discretized also within a finite-volume framework. Also in this case several energy-preserving formulations have been proposed [3,13,17]. In many cases, starting from the specification of fluxes for primary variables, the procedures allow also the identification of a kinetic energy flux, the global balance following by means of the telescoping property. However, a systematic procedure to generate a locally energy-preserving method starting from a locally conservative formulation for primary variables seems to be lacking.

Note that in this paper we will use the term 'finite-volume' in a quite liberal way to refer to discrete formulations for which numerical fluxes are explicitly specified. In particular, we will consider cell-centered finite-volume formulations, in which the grid nodes are located in the 'center' (liberally interpreted) of the control volumes [18]. We will not (yet) make a link with the cell-vertex finite-volume formulations [19,20] in which the grid nodes coincide with the corners of the control volumes. Neither will we dwell on the interpretation of the discrete time derivative as the evolution of either a nodal value or of a cell average. Over the years this interpretation has led to much confusion in the literature; e.g., see the discussion on the order of accuracy of the QUICK method [21,22].

Even more important, for flows involving shocks, is the problem of local preservation of primary invariants. Locally conservative formulations require that the discretization of advective derivatives can be expressed as the difference between numerical fluxes evaluated at adjacent nodes. This property is naturally reproduced in finite-volume methods, which are focused exactly on the specification of the numerical flux at each face. In finite-difference methods the discretization is obtained by approximating derivative operators and whether a particular discretization can be cast as difference of fluxes is not evident in many cases, especially on non-uniform meshes. Surprisingly enough, only recently this problem has been solved for the case of central explicit schemes on uniform meshes [14], for which numerical fluxes corresponding to general split forms have been identified.

In this paper, the problem of local and global conservation of primary and secondary invariants of transport equations is studied with reference to both the finite-difference and finite-volume methods on non-uniform, rectangular (Cartesian) meshes. It is found that sharp criteria can be found for global conservation in general conditions and that in many cases the concepts of local and global conservation are actually equivalent, for both linear and quadratic invariants. Explicit fluxes can be derived in all cases in which global conservation of these invariants is guaranteed (including non-uniform meshes). The treatment sheds also light on an intimate relation (almost an equivalence) between finite-difference and finite-volume approaches, in the sense that any globally conservative finite-difference formulation of the advective derivatives (of arbitrary order) can be expressed as a difference of numerical fluxes, and (almost) any formulation based on numerical fluxes built by using polynomial interpolations has a corresponding (generalized) finite-difference formulation. The complete equivalence is established, for the moment, for second-order formulations. This strict analogy suggests also the design of a wider class of finite-difference discretizations which locally preserve primary and secondary invariants. Our analysis is carried out for the semi-discretized equations, i.e. exact time integration is assumed in order not to interfere with the discrete conservation properties; for compressible flow this is not straightforward. For instance, Subbareddy and Candler [13] introduced special 'square-root' variables to construct an energy-preserving time-integration scheme. These variables can also be used to achieve supraconservative space discretization [2,23–25], but here we will not explicitly study this form of the equations, though our theory can be reformulated for this change of variables.

A final mention has to be made on the notation we use for the derivation of the presented results. The analysis is carried out by employing a matrix representation of discrete operators, which we found an ideal tool for the derivation of all the connections between the various concepts exposed. Matrix formalism is useful because many global conservation properties can be easily derived by studying the schemes globally, and the connections between global and local conservation can be obtained by using decomposition theorems. The reasoning in terms of matrices is independent of the discretization method with which the discrete equations are obtained (finite volumes/differences/elements/...), and allows to derive sharp, necessary and sufficient, conditions for the desired properties to hold. In this respect our approach essentially differs from, and generalizes, the more analytical studies found in the literature mentioned above.

It would be interesting to perform a similar analysis to other, analytically equivalent, forms of the equations (streamfunction-vorticity [26], rotational [27], velocity-vorticity [28], weak formulations [29], etc.), which are conveniently

related to other secondary invariants (enstrophy [26], helicity [30], angular momentum [31], etc.). Ultimately, this may lead to guide lines on how to choose between these formulations and invariants, depending on the application envisaged.

Outline of paper

After an overview of analytical formulations of the transport equations the matrix-vector notation is introduced (Section 2) (all technical proofs are gathered in an appendix). It is explained how conservation of linear invariants is equivalent to vanishing column sums of the relevant operators. Sections 3 and 4 study the conservation of the linear invariants mass and momentum, respectively. Conditions on the discretization are formulated and it is shown that a conservative finite-difference method can be rewritten as a (cell-centered) finite-volume method. Section 5 studies the conservation of the quadratic invariant energy. More conditions have to be satisfied, yet there remains a large freedom in choosing the mass flux. The close relation between conservative finite-difference and finite-volume methods is analyzed in more detail in Section 5.4. Higher-order discretization schemes are shortly discussed in Section 6. Several numerical experiments with the discretized transport equations are presented in Section 7, to illustrate our theoretical considerations. Thereafter, in Section 8, we make a first step toward the compressible Euler equations, accompanied by some preliminary numerical demonstrations. Finally, in Section 9 our findings are summarized.

2. Setting the scene

2.1. Conservation laws

Consider a conservation law for a quantity ϕ which is transported by a flow with mass density ρ and velocity $\boldsymbol{u}$:

$$\frac{\partial \rho}{\partial t} + \mathcal{M}_{\text{mass}} \equiv \frac{\partial \rho}{\partial t} + \nabla\cdot(\rho \boldsymbol{u}) = 0\,; \quad \frac{\partial \rho\phi}{\partial t} + \mathcal{C}_{\text{mom}}\phi \equiv \frac{\partial \rho\phi}{\partial t} + \nabla\cdot(\rho \boldsymbol{u}\phi) = 0\,. \tag{1a,b}$$

These transport equations are used as a first step towards the compressible Euler equations. The main message of the paper can be explained using this simplified set of equations. A preliminary exploration of the Euler equations will follow near the end of the paper; for further explorations we refer to a recent paper by De Michele and Coppola [10].

In Eqs. (1a,b) the rate of variation in time of the quantities $\rho\psi$ (ψ being 1 or ϕ), is driven by the divergence of a flux vector. Application of Gauss' divergence theorem shows that the integral of $\rho\psi$ over the entire domain Ω does not change in time in the case of periodic and/or homogeneous boundary conditions. In other words, the total amount of ρ and $\rho\phi$ inside the entire domain is preserved by advection (also termed convection, whence the notation $\mathcal{C}$), i.e. these quantities are *globally conserved*. From a mathematical point of view, the function $\mathcal{M}_{\text{mass}}$ and the operator $\mathcal{C}_{\text{mom}}$ satisfy the following constraints:

$$\int_\Omega \mathcal{M}_{\text{mass}}\,\mathrm{d}\Omega = 0; \quad \int_\Omega \mathcal{C}_{\text{mom}}\phi\,\mathrm{d}\Omega = 0 \quad \forall\phi. \tag{2a,b}$$

In fact, the structure of the operators at the right hand sides of Eqs. (1a,b) implies that the evolution of ρ and $\rho\phi$ integrated over an arbitrary domain Ω_h depends only on the flux at the boundary Γ_h:

$$\int_{\Omega_h} \frac{\partial \rho\psi}{\partial t}\,\mathrm{d}\Omega_h = -\int_{\Gamma_h} \rho\psi \boldsymbol{u}\cdot\boldsymbol{n}\,\mathrm{d}\Gamma_h \quad (\psi \in \{1,\phi\}). \tag{3}$$

Global conservation (2) follows as a particular case. We will express this circumstance by saying that ρ and $\rho\phi$ are also *locally conserved*. Owing to these properties, the quantities ρ and $\rho\phi$ are called *primary invariants.*

Analytically, the equations (1) admit even more invariants, such as the 'kinetic energy' $\frac{1}{2}\rho\phi^2$. By combining Eqs. (1a,b), an evolution equation for the energy can be obtained [16]:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \frac{1}{2}\rho\phi^2\,\mathrm{d}\Omega = -\int_\Omega \left(\phi\,\mathcal{C}_{\text{mom}}\phi - \frac{1}{2}\phi^2\mathcal{M}_{\text{mass}}\right)\mathrm{d}\Omega = -\int_\Omega \phi\mathcal{A}_{\text{kin}}\phi\,\mathrm{d}\Omega\,, \tag{4}$$

where the operator $\mathcal{A}_{\text{kin}}$ is defined as

$$\mathcal{A}_{\text{kin}} : \phi \mapsto \left(\mathcal{C}_{\text{mom}} - \frac{1}{2}\mathcal{M}_{\text{mass}}\right)\phi = \nabla\cdot(\rho\boldsymbol{u}\phi) - \frac{1}{2}(\nabla\cdot(\rho\boldsymbol{u}))\phi\,. \tag{5}$$

It follows that global energy conservation (left-hand side of Eq. (4) equal to zero) is equivalent to the skew symmetry of $\mathcal{A}_{\text{kin}}$ (right-hand side of Eq. (4) equal to zero). It is easily verified that analytically the operator $\mathcal{A}_{\text{kin}}$ from (5) indeed is skew-symmetric, hence the equations (1a,b) do globally conserve the secondary invariant energy, next to the primary invariants mass and momentum.

Note that also in this case even more than global conservation can be inferred. Since the product $\phi\mathcal{A}_{\text{kin}}\phi$ can be expressed as the divergence of the quantity $\rho\boldsymbol{u}\phi^2/2$, the equation for the generalized local kinetic energy has the divergence structure

$$\frac{\partial \rho\phi^2/2}{\partial t} = \phi\mathcal{A}_{\text{kin}}\phi = \nabla\cdot(\rho\boldsymbol{u}\phi^2/2) \equiv \mathcal{K}_{\text{kin}}(\phi), \tag{6}$$

where $\mathcal{K}_{\text{kin}}$ is a divergence-type operator as are occurring in Eq. (1). This implies that the evolution of $\rho\phi^2/2$ integrated over an arbitrary domain depends on the flux at its boundary, i.e. $\rho\phi^2/2$ is also *locally conserved*. We will express this circumstance by saying that the generalized kinetic energy $\rho\phi^2/2$ is a *secondary (quadratic) invariant*.

2.2. Various analytical formulations

The equations (1) can be written in other formulations, which analytically are equivalent, but where conservation is less obvious at first sight. We will study these formulations and their discretizations in the sequel. Coppola et al. [7,15] consider a large family of analytical formulations for Eq. (1), for which they study the conservation properties of central finite-difference discretizations on uniform grids. In the continuity equation (1a) they write the mass-transport term $\nabla\cdot(\rho\boldsymbol{u})$ in two ways as

$$\mathcal{M}^D_{\text{mass}} \equiv \nabla\cdot(\rho\boldsymbol{u}); \qquad \mathcal{M}^A_{\text{mass}} \equiv \boldsymbol{u}\cdot\nabla\rho + \rho\nabla\cdot\boldsymbol{u}. \tag{7a,b}$$

For the advective term $\nabla\cdot(\rho\boldsymbol{u}\phi)$ in Eq. (1b), they consider the following analytical formulations (see Eqs. (7-11) from [15]):

$$\begin{aligned}&\mathcal{C}^D_{\text{mom}}\phi \equiv \nabla\cdot(\rho\boldsymbol{u}\phi); \quad \mathcal{C}^\phi_{\text{mom}}\phi \equiv \phi\nabla\cdot(\rho\boldsymbol{u}) + \rho\boldsymbol{u}\cdot\nabla\phi; \quad \mathcal{C}^u_{\text{mom}}\phi \equiv \boldsymbol{u}\cdot\nabla(\rho\phi) + \rho\phi\nabla\cdot\boldsymbol{u};\\ &\mathcal{C}^\rho_{\text{mom}}\phi \equiv \rho\nabla\cdot(\boldsymbol{u}\phi) + \boldsymbol{u}\cdot\phi\nabla\rho; \quad \mathcal{C}^L_{\text{mom}}\phi \equiv \rho\phi\nabla\cdot\boldsymbol{u} + \rho\boldsymbol{u}\cdot\nabla\phi + \boldsymbol{u}\cdot\phi\nabla\rho.\end{aligned} \tag{8a-e}$$

In particular, Coppola et al. [7,15] consider weighted combinations of these formulations as

$$\mathcal{M}_{\text{mass}} = \xi\mathcal{M}^D_{\text{mass}} + (1-\xi)\mathcal{M}^A_{\text{mass}}; \tag{9a}$$

$$\mathcal{C}_{\text{mom}} = \alpha\mathcal{C}^D_{\text{mom}} + \beta\mathcal{C}^\phi_{\text{mom}} + \gamma\mathcal{C}^u_{\text{mom}} + \delta\mathcal{C}^\rho_{\text{mom}} + \varepsilon\mathcal{C}^L_{\text{mom}}, \tag{9b}$$

with $\alpha+\beta+\gamma+\delta+\varepsilon = 1$. They investigate which of these combinations will result in discrete global energy conservation in combination with a central finite-difference discretization. Their analysis reveals the following conditions on the weights:

$$\alpha - \varepsilon = \beta = \frac{1}{2}\xi \ \text{ and } \ \gamma = \delta = \frac{1}{2}(1-\xi) - \varepsilon, \tag{10a,b}$$

herewith defining a two-parameter family of formulations that globally preserve energy under central finite-difference discretization on uniform grids. As a special case, when $\varepsilon = 0$ the members of this family are also locally conservative and preserve mass and momentum, at least under central discretization on uniform grids. In this respect, an interesting relation, valid under slightly different conditions, is

$$\mathcal{C}_{\text{mom}}1 = \mathcal{M}_{\text{mass}} \quad \Leftrightarrow \quad \alpha+\beta = \xi \text{ and } \gamma+\delta+\varepsilon = 1-\xi. \tag{11}$$

These conditions are compatible with the conditions Eq. (10) for energy preservation when $\varepsilon = 0$. We will encounter these conditions again in a discrete setting in Section 5.1.

The family of combinations satisfying Eq. (10) contains several special cases known from the literature. The case $\xi = 1$ corresponds with the classical Feiereisen form [11], whereas $\xi = 1/2$ gives the splitting introduced by Kennedy and Gruber [32] and shown to be energy preserving by Pirozzoli [14]. The case $\xi = 0$ correspond with a new splitting introduced by Coppola et al. [15]. They all satisfy $\varepsilon = 0$. Further, the case $\xi = \alpha = 1$ with $\beta = \gamma = \delta = \epsilon = 0$ is the conservation form (1) which is the basis for finite-volume discretizations. In what follows, the mentioned results will be extended and reformulated in a more general setting. The analysis will also highlight important relations between discrete local and global conservation of linear and quadratic invariants and between finite-difference and finite-volume formulations.

2.3. Matrix-vector notation

To develop our discrete theory, we use matrix-vector notation to be explained next, which in principle is valid in any dimension. To simplify the mathematical formulations, it will be presented on a one-dimensional grid, but it can be generalized to more dimensions (although the notation then becomes somewhat cumbersome). General grid vectors (lower case) and matrices (upper case) are written in a sans-serif font. Quantities with a volume-consistent scaling (see below) are written in a Fraktur font. Unknowns like ρ, u and ϕ are represented by grid vectors and diagonal matrices:

$$\mathsf{R} = \text{diag}(\rho), \ \mathsf{U} = \text{diag}(\mathsf{u}) \ \text{ and } \ \Phi = \text{diag}(\phi). \tag{12}$$

The various realizations of the derivative operator $\mathfrak{D}$ that we are going to study will be defined in terms of the circulant permutation matrix

$$\mathsf{E} \equiv \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 \end{bmatrix}, \tag{13}$$

which is the matrix version of the *shift* operator e acting on grid variables: $e\phi_i = \phi_{i+1}$. The powers of e (both positive and negative) are naturally defined through the composition rule as $e^{\pm k}\phi_i = \phi_{i\pm k}$ and constitute a group of transformations which allows the specification of all the ordinary finite-difference formulas. As an example, the usual second-order approximation for a first-order derivative on a uniform grid with mesh size $h = 1$ is given by

$$\phi_i' = \frac{1}{2}(\phi_{i+1} - \phi_{i-1}) = \frac{1}{2}\left(e - e^{-1}\right)\phi_i.$$

In matrix notation the corresponding derivative operator on a periodic mesh is expressed as

$$\mathfrak{D}\phi = \mathfrak{D}^{\text{cen}}\phi \equiv \frac{1}{2}(\mathsf{E} - \mathsf{E}^{-1})\phi,$$

where the inverse of E is readily seen to be its transpose (i.e. E is orthogonal)

$$\mathsf{E}^{-1} = \mathsf{E}^{\mathrm{T}} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}.$$

From the obvious relations

$$\phi_i\psi_{i+k} = e^k\left(\phi_{i-k}\psi_i\right) \quad \text{and} \quad (\phi\psi)_{i+k} = \phi_{i+k}\psi_{i+k}$$

one obtains the matrix identities

$$\Phi\mathsf{E}^k = \mathsf{E}^k\text{diag}(\mathsf{E}^{-k}\phi) \quad \text{and} \quad \mathsf{E}^k\Phi = \text{diag}(\mathsf{E}^k\phi)\mathsf{E}^k. \tag{14a,b}$$

To better understand Eqs. (14a-b) and some other formulas in the subsequent sections, we explicitly note that the product between a diagonal matrix $\text{diag}(\phi)$ and a vector ψ is the matrix-vector expression for the Hadamard product (i.e. the componentwise product) between ϕ and ψ, giving as a result the vector with i-th component $\phi_i\psi_i$.

With these building blocks, any matrix can be expressed as a weighted sum of powers of E:

$$\mathfrak{D} = \sum_{k=-L}^{L} \mathsf{A}_k\mathsf{E}^k, \tag{15}$$

where L is (larger or equal to) its semi-bandwidth. The A_k are diagonal matrices $\mathsf{A}_k = \text{diag}(\mathsf{a}_k)$, built from suitably chosen vectors a_k, and each term $\mathsf{A}_k\mathsf{E}^k$ in Eq. (15) is a matrix carrying the vector a_k on its k-th diagonal. The components of a_k that are not used are assumed to be zero.

For matrices of the form (15), the row and column sums can be calculated as

$$\text{row sum:} \quad \sum_k \mathsf{A}_k\mathsf{E}^k\mathbf{1} = \sum_k \mathsf{A}_k\mathbf{1} = \sum_k \mathsf{a}_k; \tag{16a}$$

$$\text{column sum:} \quad \sum_k \left(\mathsf{A}_k\mathsf{E}^k\right)^{\mathrm{T}}\mathbf{1} = \sum_k \mathsf{E}^{-k}\mathsf{A}_k\mathbf{1} = \sum_k \mathsf{E}^{-k}\mathsf{a}_k, \tag{16b}$$

where $\mathbf{1}$ is the grid vector consisting of only ones and the obvious properties $\mathsf{E}^k\mathbf{1} = \mathbf{1}$ and $\mathsf{A}_k\mathbf{1} = \mathsf{a}_k$ have been used.

Note that the vectors a_k do not need to be "scalar" vectors (i.e. such that $\mathsf{a}_k = a_k\mathbf{1}$ with a_k scalar). This means that, although defined by using the circulant matrices E^k, the matrix $\mathfrak{D}$ in Eq. (15) does not need to be circulant in general, and is completely arbitrary. The use of the circulant version of the shift matrix is useful because the theory we are going to develop will be detailed for the case of periodic boundary conditions. However, most results will be stated as global properties of the relevant operators involved, and the extension to the case of non-periodic boundary conditions can be carried out by

suitably defining the basic derivative operators, e.g., as in the summation-by-parts (SBP) discretizations [33,34]. Especially, Svärd and Nordström [34] give an extensive discussion of the effect of boundary conditions.

Special matrices that we will encounter in the sequel are:

$$\text{skew-symmetric: } \mathfrak{D} = \sum_{k=1}^{L} (\mathsf{A}_k \mathsf{E}^k - \mathsf{E}^{-k} \mathsf{A}_k); \qquad \text{circulant: } \mathfrak{D} = \sum_{k=-L}^{L} a_k \mathsf{E}^k.$$

It is remarked that the precise value of the upper index bounds of the summations is not used in our considerations, so often they will not be indicated. The usual case of central derivative schemes on a uniform periodic mesh leads to derivative matrices which are both circulant and skew-symmetric; i.e. they have the form

$$\mathfrak{D} = \sum_{k=1}^{L} a_k \left(\mathsf{E}^k - \mathsf{E}^{-k} \right). \tag{17}$$

Equation (17) is the matrix version of the more familiar formula $\phi'_i = \sum_{k=1}^{L} a_k \left(\phi_{i+k} - \phi_{i-k} \right)$. The theory developed in [14,15] can be rephrased in matrix notation by using derivative matrices of the form (17), whereas the more general theory here illustrated will be developed by using the general form in Eq. (15).

2.4. Discrete conservation

To obtain a discrete analogue of the reasoning in Section 2.1, a *volume-consistent* scaling [16] of the equations is required. This scaling takes care that a summation over the grid cells, as in Eq. (20) below, corresponds with a discrete approximation of a volume integral. It is noted that the SBP discretization [33,34] also employs this scaling.

A PDE in conservation form can be denoted as

$$\frac{\partial \phi}{\partial t} + \nabla \cdot \mathcal{F}(\phi) = 0 \quad \text{discretized as} \quad \mathfrak{H} \frac{\mathrm{d}\phi}{\mathrm{d}t} + \mathfrak{C}\phi = 0, \tag{18a,b}$$

where $\mathfrak{H}$ is a diagonal matrix containing the sizes of the control volumes (positioned around the nodal points). The matrix $\mathfrak{C}$ represents the discretization of the convective operator $\nabla \cdot \mathcal{F}(\boldsymbol{.})$, which could depend on ϕ itself. As an example, for the 1D inviscid Burgers equation in divergence or advective forms:

$$\frac{\partial \phi}{\partial t} + \frac{\partial}{\partial x} \frac{\phi^2}{2} = 0, \qquad \frac{\partial \phi}{\partial t} + \phi \frac{\partial \phi}{\partial x} = 0$$

the matrix $\mathfrak{C}$ is $\frac{1}{2}\mathfrak{D}\Phi$ or $\Phi\mathfrak{D}$, respectively.

A volume-consistent scaling implies that under grid refinement, i.e. $|\mathfrak{H}|_\infty \to 0$, the discretization (18b) is consistent with the analytic differential equation (18a):

$$\forall \phi \text{ and } |\mathfrak{H}|_\infty \to 0: \quad \mathbf{1}^{\mathrm{T}} \mathfrak{H} \frac{\mathrm{d}\widehat{\phi}}{\mathrm{d}t} \to \int_\Omega \frac{\partial \phi}{\partial t} \mathrm{d}\Omega \quad \text{and} \quad \mathfrak{H}^{-1} \mathfrak{C} \widehat{\phi} \to \nabla \cdot \mathcal{F}(\phi), \tag{19}$$

where $\widehat{\phi}$ is the restriction of ϕ to the nodal grid points. We will also see that symmetries are directly expressed in such volume-consistent operators, thereby preventing 'pollution' of the equations by a frequent appearance of the local control volumes $\mathfrak{H}$. Below we will give some examples of volume-consistent discretizations; for instance in Eqs. (33)-(35).

Discrete linear invariants

A discrete analogue of the criterion for *global conservation* in Eq. (2) can be formed from a summation over grid cells:

$$\mathbf{1}^{\mathrm{T}} \mathfrak{H} \frac{\mathrm{d}\phi}{\mathrm{d}t} = -\mathbf{1}^{\mathrm{T}} \mathfrak{C}\phi = 0. \tag{20}$$

With the volume-consistent scaling, the summation in the left-hand side of Eq. (20) corresponds with a discrete approximation of the volume integral in Eq. (19). It follows that discrete *global conservation* of primary invariants is associated with the matrix $\mathfrak{C}$ having vanishing column sums.

Discrete *local conservation* of primary invariants as given in Eq. (3) is associated to a stricter property for $\mathfrak{C}$. We will consider here formulations (including finite differences and cell-centered finite volumes) for which local conservation is equivalent to the decomposition of $\mathfrak{C}$ as a 'difference of fluxes', i.e. as the difference between a (flux) vector and its 'shifted' version. This automatically leads to a telescoping property when summing over the domain in Eq. (20).

For its discrete analogue we need the shift operator E as defined in Eq. (13), to distinguish between fluxes in $i + 1/2$ and $i - 1/2$ (in cell-centered finite-volume terms). With this matrix shift operator, local conservation amounts to write the discrete transport terms in Eq. (18b) in a flux-type format as

$$\mathfrak{C}\phi = (\mathsf{I} - \mathsf{E}^{-1})\mathfrak{F}(\phi)\phi, \tag{21}$$

where $\mathfrak{F}(\phi)\phi$ is a consistent approximation of the flux $\mathcal{F}(\phi)$, carrying the flux in $i+1/2$ at its i-th component. The difference of fluxes between the points $i+1/2$ and $i-1/2$ can be written as $\mathcal{F}_{i+1/2} - \mathcal{F}_{i-1/2} = (1-e^{-1})\mathcal{F}_{i+1/2}$, which is the indicial version of Eq. (21). Note that $\mathsf{I} - \mathsf{E}^{-1}$ is not invertible, hence the flux $\mathfrak{F}(\phi)\phi$ cannot be found by inversion. Finite-volume discretizations automatically lead to this formulation, but for finite-difference discretizations this puts additional requirements. An important result in this respect is provided by Lemma A.1 (see the appendix): any matrix with vanishing column sums, i.e. a matrix which corresponds with global conservation, can be written in a local flux form (21). This result establishes a strong link between the concepts of discrete global and local conservation.

In the next sections we will discuss which consequences (global and local) discrete conservation has for the finite-volume and finite-difference discretizations of the family of split forms in (7) and (8). In particular we will show which relations have to exist between the discretizations of the individual terms in the split forms.

Remark 1. Note that Perot [35, p. 303] already observed that "any [discretization] method that is globally conservative, and that has weight functions with local support, must also be locally conservative". Also, compare the observation in [36] "... some discrete split forms of conservative laws can be manipulated into an equivalent, consistent and telescoping form...".

Remark 2. The matrix $\mathfrak{H}$ does not need to be diagonal, although in a cell-centered finite-volume method this will be the case. For global conservation, the only property required is that the summation over the grid cells in Eq. (19) gives an approximation of a volume integral. For diagonal matrices $\mathfrak{H}$, as we are considering here, the discrete volume integral of the time derivative in the left-hand side of Eq. (20) corresponds with a (simple) midpoint rule. A non-diagonal $\mathfrak{H}$ induces a more complex quadrature rule.

Consistency of fluxes

It has to be ensured that the flux $\mathfrak{F}\phi$ in Eq. (21) describes physically-consistent (first- or higher-order) approximations related to the underlying PDE Eq. (18a). We would like to prove that for a PDE with a constant flux, conservation and (volume) consistency imply that the discrete flux from the decomposition (21) becomes equal to that constant flux value. However, at this stage of generality of the theory, we can not (yet) prove this; therefore we formulate it as an assumption postponing to Section 3 some considerations justifying this assumption. In particular, for a PDE in conservation form, and with a volume-consistent scaling, we assume the following consistency condition to hold for fluxes $\mathfrak{F}\phi$:

$$\mathfrak{H}\frac{\mathrm{d}\phi}{\mathrm{d}t} + (\mathsf{I} - \mathsf{E}^{-1})\mathfrak{F}\phi = 0 \;\rightarrow\; \frac{\partial \phi}{\partial t} + \nabla \cdot \mathcal{F}(\phi) = 0 \qquad \Rightarrow \qquad \mathfrak{F}\phi_a \mathbf{1} = \mathcal{F}(\phi_a) \;\; \forall \phi_a, \tag{22a,b}$$

where ϕ_a is a (constant) scalar. This is the 'classical' Lax–Wendroff condition [37,38], which basically says that conservation and consistency imply that the numerical fluxes are a fair interpolation between the adjacent nodal values.

The above findings can be summarized as follows:

Observation 1. *For volume-consistent discretizations of transport equations like (18a), or their analytical equivalents, the discrete concepts of global and local conservation are equivalent, and are characterized by vanishing column sums of the discrete transport term.*

2.5. The control volumes $\mathfrak{H}$

The size for the control volumes $\mathfrak{H}$ is chosen such that on a non-uniform grid the discrete derivative of a linear function is exact. Thereto, assume that the x-grid is obtained from a smooth transformation $x = f(\eta)$, where η is covered by a uniform grid (with grid size $\Delta\eta$). Then we can write for a first-order derivative

$$\frac{\mathrm{d}\phi}{\mathrm{d}x} = \frac{\mathrm{d}\phi}{\mathrm{d}\eta}\frac{\mathrm{d}\eta}{\mathrm{d}x} = \frac{\mathrm{d}\phi}{\mathrm{d}\eta} \Big/ \frac{\mathrm{d}x}{\mathrm{d}\eta} \;\approx\; \mathrm{diag}(\mathfrak{D}\mathsf{x})^{-1}\mathrm{diag}(\mathfrak{D}\phi) \quad \Rightarrow \quad \mathfrak{H} = \mathrm{diag}(\mathfrak{D}\mathsf{x}). \tag{23}$$

Both numerator and denominator can be discretized on a uniform η-grid, i.e. in computational space (see also [39]). Then the order of the discretization on the non-uniform x-grid follows from the discretization on the uniform η-grid. When the control volumes are chosen as $\mathfrak{H} \equiv \mathrm{diag}(\mathfrak{D}\mathsf{x})$, the discrete derivative $\mathfrak{H}^{-1}\mathfrak{D}$ of a linear function is exact [40, p. 1885]. In this way, the choice of $\mathfrak{H}$ influences the order of accuracy of a discretization with a given $\mathfrak{D}$. We will demonstrate the influence of this choice for the control volumes in Section 7.2.

Example. The usual 4th-order discretization of a first-order derivative reads

$$\frac{\mathrm{d}\psi}{\mathrm{d}\eta} = \frac{-\psi_{i+2} + 8\psi_{i+1} - 8\psi_{i-1} + \psi_{i-2}}{12\Delta\eta} + \mathcal{O}(\Delta\eta^4). \tag{24}$$

Applying it with $\psi = x$ and $\psi = \phi$, respectively, gives (with f' the derivative of the transformation)

$$h_{4\text{th}-\text{order}} \equiv \frac{1}{12}(-x_{i+2} + 8x_{i+1} - 8x_{i-1} + x_{i-2}) = f'\Delta\eta + \mathcal{O}(\Delta\eta^5), \tag{25}$$

which leads to the 4th-order approximation (see Section 6)

$$\frac{\mathrm{d}\phi}{\mathrm{d}x} \approx \frac{-\phi_{i+2} + 8\phi_{i+1} - 8\phi_{i-1} + \phi_{i-2}}{-x_{i+2} + 8x_{i+1} - 8x_{i-1} + x_{i-2}}. \tag{26}$$

In contrast, when in Eq. (26) the grid size (25) is replaced by the 'usual' grid size we obtain

$$h_{2\text{nd}-\text{order}} \equiv \frac{1}{2}(x_{i+1} - x_{i-1}) = f'\Delta\eta + \mathcal{O}(\Delta\eta^3),$$

suggesting a 2nd-order approximation of the derivative $\mathrm{d}\phi/\mathrm{d}x$, which is confirmed in Section 7.2. □

The above lets us formulate

Observation 2 *(Choice of control volumes). For a consistent discretization of linear functions, the control volumes are preferably chosen according to* $\mathfrak{H} = \mathrm{diag}(\mathfrak{D}\mathsf{x})$.

3. Conservation of mass

The various formulations of the transport equations will first be inspected for their potential to globally and locally conserve the linear (primary) invariants mass and momentum. We will first do so on hand of the conservation equation for mass (7). The conservation equation for momentum (8) can be handled in a similar way, and will be discussed in Section 4 in a more general setting.

We will start by studying finite-volume methods for the discrete form of the equation for mass conservation (1a):

$$\mathfrak{H}\frac{\mathrm{d}\rho}{\mathrm{d}t} + \mathfrak{d}_{\text{mass}} = 0. \tag{27}$$

The volume-consistent grid vector $\mathfrak{d}_{\text{mass}}$ corresponds with the discrete version of $\mathcal{M}_{\text{mass}} = \nabla \cdot \boldsymbol{m} \equiv \nabla \cdot (\rho \boldsymbol{u})$ in the nodal points. As local and global conservation for these methods holds by design, in particular we inventorize the freedom that is left in the details of the flux discretization. We will use this freedom when additionally requiring energy conservation (Section 5). Thereafter, finite-difference methods will be analyzed for both the divergence (7a) and advective (7b) forms of the equation for mass. Finally, the link between the presented finite-volume and finite-difference methods for mass conservation is analyzed.

3.1. Finite-volume methods

First, let us consider finite-volume discretizations of Eq. (1a). To distinguish them from finite-difference discretizations, finite-volume discretizations are indicated by the superscript $(\boldsymbol{.})^{\text{fv}}$. In particular, we will consider cell-centered formulations [18], where the control volumes are located around the nodal grid points. The control faces are positioned between the nodal points; their precise position is not relevant for numerical stability.

A finite-volume formulation starts with the specification of the mass fluxes $m \equiv \rho u$ at cell faces which, in a one-dimensional setting, allows to express the convective term of the mass equation as $\mathcal{M}_{\text{mass}} \equiv (m_{i+1/2} - m_{i-1/2})$. In matrix notation one has:

$$\mathfrak{H}\frac{\mathrm{d}\rho}{\mathrm{d}t} = -\mathfrak{d}^{\text{fv}}_{\text{mass}} \equiv -(\mathsf{I} - \mathsf{E}^{-1})\,\mathsf{m}^{\text{fv}}_f(\mathsf{u}, \rho). \tag{28}$$

Here, m^{fv}_f is the vector containing the discrete mass fluxes $m_{i+1/2}$ through the faces of control volumes, which can be written in terms of the values of u and ρ in neighboring grid nodes. It is noted that the local flux form (21), which leads to global and local conservation, comes out naturally. There exists a large freedom in choosing the mass fluxes m^{fv}_f. In fact, consistency (22) of the flux is all we demand. Let us start exploring and exploiting this freedom.

Already within the three-point discretization stencils, i.e. two-point stencils for the mass flux, an interesting freedom can be recognized, which we will encounter again when discussing finite-difference methods. Anticipating that in the alternative formulations (7b) and (8b) the variables ρ and u are individually visible, the mass flux $m_{i+1/2} \equiv (\rho u)_{i+1/2}$ at the face $i+1/2$ can be built from values of ρ and u in the grid points i and $i+1$. In this setting, the most general form of a bilinear mass flux using a two-point stencil reads

$$\mathsf{m}^{\text{fv}}_f(\mathsf{u}, \rho): \quad m_{i+1/2} = c^{(1,1)}_{i+1/2}\,\rho_{i+1}u_{i+1} + c^{(1,0)}_{i+1/2}\,\rho_{i+1}u_i + c^{(0,1)}_{i+1/2}\,\rho_i u_{i+1} + c^{(0,0)}_{i+1/2}\,\rho_i u_i, \tag{29}$$

where the $c_{i+1/2}^{(p,q)}$ denotes the coefficient of $\rho_{i+p}u_{i+q}$ in the flux at the face $i+1/2$. For consistency with the physical flux $m \equiv \rho u$, we let these coefficients add up to unity, which corresponds with Eq. (22b).

Thus, in this two-point flux setting, we have a three-parameter family of coefficients that all produce a valid, consistent finite-volume discretization. By varying these coefficients, a directional bias can be given to the mass flux which we will encounter again below when analyzing finite-difference discretizations. Even more so, the coefficients can be chosen grid-point dependent and/or as more general functions of the neighboring nodal values.

Examples.

- A special case, which corresponds with the usual second-order central finite-volume discretization, is given by $c^{(1,1)} = c^{(0,0)} = \frac{1}{2}$, leading to

$$m_{i+1/2} = \frac{1}{2}(\rho_{i+1}u_{i+1} + \rho_i u_i) \quad \longleftrightarrow \quad \mathsf{m}_f^{\mathrm{fv}} = \frac{1}{2}(\mathsf{I}+\mathsf{E})\mathsf{R}\mathsf{u}, \tag{30}$$

  and corresponds with the form used by Feiereisen [11] represented by $\xi = 1$ in Coppola's family of split forms (9a). See Table 1 for more references of its use.
- Another popular choice for the mass flux is given by splitting m in ρ and u and choosing all four coefficients as $c^{(\cdot,\cdot)} = \frac{1}{4}$:

$$m_{i+1/2} = \frac{1}{4}(\rho_i + \rho_{i+1})(u_i + u_{i+1}) \quad \longleftrightarrow \quad \mathsf{m}_f^{\mathrm{fv}} = \frac{1}{4}\,\mathrm{diag}\big((\mathsf{I}+\mathsf{E})\rho\big)\,(\mathsf{I}+\mathsf{E})\mathsf{u}. \tag{31}$$

  It is the choice by Kennedy–Gruber [32] and Pirozzoli [14] corresponding with $\xi = 1/2$ in Coppola's family of split forms (9a). □

3.2. Finite-difference methods – divergence form

Our pursuit of locally and globally conservative discrete formulations is continued with a volume-consistent finite-difference discretization of the divergence form (7a). Thus, as in the previous section, we will study the discrete setting from Eq. (27). Since local conservation implies also global conservation by means of the telescoping property, we start the analysis considering global conservation and move later towards local conservation.

Global mass conservation As we have seen in Section 2.4, for a discretization with volume-consistent scaling as in Eq. (27), global conservation (20) is equivalent to the vanishing of the column sums. For the discrete formulation (27) with the divergence form

$$\mathfrak{H}\frac{\mathrm{d}\rho}{\mathrm{d}t} = -\mathfrak{D}\mathsf{m}, \tag{32}$$

with $\mathfrak{D}$ a scaled first-order derivative matrix, for which we will assume vanishing row sums. Global conservation boils down to $\mathbf{1}^{\mathrm{T}}\mathfrak{D}\mathsf{m} = 0$, which implies that global mass conservation is associated with a derivative matrix having vanishing column sums.

Matrices $\mathfrak{D}$ with vanishing row and column sums are sometimes called double-centered matrices [41,42]. Derivative matrices (having vanishing row sums) have also vanishing column sums if they are

1. symmetric or skew symmetric;
2. circulant or Toeplitz [43, Ch. 4.7].

For three-point stencils both classes coincide, but for wider stencils the two classes are not identical. As a particular case, apart from boundary effects, the summation-by-parts (SBP) matrices fit into the first class [33,34]. Obviously, also any linear combination of such matrices possesses zero row and column sums.

Examples. To make the above concrete, we present some examples:

- Any mesh-independent, volume-consistent numerical derivative formula (both central or unsymmetric) on a periodic mesh generates a circulant matrix $\mathfrak{D}$, which guarantees global conservation when used inside the divergence form, even on non-uniform meshes. Circulant examples are the central skew-symmetric discretization

$$\frac{1}{2}(x_{i+1} - x_{i-1})\left.\frac{\partial\phi}{\partial x}\right|_i = \frac{1}{2}(\phi_{i+1} - \phi_{i-1}) \quad \longleftrightarrow \quad \mathfrak{D}^{\mathrm{cen}}\phi = \frac{1}{2}(\mathsf{E} - \mathsf{E}^{-1})\phi, \tag{33}$$

  and a directionally biased discretization

$$(x_i - x_{i-1}) \left.\frac{\partial \phi}{\partial x}\right|_i = (\phi_i - \phi_{i-1}) \quad \longleftrightarrow \quad \mathfrak{D}^{\text{upw}}\phi = (\mathsf{I} - \mathsf{E}^{-1})\phi. \tag{34}$$

Of course, the corresponding, non-scaled derivative matrices $\mathsf{D} = \mathfrak{H}^{-1}\mathfrak{D}$ are not circulant in general for non-uniform meshes – the scaling of the equations is essential when studying conservation.

- The 'classical' second-order Lagrangian derivative, scaled with the volumes $\frac{1}{2}(x_{i+1} - x_{i-1})$, is given by

$$\text{Lagrangian:} \quad \mathfrak{D}^{\text{Lag}}\phi_i \equiv \frac{1}{2}\left[\frac{x_i - x_{i-1}}{x_{i+1} - x_i}(\phi_{i+1} - \phi_i) + \frac{x_{i+1} - x_i}{x_i - x_{i-1}}(\phi_i - \phi_{i-1})\right], \tag{35}$$

which can be written as

$$\mathfrak{D}^{\text{Lag}} = \mathsf{A}_1\mathsf{E} + (\mathsf{A}_{-1} - \mathsf{A}_1) - \mathsf{A}_{-1}\mathsf{E}^{-1} \text{ with } (\mathsf{a}_1)_i = \frac{1}{2}\frac{x_i - x_{i-1}}{x_{i+1} - x_i} \text{ and } (\mathsf{a}_{-1})_i = \frac{1}{2}\frac{x_{i+1} - x_i}{x_i - x_{i-1}}.$$

On a non-uniform grid, in general this volume-consistent discretization is not skew-symmetric nor circulant. It is well known that in these cases it does not globally conserve invariants. Therefore we will not consider this discretization any further. □

Local mass conservation In Lemma A.1 it is shown that any matrix $\mathfrak{D} = \sum_k \text{diag}(\mathsf{a}_k)\mathsf{E}^k$ with vanishing column sums can be written in a locally conservative form

$$\mathfrak{D} = (\mathsf{I} - \mathsf{E}^{-1}) \sum_{k=-L}^{L} \text{diag}(\mathsf{b}_k)\mathsf{E}^k \equiv (\mathsf{I} - \mathsf{E}^{-1})\mathfrak{F}, \tag{36}$$

where

$$\mathsf{b}_k = \sum_{h=k}^{L} \mathsf{E}^{k-h}\mathsf{a}_h. \tag{37}$$

This is our first explicit 'difference of fluxes' decomposition of a finite-difference discretization. We will introduce another such decomposition in the sequel associated to the advective form. Eq. (36) defines the flux associated to the divergence form as

$$\mathfrak{F}\mathsf{m} = \sum_{k=-L}^{L} \text{diag}(\mathsf{b}_k)\mathsf{E}^k\mathsf{m}. \tag{38}$$

Above we assumed that for a volume-consistent discretization Eq. (22b) holds, which in our case reads

$$\mathfrak{F}\mathbf{1} = \sum_{k=-L}^{L} \mathsf{b}_k = \mathbf{1}. \tag{39}$$

It is remarked that the vectors b_k defined in Eq. (36) play an essential role in the relation between finite-difference and finite-volume discretizations; see Section 3.4, especially Eq. (53). This relation is already emerging in Eq. (38), showing that these vectors are closely related to the fluxes.

Observation 3. *With a volume-consistent scaling, the discrete divergence form (28) globally and locally conserves linear invariants if and only if it has vanishing column sums.*

Note that, since the matrix $\mathfrak{D}$ is a derivative matrix, it has vanishing row sums ($\mathfrak{D}\mathbf{1} = \mathbf{0}$) and Eq. (36) immediately shows that $\mathfrak{F}\mathbf{1} = c\mathbf{1}$, with c an arbitrary constant. This result is close to the assumed consistency relation (39), although the value of the constant c is undetermined. This actually should not be surprising, since we only assumed vanishing row and column sums for the matrix $\mathfrak{D}$, whereas the consistency condition (39), which asserts that $\mathfrak{F}$ is an interpolation operator, implies that $\mathfrak{D}$ needs to be a *first-order* derivative matrix. Since the condition (39) turns out to be a consistency condition for the fluxes, we can actually use it as a characterization of $\mathfrak{D}$ as a first-order derivative matrix. In fact, in the cases in which $\mathfrak{D}$ is defined from the beginning as a first-order derivative matrix, as in the *Example* below, Eq. (39) is automatically satisfied.

Example. As a concrete example, in the common case in which the *same numerical formula* for a first-order derivative is used on all nodes of the mesh with periodic boundary conditions, the derivative matrix is circulant. In this last case, Eq. (39) can be rearranged to the consistency condition (see Corollary A.2)

$$\sum_{k=-L}^{L} k a_k = 1. \tag{40}$$

This condition is always satisfied by a volume-consistent numerical scheme for a first-order derivative on a uniform grid. Furthermore, in Corollary A.2 it is shown that for circulant derivative matrices which are also skew-symmetric (i.e. associated to central schemes), the flux can be written as

$$\mathfrak{F}\mathsf{m} = 2\sum_{k=1}^{L} a_k \sum_{h=0}^{k-1} \mathsf{E}^{-h}\left[\frac{\mathsf{I}+\mathsf{E}^k}{2}\mathsf{m}\right]. \tag{41}$$

This is the matrix form of Eq. (A.1) in [15], with the interpolation operator (there defined in the first of Eqs. (A.2)) expressed by the term in the square brackets. This analysis also shows that the present treatment gives a generalization of the fluxes derived in [14]. In Section 3.4 we will show how the fluxes presented in this section relate to a finite-volume formulation. □

3.3. Finite-difference methods – advective form

To analyze the conservative properties of the advective form, we consider a general discretization of Eq. (7b) in which the two occurrences of the same derivative $\mathfrak{D}$ are replaced by two, possibly different, scaled derivative matrices. In particular we consider the advective form with volume-consistent scaling

$$\mathfrak{H}\frac{\mathrm{d}\rho}{\mathrm{d}t} = -(\mathsf{R}\mathfrak{D}^{\mathsf{u}}\mathsf{u} + \mathsf{U}\mathfrak{D}^{\rho}\rho). \tag{42}$$

Global mass conservation By applying the usual principle (20) that discrete global conservation is obtained when the sum over the grid cells of Eq. (42) vanishes, one obtains the condition

$$\mathbf{1}^{\mathsf{T}}\mathsf{R}\mathfrak{D}^{\mathsf{u}}\mathsf{u} + \mathbf{1}^{\mathsf{T}}\mathsf{U}\mathfrak{D}^{\rho}\rho = 0 \quad \Leftrightarrow \quad \rho^{\mathsf{T}}\mathfrak{D}^{\mathsf{u}}\mathsf{u} + \mathsf{u}^{\mathsf{T}}\mathfrak{D}^{\rho}\rho = 0.$$

Since each term is a scalar quantity, the second term can be transposed, leading to

$$\rho^{\mathsf{T}}(\mathfrak{D}^{\mathsf{u}} + \mathfrak{D}^{\rho\mathsf{T}})\mathsf{u} = 0 \quad \forall\rho \text{ and } \mathsf{u} \quad \Leftrightarrow \quad \mathfrak{D}^{\rho} = -\mathfrak{D}^{\mathsf{u}\mathsf{T}}. \tag{43}$$

This duality condition is the most general necessary and sufficient condition for global conservation of linear invariants for the advective form (42). Observe that $\mathfrak{D}^{\mathsf{u}}$ corresponds with a divergence operator and $\mathfrak{D}^{\rho}$ with a gradient operator, together forming a discrete product rule; see for example Lemma A.3. In our applications we will usually consider derivative matrices with strictly vanishing row sums. In that case both $\mathfrak{D}^{\mathsf{u}}$ and $\mathfrak{D}^{\rho}$ have vanishing row and column sums: they can be either (skew-)symmetric or circulant (or a linear combination). In case one of the discrete operators possesses a directional bias, the dual operator should possess the opposite directional bias; we call such discretizations *dual-sided*. When we demand that the matrix operators are equal, i.e. $\mathfrak{D}^{\rho} = \mathfrak{D}^{\mathsf{u}}$, the duality condition (43) implies that they are skew-symmetric.

Local mass conservation Under the necessary condition (43) for global conservation, Lemma A.3 shows that the advective form can be cast in a local conservation format (21);

$$\mathsf{R}\mathfrak{D}^{\mathsf{u}}\mathsf{u} + \mathsf{U}\mathfrak{D}^{\rho}\rho = (\mathsf{I}-\mathsf{E}^{-1})\mathsf{m}_f, \tag{44}$$

with mass flux m_f. When $\mathfrak{D}^{\mathsf{u}} = \sum_k \mathsf{A}^u_k\mathsf{E}^k$ and $\mathfrak{D}^{\rho} = \sum_k \mathsf{A}^{\rho}_k\mathsf{E}^k$ satisfying (43), i.e. $\mathsf{A}^{\rho}_k\mathsf{E}^k = -\mathsf{E}^k\mathsf{A}^u_{-k}$, the flux vector is given by

$$\mathsf{m}_f = \sum_{k>0}\left(\sum_{h=0}^{k-1}\mathsf{E}^{-h}\right)\left(\mathsf{R}\mathsf{A}^u_k\mathsf{E}^k\mathsf{u} - \mathsf{U}\mathsf{E}^k\mathsf{A}^u_{-k}\rho\right) \overset{(43)}{=} \sum_{k>0}\left(\sum_{h=0}^{k-1}\mathsf{E}^{-h}\right)\left(\mathsf{R}\mathsf{A}^u_k\mathsf{E}^k\mathsf{u} + \mathsf{U}\mathsf{A}^{\rho}_k\mathsf{E}^k\rho\right). \tag{45}$$

Consistency (22) of the flux amounts to

$$\sum_{k>0}\left(\sum_{h=0}^{k-1}\mathsf{E}^{-h}\right)\left(\mathsf{a}^u_k - \mathsf{E}^k\mathsf{a}^u_{-k}\right) \overset{(43)}{=} \sum_{k>0}\left(\sum_{h=0}^{k-1}\mathsf{E}^{-h}\right)\left(\mathsf{a}^u_k + \mathsf{a}^{\rho}_k\right) = 1. \tag{46}$$

We will encounter this (rather complicated) expression again when making the comparison with finite-volume discretizations in Section 3.4.

For the moment we conclude:

Observation 4. *With a volume-consistent scaling, the discrete advective form (42) globally and locally conserves linear invariants if and only if the two discrete derivative operators satisfy the duality relation (43), which implies that they have both vanishing row and column sums.*

Combining Observations 3 and 4 inspires to study the following generalized discrete version of the analytic mass transport operator from Eq. (9a):

$$\mathfrak{d}_{\text{mass}} = \xi \mathfrak{D}^{\rho u} \mathsf{R}\mathsf{u} + (1-\xi)(\mathsf{U}\mathfrak{D}^{\rho}\rho + \mathsf{R}\mathfrak{D}^{\mathsf{u}}\mathsf{u}). \tag{47}$$

Examples.

- As an illustration of the flux decomposition corresponding with Eq. (45), the second-order central discretization of the advective form can be rewritten as
$$\frac{1}{2}\rho_i(u_{i+1}-u_{i-1}) + \frac{1}{2}u_i(\rho_{i+1}-\rho_{i-1}) = \frac{1}{2}(\rho_i u_{i+1} + \rho_{i+1}u_i) - \frac{1}{2}(\rho_{i-1}u_i + \rho_i u_{i-1}),$$
where we recognize a discrete product rule. In matrix-vector notation it has the form (44) with mass flux $\mathsf{m}_f = \frac{1}{2}(\mathsf{R}\mathsf{E}\mathsf{u} + \mathsf{U}\mathsf{E}\rho)$. It corresponds with the case $\xi = 0$ of Coppola's family of split forms (9).
- A simple illustration of a directionally-biased discretization of the advective form can be re-formulated in finite-volume flux form as
$$u_i(\rho_i - \rho_{i-1}) + \rho_i(u_{i+1}-u_i) = \rho_i u_{i+1} - \rho_{i-1}u_i \quad \leftrightarrow \quad \mathsf{m}_f = \mathsf{R}\mathsf{E}\mathsf{u}.$$
- In the common case in which the derivative matrices $\mathfrak{D}^{\rho}$ and $\mathfrak{D}^{u}$ are the same (skew-symmetric) circulant matrix, Eq. (45) can be rewritten as
$$\mathsf{m}_f = 2\sum_{k>0} a_k \left(\sum_{h=0}^{k-1} \mathsf{E}^{-h}\right)\left[\frac{\mathsf{R}\mathsf{E}^k\mathsf{u} + \mathsf{U}\mathsf{E}^k\rho}{2}\right], \tag{48}$$
which is again the matrix form of Eq. (A.1) in [15], with the interpolation operator (there defined in the second of Eqs. (A.2)) expressed in Eq. (48) by the term in the square brackets. Under the same hypotheses, Eq. (46) can be written as
$$\sum_{k>0} 2ka_k = 1, \tag{49}$$
which is the skew-symmetric case of Eq. (40) and is satisfied by all central schemes for first-order derivatives on uniform grids.
- Another interesting case is obtained by averaging the above divergence and advective forms. For a circulant discretization, it admits a flux given by the average of Eqs. (41) and (48):
$$\mathsf{m}_f = \sum_{k>0} a_k \left(\sum_{h=0}^{k-1} \mathsf{E}^{-h}\right) \operatorname{diag}\big[(\mathsf{I}+\mathsf{E}^k)\rho\big](\mathsf{I}+\mathsf{E}^k)\mathsf{u}.$$
It gives the matrix form of Eq. (A.1) in [15], with the interpolant from Eq. (A.4). Compare Eq. (69) below. □

3.4. Link between finite-volume and finite-difference discretizations

The above finite-difference discretizations Eq. (32) for the divergence form and Eq. (42) for the advective form can be recast in a cell-centered finite-volume form. In particular, a link can be made with the general mass fluxes introduced in Eq. (29).

When 'translated' into grid-point notation, the discretization of the divergence form corresponds with a mass flux (38) through a face $i+1/2$ given by

$$m_{i+1/2} = \sum_{k=-L}^{L}\sum_{h=k}^{L} (\mathsf{a}_h^{\rho u})_{i+k-h}(\rho u)_{i+k}, \tag{50}$$

where $\mathsf{a}_h^{\rho u}$ are the vectors defining the matrix $\mathfrak{D}^{\rho u}$ as in Eq. (15) and $\sum_{h=k}^{L}(\mathsf{a}_h^{\rho u})_{i+k-h}$ is the i-th component of the associated vector b_k (cf. Eq. (37)). The flux $m_{i+1/2}$ is built from factors $(\rho u)_j = \rho_j u_j$ in the neighboring grid points, thereby generalizing the two-point flux family (29) to larger stencils. On the other hand, the discretization of the advective form corresponds with a mass flux (45) given by

$$m_{i+1/2} = \sum_{k=1}^{L}\sum_{h=0}^{k-1}\big[(\mathsf{a}_k^{\mathsf{u}})_{i-h}\rho_{i-h}u_{i+k-h} + (\mathsf{a}_k^{\rho})_{i-h}u_{i-h}\rho_{i+k-h}\big]. \tag{51}$$

It is observed that this flux consists of products $\rho_{i+p}u_{i+q}$ with factors evaluated in different neighboring grid points $i+p$ and $i+q$ where $p \neq q$ (since $k > 0$). More precise, $\mathsf{a}^{\mathsf{u}}_k$ corresponds with $p < q$, whereas a^{ρ}_k corresponds with $p > q$. Hence, in this respect the advective form (51) is complementary to the divergence form (50) where ρ_{i+p} and u_{i+q} are evaluated in the same points (i.e. $p = q$).

Thus, these fluxes fit in the formulation

$$m_{i+1/2} = \sum_p \sum_q c^{(p,q)}_{i+1/2} \rho_{i+p} u_{i+q} \quad \text{with} \quad \sum_p \sum_q c^{(p,q)}_{i+1/2} = 1, \tag{52a,b}$$

which generalizes Eq. (29). In particular, the relations between the two types of notation are (compare Eq. (36))

$$\mathsf{c}^{(k,k)} = \sum_{h=k}^{L} \mathsf{E}^{k-h} \mathsf{a}^{\rho u}_h = \mathsf{b}_k, \quad \mathsf{c}^{(-h,-h+k)} = \mathsf{E}^{-h} \mathsf{a}^{\mathsf{u}}_k \text{ and } \mathsf{c}^{(-h+k,-h)} = \mathsf{E}^{-h} \mathsf{a}^{\rho}_k. \tag{53}$$

Combining these relations with the finite-volume consistency condition (52b), shows that the consistency conditions (39) and (46) imply that the coefficients of the products of ρ and u add up to unity. We will come back to this correspondence in Section 5.4, and for now conclude:

Observation 5 *(Finite-difference versus finite-volume). Any finite-difference discretization with volume-consistent scaling, constructed as a combination of divergence and advective forms, that globally conserves mass is also locally conservative. In particular, it can be reformulated as a (cell-centered) finite-volume discretization.*

Whether the opposite is also true, i.e. can any finite-volume discretization be reformulated as a locally conservative finite-difference discretization constructed as a combination of divergence and advective forms, is still not clear. In Section 5.4 we will further investigate this aspect.

4. Conservation of momentum

Turning next to the discrete conservation of the momentum transport equation, we again first investigate finite-volume discretizations. Inspired by this inventory, thereafter we will design a wider class of finite-difference discretizations. The discretizations studied have the generic form

$$\mathfrak{H} \frac{\mathrm{d}\mathsf{R}\phi}{\mathrm{d}t} + \mathfrak{C}_{\mathrm{mom}} \phi = 0. \tag{54}$$

As argued above, global conservation of momentum boils down to vanishing column sums

$$\mathbf{1}^{\mathrm{T}} \mathfrak{C}_{\mathrm{mom}} = \mathbf{0}^{\mathrm{T}}.$$

Recalling that the advective operator $\mathfrak{C}_{\mathrm{mom}}$ can be built from various split forms (8), this condition puts constraints on the discretization of the individual terms. In this section we will focus on these constraints.

It is emphasized that the factor ρ, discretized as R, is absorbed in the operator $\mathfrak{C}_{\mathrm{mom}}$. As we will explain below, the advective argument is split according to $\rho \boldsymbol{u}\phi = \rho \boldsymbol{u} * \phi = \boldsymbol{m} * \phi$, in contrast with $\boldsymbol{u} * \rho\phi$ which is the more usual way. The discretizations in [8,32,44] belong to the few that explicitly recognize the importance of this type of splitting of the advective argument.

4.1. Finite-volume methods

The advective fluxes in a finite-volume discretization are built from the product $m\phi$, to be interpolated from the neighboring nodal values. In general, a discretization of the advective term will look like (compare Eq. (28))

$$\mathfrak{C}^{\mathrm{fv}}_{\mathrm{mom}} \phi = (\mathsf{I} - \mathsf{E}^{-1}) \mathfrak{F}^{\mathrm{fv}} \phi \quad \text{with flux} \quad \mathfrak{F}^{\mathrm{fv}} \phi \equiv \left(\mathrm{diag}(\mathsf{m}^{\mathrm{fv}}) \phi \right)_f. \tag{55}$$

A choice has to be made whether the product $m\phi$ in the facial fluxes are formed from the interpolation of the product $(m\phi)_f$, from the product of the interpolations $m_f \phi_f$ or from a more complicated expression. For the conservation of momentum all choices are fine. The freedom for m^{fv} that we encountered in choosing the mass fluxes in Section 3 is still available here; moreover, there is freedom in choosing ϕ_f. Any consistent interpolation between the neighboring nodal points is acceptable for discrete momentum conservation. However, we will see in Section 5 that for discrete energy conservation the choice does matter.

4.2. Generalized finite-difference methods

For finite-difference methods conservation is not immediate. Inspired by the freedom in the finite-volume discretizations and in the formulation of the equation for mass conservation (Section 3) we will generalize the discretization of the momentum transport equation. In a similar way as the discrete mass operator in Eq. (47), the discrete version of the advective transport operator (9b) is generalized as

$$\begin{aligned}\mathfrak{C}_{\text{mom}} = {} & \alpha\mathfrak{D}^{\rho u}\mathsf{R}\mathsf{U} + \beta\mathsf{R}\mathsf{U}\mathfrak{D}^{0} + \gamma\mathsf{U}\mathfrak{D}^{\rho}\mathsf{R} + \delta\mathsf{R}\mathfrak{D}^{u}\mathsf{U} + \varepsilon\mathsf{R}\mathsf{U}\mathfrak{D}^{0} \\ & + \operatorname{diag}\big[\beta\mathfrak{D}^{\rho u}\mathsf{R}u + \gamma\mathsf{R}\mathfrak{D}^{u}u + \delta\mathsf{U}\mathfrak{D}^{\rho}\rho + \varepsilon(\mathsf{R}\mathfrak{D}^{u}u + \mathsf{U}\mathfrak{D}^{\rho}\rho)\big],\end{aligned} \tag{56}$$

with $\alpha+\beta+\gamma+\delta+\varepsilon=1$. The discrete matrices in (47) and (56) are chosen the same to provide the necessary consistency for obtaining discrete energy conservation, as we will see below in Section 5.3.

In many of the preceding sections, the finite-difference matrices $\mathfrak{D}^{(\cdot)}$ were equal to the central discretization $\mathfrak{D}^{\text{cen}} = \frac{1}{2}(\mathsf{E}-\mathsf{E}^{-1})$, but now we do allow different (consistent) discrete approximations of the various ∇-operators. Thus, let us analyze the requirements that have to be invoked to ensure discrete preservation of momentum of such a generalized finite-difference approach.

Global momentum conservation Global momentum conservation boils down to $\mathbf{1}^{\mathrm{T}}\mathfrak{C}_{\text{mom}} = \mathbf{0}^{\mathrm{T}}$; see Eq. (20). Using $\mathbf{1}^{\mathrm{T}}\mathfrak{D}^{\rho u} = \mathbf{0}^{\mathrm{T}}$ (see above) and $u^{\mathrm{T}}\mathsf{R} = \rho^{\mathrm{T}}\mathsf{U}$ (as products of diagonal matrices), this amounts to:

$$\begin{aligned}\mathbf{1}^{\mathrm{T}}\mathfrak{C}_{\text{mom}} &= \alpha\mathbf{1}^{\mathrm{T}}\mathfrak{D}^{\rho u}\mathsf{R}\mathsf{U} + \beta\mathbf{1}^{\mathrm{T}}\big[\mathsf{R}\mathsf{U}\mathfrak{D}^{0} + \operatorname{diag}(\mathfrak{D}^{\rho u}\mathsf{R}u)\big] + \gamma\mathbf{1}^{\mathrm{T}}\big[\mathsf{U}\mathfrak{D}^{\rho}\mathsf{R} + \operatorname{diag}(\mathsf{R}\mathfrak{D}^{u}u)\big] + \\ &\qquad + \delta\mathbf{1}^{\mathrm{T}}\big[\mathsf{R}\mathfrak{D}^{u}\mathsf{U} + \operatorname{diag}(\mathsf{U}\mathfrak{D}^{\rho}\rho)\big] + \varepsilon\mathbf{1}^{\mathrm{T}}[\mathsf{R}\mathsf{U}\mathfrak{D}^{0} + \operatorname{diag}(\mathsf{R}\mathfrak{D}^{u}u + \mathsf{U}\mathfrak{D}^{\rho}\rho)] \\ &= \beta\big[\rho^{\mathrm{T}}\mathsf{U}\mathfrak{D}^{0} + (\mathfrak{D}^{\rho u}\mathsf{R}u)^{\mathrm{T}}\big] + \gamma\big[u^{\mathrm{T}}\mathfrak{D}^{\rho}\rho + (\mathsf{R}\mathfrak{D}^{u}u)^{\mathrm{T}}\big] + \\ &\qquad + \delta\big[\rho^{\mathrm{T}}\mathfrak{D}^{u}\mathsf{U} + (\mathsf{U}\mathfrak{D}^{\rho}\rho)^{\mathrm{T}}\big] + \varepsilon[\rho^{\mathrm{T}}\mathsf{U}\mathfrak{D}^{0} + (\mathsf{R}\mathfrak{D}^{u}u)^{T} + (\mathsf{U}\mathfrak{D}^{\rho}\rho)^{T}] \\ &= \beta\rho^{\mathrm{T}}\mathsf{U}[\mathfrak{D}^{0} + \mathfrak{D}^{\rho u\mathrm{T}}] + \gamma u^{\mathrm{T}}\big[\mathfrak{D}^{\rho} + \mathfrak{D}^{u\mathrm{T}}\big]\mathsf{R} + \delta\rho^{\mathrm{T}}\big[\mathfrak{D}^{u} + \mathfrak{D}^{\rho\mathrm{T}}\big]\mathsf{U} + \\ &\qquad + \varepsilon[\rho^{\mathrm{T}}\mathsf{U}\mathfrak{D}^{0} + u^{\mathrm{T}}\mathfrak{D}^{u\mathrm{T}}\mathsf{R} + \rho^{\mathrm{T}}\mathfrak{D}^{\rho\mathrm{T}}\mathsf{U}].\end{aligned} \tag{57}$$

When $\varepsilon = 0$, it can be seen that this expression vanishes for all ρ and u, i.e. it admits global momentum conservation, if and only if

$$\mathfrak{D}^{0} = -\mathfrak{D}^{\rho u\mathrm{T}} \quad \text{and} \quad \mathfrak{D}^{\rho} = -\mathfrak{D}^{u\mathrm{T}}. \tag{58}$$

The second duality condition we have seen before in Eq. (43) when studying mass conservation of the advective form of mass transport. For derivative matrices, the duality conditions (58) imply that all four matrices $\mathfrak{D}^{(\cdot)}$ have vanishing row and column sums. In this case with $\varepsilon = 0$, there are no special conditions (yet) on the weights α, β, γ and δ. In the case $\varepsilon \neq 0$ it can be shown (not presented here) that discrete conservation of momentum requires all matrices $\mathfrak{D}^{(\cdot)}$ to be the same, and to satisfy the discrete product rule $\mathfrak{D}\mathsf{U}\rho = \mathsf{R}\mathfrak{D}u + \mathsf{U}\mathfrak{D}\rho$. Because such a $\mathfrak{D} \neq \mathbf{0}$ does not exist, discrete advective momentum conservation requires $\varepsilon = 0$.

Local momentum conservation As already used in Section 2.4, in Lemma A.1 it is proven that any matrix with vanishing column sums can be factorized in a local flux form. This implies that under the duality conditions (58) the matrix $\mathfrak{C}_{\text{mom}}$ studied in Eq. (57) contains a factor $\mathsf{I}-\mathsf{E}^{-1}$. A volume-consistent scaling invokes that the flux is consistent; see Section 2.4.

Observation 6 *(Discrete conservation of momentum). With a volume-consistent scaling, finite-difference discretizations for the split forms of the momentum equation in (9) with $\varepsilon = 0$ advectively conserve momentum, globally and locally, if and only if the difference operators satisfy the duality relations (58). In that case, the resulting discretization can be written as a finite-volume discretization.*

5. Conservation of kinetic energy

5.1. General criteria

Finally, we turn to the discrete advective conservation of kinetic energy. For our discrete transport equations, the discrete evolution of the quadratic invariant kinetic energy can be obtained by combining the discrete conservation equations for mass and momentum, and results in

$$\mathfrak{H}\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2}\mathsf{R}\Phi\phi\right) = \mathfrak{H}\left(\Phi\frac{\mathrm{d}\mathsf{R}\phi}{\mathrm{d}t} - \frac{1}{2}\Phi^{2}\frac{\mathrm{d}\rho}{\mathrm{d}t}\right) = -\Phi\left(\mathfrak{C}_{\text{mom}} - \frac{1}{2}\operatorname{diag}(\mathfrak{d}_{\text{mass}})\right)\phi. \tag{59}$$

It follows, as in Section 2.1, that discrete global energy conservation is guaranteed if and only if [9,16]

$$\mathfrak{A}_{\text{kin}} \equiv \mathfrak{C}_{\text{mom}} - \frac{1}{2}\,\text{diag}(\mathfrak{d}_{\text{mass}}) \text{ is skew symmetric,} \tag{60a}$$

which can be reformulated as

$$\text{diag}(\mathfrak{d}_{\text{mass}}) = \mathfrak{C}_{\text{mom}} + \mathfrak{C}_{\text{mom}}^{\text{T}}. \tag{60b}$$

This necessary and sufficient condition shows that outside its diagonal a discrete advective operator $\mathfrak{C}_{\text{mom}}$ must be skew-symmetric, while at its diagonal a discrete mass vector $\mathfrak{d}_{\text{mass}}$ must be recognized.

The diagonal relation between the discrete mass and momentum equations has been noticed before. E.g., Pirozzoli [45, p. 2999] writes that kinetic energy preservation is only achieved "when both the continuity and the momentum equations are split [in the same way]", while Morinishi [46, p. 278] writes "... as long as the [diagonal] terms in [the momentum equation] are discretized in the same manner as that in the discrete continuity". Our analysis gives the mathematical proof of the necessity of this relation.

As a special case, multiplying condition (60b) with the all-ones vector $\mathbf{1}$, discrete energy preservation implies that

$$\mathfrak{d}_{\text{mass}} = \text{diag}(\mathfrak{d}_{\text{mass}})\mathbf{1} = (\mathfrak{C}_{\text{mom}} + \mathfrak{C}_{\text{mom}}^{\text{T}})\mathbf{1}. \tag{61}$$

Hence, under energy preservation the mass transport term is fixed as soon as the advective operator $\mathfrak{C}_{\text{mom}}$ has been chosen. When also momentum is conserved, i.e. when the column sums of $\mathfrak{C}_{\text{mom}}$ vanish implying $\mathfrak{C}_{\text{mom}}^{\text{T}}\mathbf{1} = (\mathbf{1}^{\text{T}}\mathfrak{C}_{\text{mom}})^{\text{T}} = \mathbf{0}$, this expression for the mass transport term can be simplified to

$$\mathfrak{d}_{\text{mass}} = \mathfrak{C}_{\text{mom}}\mathbf{1} \quad \Rightarrow \quad \mathbf{1}^{\text{T}}\mathfrak{d}_{\text{mass}} = \big(\mathbf{1}^{\text{T}}\mathfrak{C}_{\text{mom}}\big)\mathbf{1} = 0, \tag{62a,b}$$

showing discrete mass conservation. The conditions under which our forms for mass and momentum transport do satisfy relation (62a) can be found starting from Eq. (56):

$$\mathfrak{C}_{\text{mom}}\mathbf{1} = (\alpha+\beta)\mathfrak{D}^{\rho u}\mathsf{R}\mathsf{u} + (\gamma+\delta+\varepsilon)(\mathsf{U}\mathfrak{D}^{\rho}\rho + \mathsf{R}\mathfrak{D}^{\mathsf{u}}\mathsf{u}).$$

A comparison with $\mathfrak{d}_{\text{mass}}$ in Eq. (47) shows that Eq. (62a) is satisfied if and only if

$$\alpha+\beta = \xi \quad \text{and} \quad \gamma+\delta+\varepsilon = 1-\xi. \tag{63}$$

These are conditions that we have seen earlier in Eq. (11) for the analytic case.

In the next (sub)sections we study how the freedom in discretization choices that is still open after requiring conservation of mass and momentum can be exploited to achieve conservation of energy. We will first restrict ourselves to two-point fluxes, i.e. three-point stencils for the discrete difference operators.

Staggered grid

In the above formulation of (59) it has been assumed that both conservation equations, i.e. for mass and for momentum, use the same control volume $\mathfrak{H}$. For a staggered positioning both control volumes are different, and the density and mass flux are defined at other positions than in a centered grid. To bring the two terms in (59) together will require an interpolation between the staggered and centered versions of the quantities involved. For example, an equal-weighted $\frac{1}{2}$-$\frac{1}{2}$ interpolation $\mathcal{I}_{1/2}$ of the product $\mathfrak{H}\rho$ and of the components of m suffices to generate staggered quantities for the momentum equation that leave the above reasoning intact [9,47]:

$$(\mathfrak{H}\rho)_{\text{mom}} = \mathcal{I}_{1/2}(\mathfrak{H}\rho)_{\text{mass}} \quad \text{and} \quad \mathsf{m}_{\text{mom}} = \mathcal{I}_{1/2}\mathsf{m}_{\text{mass}}$$

(the subscripts refer to the equation concerned). Obviously, the notation for staggered grids becomes more complex, but the notation from [47] has been designed sufficiently general to allow application to (collocated and staggered) unstructured grids.

5.2. Finite-volume methods

Global energy conservation

While studying advective momentum conservation in the previous section, still a choice had to be made for the computation of the fluxes $m\phi$: interpolation of the product, product of the interpolations or a more complicated expression. The condition (60) to assure global energy conservation will determine this choice, as we will see now.

The former option, in which $(m\phi)_f$ is found from an interpolation of the product $m\phi$ in the adjacent nodes, corresponds with a flux given by (compare Eq. (30))

$$(m\phi)_{i+1/2} = \frac{1}{2}(m_i\phi_i + m_{i+1}\phi_{i+1}) \leftrightarrow (\mathsf{M}\phi)_f = \frac{1}{2}(\mathsf{I}+\mathsf{E})\mathsf{M}\phi. \tag{64}$$

It leads to a discretization for which the advection matrix has a zero diagonal, as in the usual central finite-difference discretizations. Hence, it cannot cancel the contribution from the mass transport in $\mathfrak{A}_{\text{kin}}^{\text{fv}}$ as required by Eq. (60). Of course,

in incompressible flow, where $\mathfrak{d}_{\text{mass}} = 0$, this discretization is fine and does satisfy Eq. (60), i.e. it conserves (kinetic) energy. But for compressible flow the fluxes must be computed in a different way.

Thus, we consider the second interpolation option $m_f \phi_f$, leading to

$$\mathfrak{C}^{\text{fv}}_{\text{mom}}\phi \equiv (\mathsf{I} - \mathsf{E}^{-1})\mathfrak{F}^{\text{fv}}\phi \quad \text{with flux} \quad \mathfrak{F}^{\text{fv}}\phi = \mathsf{M}^{\text{fv}}_f \phi_f, \tag{65}$$

where $\mathsf{M}^{\text{fv}}_f = \text{diag}(\mathsf{m}^{\text{fv}}_f)$. Outside the diagonal the operator $\mathfrak{C}^{\text{fv}}_{\text{mom}}$ has to be skew symmetric according to Eq. (60). This is only possible when in the interpolation the coefficients of neighboring nodes contain some symmetry. The simplest (linear) option is to choose the face values of ϕ as

$$\phi_{i+1/2} = \frac{1}{2}(\phi_i + \phi_{i+1}) \quad \leftrightarrow \quad \phi_f \equiv \frac{1}{2}(\mathsf{I} + \mathsf{E})\phi. \tag{66}$$

For three-point discretization stencils this is the *only possible* choice leading to skew-symmetry [16], and is in line with the interpretation of the nodal values as averages over the control volumes [18]. Observe that this interpolation also allows for a discrete product rule:

$$\Phi_f(\mathsf{E} - \mathsf{I})\psi + \Psi_f(\mathsf{E} - \mathsf{I})\phi = (\mathsf{E} - \mathsf{I})\Phi\psi.$$

Interpolations with larger and more complicated stencils to build higher-order approximations are discussed in Section 6.

The choices (65) and (66) result in the advection operator

$$\mathfrak{C}^{\text{fv}}_{\text{mom}} = \frac{1}{2}(\mathsf{I} - \mathsf{E}^{-1})\mathsf{M}^{\text{fv}}_f(\mathsf{I} + \mathsf{E}) = \frac{1}{2}(\mathsf{M}^{\text{fv}}_f\mathsf{E} - \mathsf{E}^{-1}\mathsf{M}^{\text{fv}}_f) + \frac{1}{2}\,\text{diag}\big((\mathsf{I} - \mathsf{E}^{-1})\mathsf{m}^{\text{fv}}_f\big), \tag{67}$$

where the identity (14b) has been used. Observe the term on the diagonal which equals half the mass transport (28). The importance of the equal-weighted $\frac{1}{2}$-$\frac{1}{2}$ interpolation in Eq. (66) is clearly visible, independent of the geometric position of the face $i + 1/2$ in between the nodes i and $i + 1$ (see also [48, Th. 2.2]). The operator $\mathfrak{A}^{\text{fv}}_{\text{kin}}$ governing discrete energy conservation now follows as

$$\mathfrak{A}^{\text{fv}}_{\text{kin}} = \mathfrak{C}^{\text{fv}}_{\text{mom}} - \frac{1}{2}\,\text{diag}(\mathfrak{d}^{\text{fv}}_{\text{mass}}) = \frac{1}{2}(\mathsf{M}^{\text{fv}}_f\mathsf{E} - \mathsf{E}^{-1}\mathsf{M}^{\text{fv}}_f). \tag{68}$$

Its obvious skew-symmetry satisfies the condition (60) for global energy conservation.

Example. As an example, an appropriate central finite-volume discretization uses the mass flux

$$m_{i+1/2} \equiv \frac{1}{2}(m_i + m_{i+1}) \quad \overset{(66)}{\leftrightarrow} \quad \mathsf{M}^{\text{fv}}_f \phi_f = \frac{1}{4}\,\text{diag}\big((\mathsf{I} + \mathsf{E})\mathsf{m}\big)(\mathsf{I} + \mathsf{E})\phi. \tag{69}$$

In more familiar grid-point terminology, this discretization reads

$$\nabla \cdot (m\phi)|_i \approx \frac{(m_{i+1} + m_i)\phi_{i+1} + (m_{i+1} - m_{i-1})\phi_i - (m_i + m_{i-1})\phi_{i-1}}{2(x_{i+1} - x_{i-1})}.$$

On the diagonal one recognizes half the discrete mass transport. □

Remark. It is emphasized that an unequal-weighted average in the interpolation (66) would lead to an upwind/downwind-biased discretization of the advective term. As is well-known, this influences the kinetic energy; consistent herewith, the skew-symmetry of the operator $\mathfrak{A}_{\text{kin}}$ is lost and condition (60) for energy preservation is not satisfied.

Local energy conservation

Local conservation of energy can be studied by rewriting the evolution of kinetic energy, starting from Eq. (68) and using Eq. (14b), as

$$\mathfrak{H}\frac{\mathrm{d}}{\mathrm{d}t}(\frac{1}{2}\mathsf{R}\Phi\phi) = -\Phi\mathfrak{A}^{\text{fv}}_{\text{kin}}\phi = -\frac{1}{2}(\mathsf{I} - \mathsf{E}^{-1})\mathsf{M}^{\text{fv}}_f\Phi\mathsf{E}\phi, \tag{70}$$

which constitutes a further decomposition from which the 'difference of fluxes' term $(\mathsf{I} - \mathsf{E}^{-1})$ emerges. This decomposition shows that, next to global energy conservation, discrete energy is also locally conserved. As a special case, with the above choice (66), at the face $i + 1/2$ it has the energy flux function $\big(\frac{1}{2}\rho u\phi^2\big)_{i+/2} = \frac{1}{2}m_{i+1/2}\phi_i\phi_{i+1}$. This choice for the kinetic energy flux equals the one used by e.g. Subbareddy and Candler [13, Eq. (6)], Chandrashekar [17, Section 4.7(1)] and Kuya et al. [8, Eqs. (32, 33)].

A comparison of the finite-volume discretizations for mass Eq. (28), momentum Eq. (65) and energy Eq. (70) reveals that their advective terms all fit in the same discrete framework:

$$\nabla\cdot(\boldsymbol{m}\psi) \leftrightarrow (\mathsf{I}-\mathsf{E}^{-1})\mathsf{M}_f^{\mathrm{fv}}\psi \quad \text{with} \quad \psi\in\{\mathbf{1},\ \frac{1}{2}(\mathsf{I}+\mathsf{E})\phi,\ \frac{1}{2}\Phi\mathsf{E}\phi\}. \tag{71}$$

In fact, this framework describes *all* three-point energy-preserving finite-volume discretizations. The mass flux $m_{i\pm1/2} = (\rho u)_{i\pm1/2} \leftrightarrow \mathsf{M}_f^{\mathrm{fv}}$ may be chosen completely arbitrary without loosing (local and global) discrete conservation, as long as it is done consistently over all conservation equations. For instance, a directional upwind-like bias of the mass flux as in Eq. (34) is allowed without jeopardizing discrete energy conservation. Also, this freedom can be used to achieve other properties, for example to induce entropy conservation [17, Section 4.7(2)]. Note that, as a finite-volume method, mass and momentum are always locally conserved from the start, although possibly in an nonphysical way when 'exotic' choices for the mass flux are made.

Observation 7 *(Discrete conservation of energy - finite-volume). Finite-volume methods for transport equations, which by design preserve mass and momentum, also globally and locally preserve kinetic energy if and only if in the discrete momentum equation the advective operator is skew-symmetric outside its diagonal. On the diagonal a consistency with the mass transport operator is required. In particular, the latter must satisfy Eq. (61). For second-order three-point stencils, i.e. two-point fluxes, these properties require that:*

1) *the two-point advective flux in Eq. (65) is written as the product of the mass flux* m *in Eq. (28) and the flux of the transported quantity* ϕ*, or as a more complicated expression.*
2) *the flux of the transported quantity* ϕ *is found from an interpolation which is symmetric in the neighboring nodal values, as in Eq. (66).*

As already mentioned in Section 3, there exists a large freedom in choosing the mass flux m while maintaining energy (and mass and momentum) conservation. This was explicitly observed earlier in [13, p. 1350] and [16]. We will explore this freedom below in Section 6.

5.3. Generalized finite-difference methods

Global energy conservation The conservation of the quadratic invariant kinetic energy is governed by the discrete matrix $\mathfrak{A}_{\mathrm{kin}}$. For global conservation of energy it has to satisfy the condition (60). Starting from the discrete flow equations (47) and (56), and with some rearrangement, we can write this discrete matrix for the family of analytic formulations (9) as

$$\begin{aligned}\mathfrak{A}_{\mathrm{kin}} = {}& \alpha\,\mathfrak{D}^{\rho u}\mathsf{R}\mathsf{U} + (\beta+\varepsilon)\,\mathsf{R}\mathsf{U}\mathfrak{D}^{0} + \gamma\,\mathsf{U}\mathfrak{D}^{\rho}\mathsf{R} + \delta\,\mathsf{R}\mathfrak{D}^{u}\mathsf{U} + (\beta-\frac{1}{2}\xi)\,\mathrm{diag}(\mathfrak{D}^{\rho u}\mathsf{R}u)\\ &+\big(\gamma+\varepsilon-\frac{1}{2}(1-\xi)\big)\,\mathrm{diag}(\mathsf{R}\mathfrak{D}^{u}u) + \big(\delta+\varepsilon-\frac{1}{2}(1-\xi)\big)\,\mathrm{diag}(\mathsf{U}\mathfrak{D}^{\rho}\rho).\end{aligned}$$

At the diagonal of $\mathfrak{A}_{\mathrm{kin}}$, the last three terms cancel for all R and U if and only if

$$\alpha-\varepsilon=\beta=\frac{1}{2}\xi \quad \text{and} \quad \gamma=\delta=\frac{1}{2}(1-\xi)-\varepsilon, \tag{72}$$

where the value of α follows from $\alpha+\beta+\gamma+\delta+\varepsilon=1$. These are the same conditions as found by Coppola et al. [15] for central finite-difference methods of split forms; see Eq. (10). Further, it is observed that these conditions are compatible with the conditions (63) for satisfying $\mathfrak{C}_{\mathrm{mom}}\mathbf{1}=\mathfrak{d}_{\mathrm{mass}}$ if and only if $\varepsilon=0$.

Using $\mathsf{U}\mathsf{R}=\mathsf{R}\mathsf{U}$, one now can write

$$\begin{aligned}\mathfrak{A}_{\mathrm{kin}}+\mathfrak{A}_{\mathrm{kin}}^{\mathrm{T}} = {}& (\frac{1}{2}\xi+\varepsilon)\Big[\big(\mathfrak{D}^{\rho u}+\mathfrak{D}^{0^{\mathrm{T}}}\big)\mathsf{R}\mathsf{U}+\mathsf{R}\mathsf{U}\big(\mathfrak{D}^{\rho u^{\mathrm{T}}}+\mathfrak{D}^{0}\big)\Big]+\\ &+[\frac{1}{2}(1-\xi)-\varepsilon]\Big[\mathsf{U}\big(\mathfrak{D}^{\rho}+\mathfrak{D}^{u^{\mathrm{T}}}\big)\mathsf{R}+\mathsf{R}\big(\mathfrak{D}^{\rho^{\mathrm{T}}}+\mathfrak{D}^{u}\big)\mathsf{U}\Big].\end{aligned}$$

The duality conditions (58) are necessary and sufficient to make this expression for the symmetric part of $\mathfrak{A}_{\mathrm{kin}}$ vanish for all R and U. In combination with condition (72), these conditions ensure global conservation of energy.

Whether or not also momentum is conserved depends on satisfying Eq. (62), which is satisfied under the conditions in Eq. (63). As mentioned above, these conditions imply Eq. (72) when $\varepsilon=0$, in which case both discrete energy and momentum are preserved. Thus, for the split forms (9) with $\varepsilon=0$, discrete energy conservation implies discrete conservation of mass and momentum, while for $\varepsilon\neq0$ momentum is not conserved as we saw earlier.

We explicitly note that the conditions in Eq. (72) have been derived in [15] under the (restrictive) assumption of skew symmetry of the derivative matrices and uniform mesh. The derivation here exposed extends the relevance of these conditions to a more general formulation on non-uniform grids in which the skew symmetry is replaced by the duality conditions (58).

Table 1
Overview of some popular and recent energy-preserving discretizations for split forms. The discretization of momentum transport is given by $(m\phi)_{i+1/2} = \frac{1}{2}m_{i+1/2}(\phi_{i+1}+\phi_i)$.

	ξ	mass flux $m_{i+1/2}$
Feiereisen [11]; Kok [3]; Kuya [8, DQ]	1	$\frac{1}{2}(\rho_{i+1}u_{i+1} + \rho_i u_i)$
KGP [14,32]; Kuya [8, QC]; Singh [44, mKEP]	$\frac{1}{2}$	$\frac{1}{4}(\rho_{i+1} + \rho_i)(u_{i+1} + u_i)$
Coppola [15]	0	$\frac{1}{2}(\rho_i u_{i+1} + \rho_{i+1}u_i)$

Local energy conservation Similar to our study of conservation of linear invariants, a globally energy-preserving finite-difference method also locally conserves energy. This is a direct consequence of the skew symmetry of $\mathfrak{A}^{\mathrm{fv}}_{\mathrm{kin}}$ as we will show next. In Section 2.3 it was mentioned that any skew-symmetric matrix $\mathfrak{A}_{\mathrm{kin}}$ can be written as $\mathfrak{A}_{\mathrm{kin}} = \sum_{k>0}(\mathsf{A}_k\mathsf{E}^k - \mathsf{E}^{-k}\mathsf{A}_k)$. Then, using Corollary A.4, the total energy can be written in a local flux formulation

$$\Phi\mathfrak{A}^{\mathrm{fv}}_{\mathrm{kin}}\phi = \sum_{k>0}\Phi\big(\mathsf{A}_k\mathsf{E}^k - \mathsf{E}^{-k}\mathsf{A}_k\big)\phi = (\mathsf{I}-\mathsf{E}^{-1})\sum_{k>0}\Big(\sum_{h=0}^{k-1}\mathsf{E}^{-h}\Big)\mathsf{A}_k\Phi\mathsf{E}^k\phi,$$

herewith generalizing Eq. (70) and revealing local conservation.

5.4. Relation between finite-difference and finite-volume discretizations

When $\varepsilon = 0$, the above energy-preserving finite-difference methods form a one-parameter family, parameterized with ξ, while $\alpha = \beta = \frac{1}{2}\xi$ and $\gamma = \delta = \frac{1}{2}(1-\xi)$. Taking the duality relations (58) into account, the most general supraconservative, i.e. also preserving mass and momentum [9], family of finite-difference methods reads

$$\begin{aligned}
&\mathfrak{H}\frac{\mathrm{d}\rho}{\mathrm{d}t} + \xi\mathfrak{D}^{\rho\mathsf{u}}\mathsf{R}\mathsf{u} + (1-\xi)\big(\mathsf{R}\mathfrak{D}^{\mathsf{u}}\mathsf{u} - \mathsf{U}\mathfrak{D}^{\mathsf{u}^{\mathsf{T}}}\rho\big) = 0;\\
&\mathfrak{H}\frac{\mathrm{d}\mathsf{R}\phi}{\mathrm{d}t} + \frac{1}{2}\xi\big(\mathfrak{D}^{\rho\mathsf{u}}\mathsf{R}\mathsf{U} - \mathsf{R}\mathsf{U}\mathfrak{D}^{\rho\mathsf{u}^{\mathsf{T}}} + \mathrm{diag}(\mathfrak{D}^{\rho\mathsf{u}}\mathsf{R}\mathsf{u})\big)\phi +\\
&\qquad + \frac{1}{2}(1-\xi)\big(\mathsf{R}\mathfrak{D}^{\mathsf{u}}\mathsf{U} - \mathsf{U}\mathfrak{D}^{\mathsf{u}^{\mathsf{T}}}\mathsf{R} + \mathrm{diag}(\mathsf{R}\mathfrak{D}^{\mathsf{u}}\mathsf{u} - \mathsf{U}\mathfrak{D}^{\mathsf{u}^{\mathsf{T}}}\rho)\big)\phi = 0.
\end{aligned} \tag{73}$$

Effectively, there is freedom to design two derivative operators ($\mathfrak{D}^{\rho\mathsf{u}}$ and $\mathfrak{D}^{\mathsf{u}}$), and to choose one parameter (ξ). In Section 3.4 we have already seen that any finite-difference discretization globally preserving linear invariants can be formulated as a cell-centered finite-volume discretization with linear fluxes, but the converse is doubtful. For instance, for the non-linear logarithmic fluxes of Chandrashekar [17] and Ranocha et al. [49,50] no finite-difference equivalent has been reported yet.

All second-order central finite-difference discretizations of the convective formulations as given in e.g. [7,15,46,51], fit into this framework. Table 1 shows a more specific relation with these existing split forms.

Three-point stencils For linear three-point stencils we can prove equivalence. The discrete operators involved start with three free coefficients, which have to satisfy two conditions: *i)* their row sums vanish; *ii)* their scaling is volume consistent. This leaves one degree of freedom per operator, with which an amount of directionality can be built in: central vs. directionally biased. With two discrete operators and one parameter (ξ) to choose, this gives a three-parameter family of three-point finite-difference discretizations which satisfy our requirements. On the other hand, in Section 4.1 we have seen that the three-point finite-volume discretizations with linear fluxes also form a three-parameter family spanned by the coefficients $c^{(\cdot,\cdot)}$. The two three-parameter families are the same, and the mapping between them is given in Eq. (53) in Section 3.4. Note that in principle the three free parameters may be grid-point dependent, thus in fact we have a $3N$-parameter family.

For, possibly higher-order, methods with larger stencils we can again refer to Eqs. (51) and (53). It is clear that for any choice of the vectors $\mathsf{a}^{\rho\mathsf{u}}_k$, $\mathsf{a}^{\mathsf{u}}_k$ and a^{ρ}_k, corresponding vectors $\mathsf{c}^{(\cdot,\cdot)}$ can be constructed. But the opposite does not need to hold. Eq. (51) shows that not all the possible products $\rho_{i+p}u_{i+q}$ with $p \neq q$ are admissible in the mass flux, if a corresponding generalized advective form is sought. Only products of the variables at nodes x_{i-h} and x_{i+k-h}, with $0 \leq h < k$, are allowed; the latter inequalities form a severe restriction on the possible index combinations. As an example, a numerical flux calculated at node $i+1/2$ with a polynomial interpolation including products of the type e.g. $\rho_{i-2}u_{i-1}$ or $\rho_{i+3}u_{i+1}$ do not have a corresponding finite-difference formulation within the family of advective discretizations (42). Thus, for larger stencils it seems that the finite-volume framework is able to produce more general methods, not achievable within the finite-difference framework we considered (i.e. built from divergence and advective forms).

Observation 8 *(Discrete conservation of energy - general finite-difference). When scaled in a volume-consistent way, the split family of (possibly directionally-biased) finite-difference methods (47) and (56) for the transport equations for mass and momentum, respectively, globally and locally conserve*

1) *mass, if and only if the duality condition (43) is satisfied;*
2) *momentum, if and only if $\varepsilon = 0$ and the extended duality conditions (58) are satisfied;*
3) *energy, if and only if next to the duality conditions (58) also Coppola's conditions (10) for the weights are satisfied.*

When $\varepsilon = 0$, the above three conditions for the split forms are nested in the sense that 3) $\Rightarrow$ 2) $\Rightarrow$ 1). *For more general discretizations, given discrete energy conservation, satisfying (62) is necessary and sufficient for discrete conservation of mass and momentum. In particular, in the case $\varepsilon \neq 0$, discrete energy conservation does not imply momentum conservation. The freedom in the choice of the mass flux which is present in the finite-volume formulation corresponds to the freedom in the choice of the derivative matrices $\mathfrak{D}^{\rho u}$ and $\mathfrak{D}^{u}$ and of the weights in the split forms.*

Remark. In all existing studies of the conservation properties for the split convective formulations, e.g. [7,8,15,46,51,52], central discretization is assumed beforehand. Our analysis shows that also directionally-biased discretizations are allowed, provided they satisfy the duality relations (58).

6. Higher-order fluxes

6.1. Product of interpolations

In Section 5.2 we have already seen that central fluxes of the form $(m\phi)_{i+1/2} = \sum_q \gamma_q (m\phi)_{i+q} = \sum_q \gamma_q m_{i+q}\phi_{i+q}$ will not be energy-preserving for compressible flow, as the flux difference $(m\phi)_{i+1/2} - (m\phi)_{i-1/2}$ gives a skew-symmetric discretization of a first-order derivative, as in Eq. (64). This makes the central coefficient (of ϕ_i) vanish, hence it cannot cancel the discrete mass equation as required by the skew-symmetry condition (60). Only for incompressible flow these discretizations may preserve energy. Pirozzoli [45] calls these DFD (divergence finite difference) schemes, and writes "It is well known that in most cases the DFD scheme leads to rapid nonlinear numerical divergence, and it requires some form of artificial stabilization." This observation is consistent with our theoretical conclusions.

So for higher-order discretizations we have to resort to more intricate expressions for the flux, like $(m\phi)_{i+1/2} = m_{i+1/2}\phi_{i+1/2}$ as in Eq. (65). With the freedom in the mass flux, $m_{i+1/2}$ can be constructed to arbitrary order by including an arbitrary number of neighboring points. But the choice for $\phi_{i+1/2}$ is determined by the skew-symmetry requirement of the discrete convective operator. In Section 5.2 we used an equal-weight interpolation (66) between the two adjacent nodal points, limiting the accuracy to second order. A higher-order extension could be attempted by generalizing the interpolation (66) using more neighboring nodes. For example, we can consider a momentum flux of the form

$$(m\phi)_{i+1/2} = m_{i+1/2}\phi_{i+1/2} = m_{i+1/2}\sum_q \alpha_q \phi_{i+q} \quad \text{with} \quad \sum_q \alpha_q = 1. \tag{74}$$

The diagonal of $\mathfrak{C}^{\mathrm{fv}}_{\mathrm{mom}} = (\mathsf{I}-\mathsf{E}^{-1})\mathsf{M}^{\mathrm{fv}}_f$, i.e. the coefficient of ϕ_i, becomes $m_{i+1/2}\alpha_0 - m_{i-1/2}\alpha_1$. Demanding energy conservation this expression has to be half a consistent discretization of the mass transport term (29), i.e. $\frac{1}{2}(m_{i+1/2} - m_{i-1/2})$. Hence $\alpha_0 = \alpha_1 = 1/2$, while all other α_q's sum up to zero. As there does not exist a higher-order (> 2) interpolation for ϕ in which the coefficients satisfy the just-mentioned criteria, using the momentum flux format of product type (74) a second-order accuracy is the most that can be achieved.

Thus, to achieve higher-order accuracy even more complicated forms of the momentum flux have to be sought. The split forms which have been studied and reviewed by various authors [7,8,14,15,45,51,52] fall in this category, and are the ones considered here. The mentioned papers pay particular attention to the split forms of the mass flux. Our theory shows that this mass splitting is not relevant for discrete energy preservation, as long as it is done consistently in both the mass and momentum equations. Of course these splittings may influence accuracy, but that is a separate study.

6.2. Richardson extrapolation

A particular example of a higher-order discretization has been proposed by Verstappen and Veldman [5,6], thus far limited to incompressible flow. They construct the flux by a Richardson extrapolation in combination with a two- or three-times coarser grid (for collocated and staggered grids, respectively). We generalize this approach for compressible flow. It starts with the standard energy-preserving central finite-volume discretization (69) on the basic fine grid, to be combined with the same discretization on a, here two-times, coarser grid. In our matrix-vector notation:

$$\text{fine grid:} \qquad \mathfrak{H}^{\mathrm{fine}}\frac{\mathrm{d}\mathsf{R}\phi}{\mathrm{d}t} = -\,\mathfrak{C}^{\mathrm{fine}}_{\mathrm{mom}}\phi = -\frac{1}{2}(\mathsf{I}-\mathsf{E}^{-1})\,\mathsf{M}^{\mathrm{fine}}_f\,(\mathsf{I}+\mathsf{E})\phi; \tag{75a}$$

$$\text{coarse grid:} \qquad \mathfrak{H}^{\mathrm{crse}}\frac{\mathrm{d}\mathsf{R}\phi}{\mathrm{d}t} = -\,\mathfrak{C}^{\mathrm{crse}}_{\mathrm{mom}}\phi = -\frac{1}{2}(\mathsf{I}-\mathsf{E}^{-2})\,\mathsf{M}^{\mathrm{crse}}_f\,(\mathsf{I}+\mathsf{E}^{2})\phi, \tag{75b}$$

where the $\mathsf{M}^{(\cdot)}_f$ are coarse and fine mass fluxes which are free to choose, for example the second-order choices

$$\mathsf{M}_f^{\text{fine}} = \frac{1}{2}\,\text{diag}[(\mathsf{I}+\mathsf{E})\mathsf{m}] \quad \text{and} \quad \mathsf{M}_f^{\text{crse}} = \frac{1}{2}\,\text{diag}[(\mathsf{I}+\mathsf{E}^2)\mathsf{m}]. \tag{76a,b}$$

The final discretization is formed by a Richardson extrapolation that cancels the leading terms in the truncation error (which is of third-order, hence the power 3 below):

$$\left[2^3 * \text{Eq. (75a) - Eq. (75b)}\right]/6. \tag{77}$$

In the left-hand side, the above extrapolation is similar to turning a trapezoidal quadrature rule into Simpson's quadrature, which makes also the volume integration of the time derivative 4th order. We obtain a finite-volume method over an effective control volume of size

$$\mathfrak{H}^{\text{4th}} = (8\,\mathfrak{H}^{\text{fine}} - \mathfrak{H}^{\text{crse}})/6, \tag{78}$$

which shows fourth-order accuracy on smooth grids [6]. Observe that this size of the control volume, found from a Richardson extrapolation, equals (25) found from defining $\mathfrak{H} = \text{diag}(\mathfrak{D}\mathsf{x})$. Because $\mathsf{I}-\mathsf{E}^{-2} = (\mathsf{I}-\mathsf{E}^{-1})(\mathsf{I}+\mathsf{E}^{-1})$, the 'coarse' flux (75b)+(76b) can also be written as a 'fine' flux. Taking the combination (77), this results in a flux of the 4th-order finite-volume method as

$$(m\phi)^{\text{4th}}_{i+1/2} = \frac{1}{3}(m_{i+1}+m_i)(\phi_{i+1}+\phi_i) - \frac{1}{24}(m_{i+2}+m_i)(\phi_{i+2}+\phi_i) - \frac{1}{24}(m_{i+1}+m_{i-1})(\phi_{i+1}+\phi_{i-1}).$$

This flux equals Pirozzoli's [14] flux in his Eq. (13) with $L=2$, $a_1 = 2/3$ and $a_2 = -1/12$. The difference is that the latter flux is used in a finite-difference context, hence the treatment of the time derivative differs (but this opens up another discussion [21,22]).

Observation 9 *(Higher-order fluxes). With fluxes of the form $(m\phi)_{i+1/2} = m_{i+1/2}\phi_{i+1/2}$ energy-preserving discretizations are at most second-order accurate. Higher-order finite-volume discretizations require more complicated expressions for the flux.*

7. Numerical examples for the transport equations

7.1. The test problem

To illustrate and further study the above theoretical considerations we present a simple model problem. The governing equations are given by (1) with $\phi = u$. When not otherwise indicated, in the analytical formulations (9) we set $\xi = 1/2$ where all four operators $\mathfrak{D}^{(\cdot)}$ are involved, whereas the parameters $\alpha, \cdots, \delta$ are chosen according to (10).

A periodic domain $x \in [0,1]$ is selected, during a time interval $t \in [0, 1.0]$ unless indicated otherwise. Initial conditions are given by

$$u(x,0) = 1 + 0.1\sin(2\pi x) \quad \text{and} \quad (\rho u)(x,0) = 2 + \sin(2\pi x),$$

which leads to a smooth, oscillating solution in space and time. The grids in the examples are either uniform or smoothly stretched, based on the transformation

$$x = 0.5 + 0.5\frac{\tanh(s(\sigma-0.5))}{\tanh(0.5\,s)},$$

where the computational coordinate $\sigma \in [0,1]$ is discretized uniformly and the parameter s controls the stretching. When $s > 0$ the grids have refinement regions near the boundaries $x = 0$ and $x = 1$. This is to keep the grid consistent with periodic boundary conditions; in particular, we do not want an abrupt change in mesh size when switching from one side boundary to the other side. In the simulations shown we used $s = 5$, to illustrate the influence of a non-uniformity of a grid. In these grids the coarsest mesh is up to 37 times larger than the smallest one. The time integration is a 4th-order Runge–Kutta method, with a very small time step (typically below 10^{-5}), to make sure that the time-integration error is smaller than the spatial discretization error we are focusing on.

In the examples to follow, several energy-preserving discretizations have been studied that satisfy the duality conditions in Eq. (58):

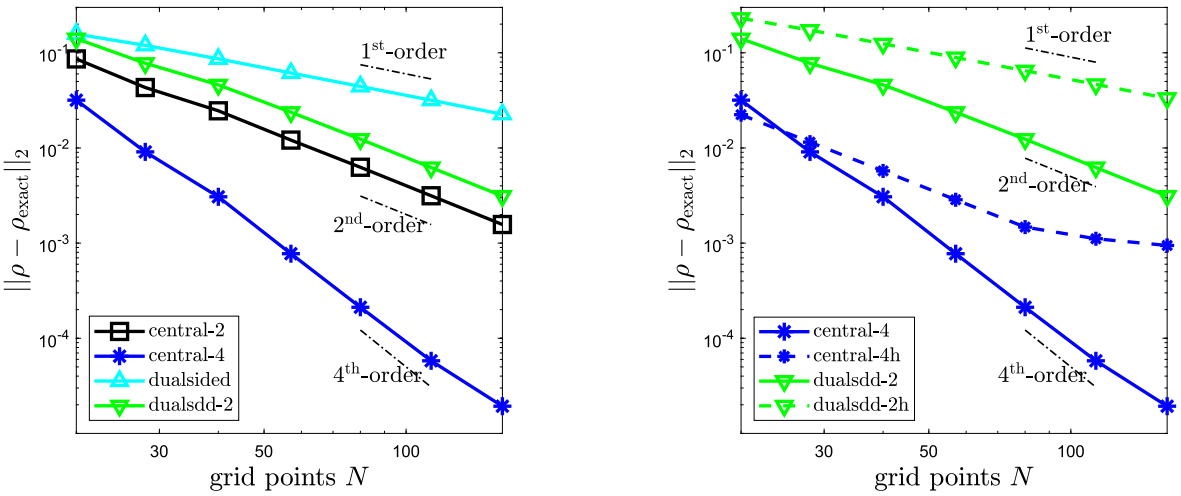
$$\text{central:}\quad \mathfrak{D}^{\rho u} = \mathfrak{D}^{\rho} = \frac{1}{2}(\mathsf{E}-\mathsf{E}^{-1}); \tag{79a}$$

$$\text{4th-order central:}\quad \mathfrak{D}^{\rho u} = \mathfrak{D}^{\rho} = \frac{1}{12}(-\mathsf{E}^2+8\mathsf{E}-8\mathsf{E}^{-1}+\mathsf{E}^{-2}); \tag{79b}$$

$$\text{dual-sided:}\quad \mathfrak{D}^{\rho u} = \mathfrak{D}^{\rho} = \mathsf{I}-\mathsf{E}^{-1}; \tag{79c}$$

$$\text{2nd-order dual-sided:}\quad \mathfrak{D}^{\rho u} = \mathfrak{D}^{\rho} = \frac{1}{2}(3\mathsf{I}-4\mathsf{E}^{-1}+\mathsf{E}^{-2}), \tag{79d}$$

with $\mathfrak{D}^0 = \mathfrak{D}^u = -(\mathfrak{D}^{\rho u})^{\mathsf{T}}$. Also, we will discuss the behavior of the traditional upwind discretization, which does not satisfy the duality relations Eq. (58).



(a) $\mathfrak{H}$ from Eq. (23). (b) $\mathfrak{H}$ from local h.

Fig. 1. Grid-refinement study for the density of various discretization methods on smoothly-stretched grids with N grid points. In (a) the control volumes are chosen according to Eq. (23), confirming the theoretical predictions. In (b) the control volumes are taken equal to the local grid size h, indicated by the extension "h", which leads to much poorer convergence behavior.

7.2. Grid-refinement study

Before focusing on the conservation properties, we first show a (traditional) grid-refinement study to show that the order of convergence of the methods discussed behaves as expected. Fig. 1 presents such a grid refinement study of the various discretizations in (79) for the smoothly-stretched non-uniform grid described above, with N the number of grid points. The equations are solved until $t = T = 0.1$, after which the solution $\rho(x, T)$ is compared in the L_2-norm with the solution on a very fine similarly-stretched grid with 5000 grid points.

In the simulations, the control volumes $\mathfrak{H}$ are first taken equal to Eq. (23), $\mathfrak{H} = \mathrm{diag}(\mathfrak{D}\mathsf{x})$, making sure that the discretization is exact for linear functions. Fig. 1(a) shows results for central discretizations of 2nd order (79a) and 4th order (79b), and 1st-order (79c) and 2nd-order (79d) dual-sided discretizations. The graphs show that the convergence behavior of energy-preserving discretizations on smooth non-uniform grids is similar to the behavior on uniform grids. This matches our discussion in Section 2.5, and is in line with the (underappreciated) theory of Manteuffel and White [53].

On non-uniform grids, the choice of $\mathfrak{H}$ is essential, as shown in Fig. 1(b) where $\mathfrak{H}$ is chosen according to the local grid size $\mathfrak{H} \leftrightarrow h = \frac{1}{2}(x_{i+1} - x_{i-1})$. The convergence rate for the higher-order methods considered, a 4th-order central discretization and a 2nd-order dual-sided discretization, deteriorates to roughly first order when the grid is non-uniform. This can be explained from Eq. (23), as $\mathrm{d}x/\mathrm{d}\xi$ uses a lower-order discretization than $\mathrm{d}\phi/\mathrm{d}\xi$. As referred to in Section 6.2, this difference is related to the different formal accuracy when interpreting a higher-order discretization as finite volume (where $\mathfrak{H}$ is given by (78)) or as finite difference (where $\mathfrak{H}$ equals the local grid size) [21,22].

7.3. Duality relations

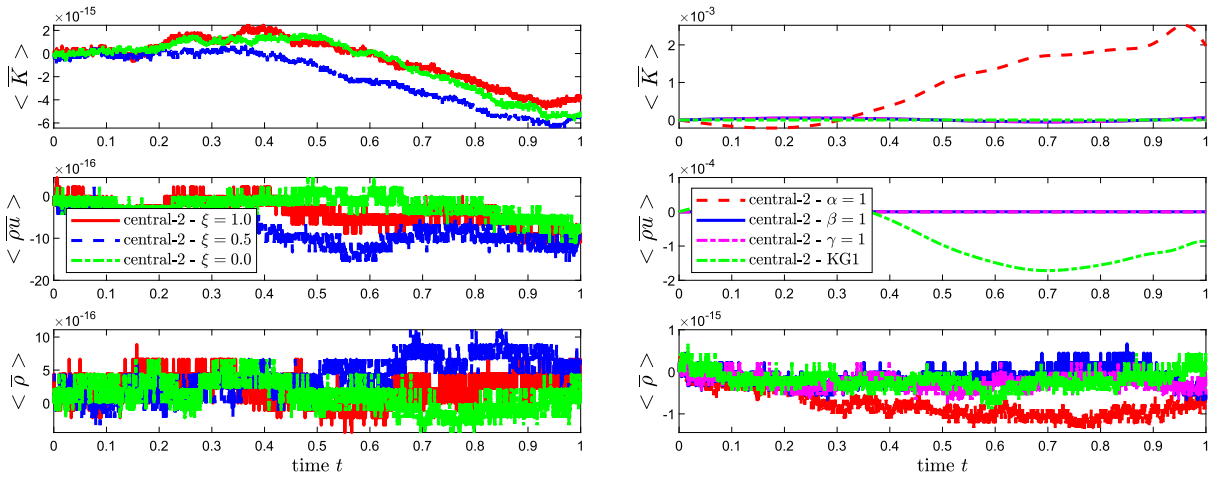
Our next examples study the relevance of the duality relations featuring in Observations 4, 6 and 8, where it is stated that discretizations should satisfy them in order to be conservative. More specifically, for mass conservation (43) needs to be satisfied, while for momentum and energy conservation additionally (58) is required.

The figures below show the discrete evolution of the invariants mass, momentum and energy of the transport equation (1) for special members of the ξ-family (9), on a grid with $N = 21$ grid points, over the time interval until $T = 1$. The global invariants have been normalized with respect to the initial value:

$$\langle f(t) \rangle = \frac{\overline{f}(t) - \overline{f}(0)}{\overline{f}(0)}, \tag{80}$$

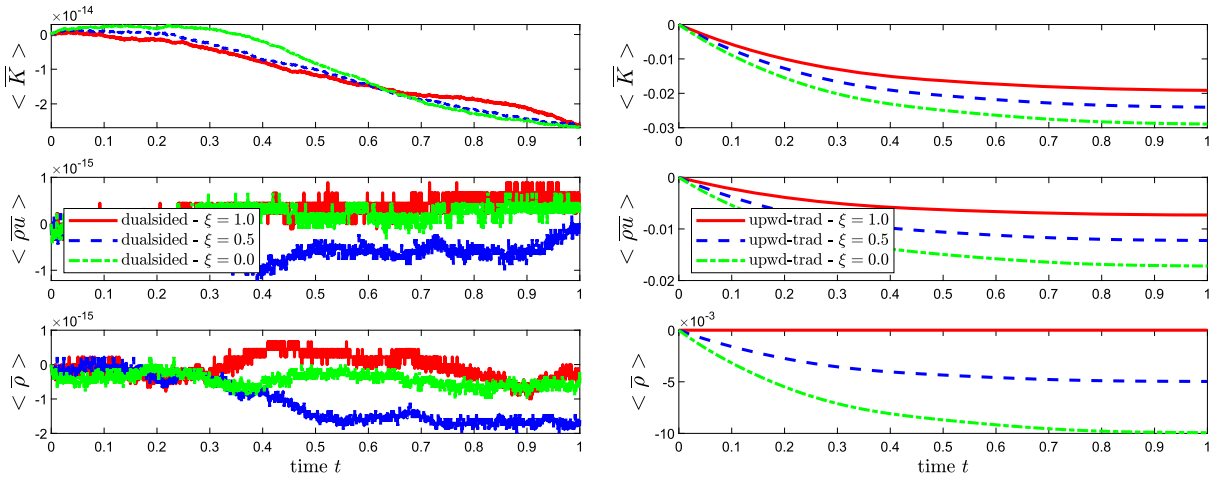
where the overbar indicates spatial integration over the domain.

Central discretization Skew-symmetric central discretizations for all derivatives in the split forms (9) with $\varepsilon = 0$ satisfy both duality conditions, hence they lead to mass and momentum conservation for all values of $\alpha, \cdots, \delta$. For energy conservation additionally the conditions (10) need to be satisfied. Fig. 2(a) shows the discrete evolution of mass, momentum and energy for the special split forms $\xi \in \{0, 1/2, 1\}$, with the parameters $\alpha, \cdots, \delta$ chosen according to (10), The graphs show that these analytic invariants are discretely preserved within machine accuracy.



(a) Central discretization for various ξ. (b) Central discretization not satisfying (10).

Fig. 2. Evolution of the invariants mass, momentum and energy for central discretization, at varying values of the parameters $\alpha, \ldots, \delta$. (a) The parameters are chosen according to (10) for several values of ξ. (b) When these parameters do not satisfy (10), energy preservation is no longer guaranteed.



(a) Dual-sided (58) discretization. (b) Traditional upwind discretization.

Fig. 3. Evolution of mass, momentum and energy for directionally-biased discretizations at various values of ξ. (a) The discretization satisfies the duality conditions (58): all invariants are preserved. (b) All derivatives are chosen equal to the traditional upwind discretization (79c) and do not satisfy the duality conditions: most invariants are no longer discretely preserved.

In Fig. 2(b) the weight parameters have been chosen not to satisfy (10): $\{\alpha = 1,\ \beta = \gamma = \delta = 0\}$, $\{\beta = 1,\ \alpha = \gamma = \delta = 0\}$ and $\{\gamma = 1,\ \alpha = \beta = \delta = 0\}$, respectively. The case $\delta = 1$ turned out not always to be stable over the time interval considered. As predicted theoretically, mass and momentum are preserved discretely. But as these parameter choices do not satisfy the conditions (10), energy preservation is not guaranteed anymore. Also a discretization promoted by Kennedy and Gruber [32] has been added with a non-zero value of ε. In this case, called KG1, the parameters are chosen as $\varepsilon = 0.5,\ \xi = 0.0$, with the other parameters satisfying the conditions for energy preservation (10). It clearly does not preserve momentum.

Directionally-biased discretization Directionally-biased discretizations, like the well-known upwind discretization, in general will not satisfy the duality conditions (58) and discrete conservation is not guaranteed. However, we should not confuse upwind discretization with the dual-sided discretization that does satisfy these duality relations. The latter discretization does preserve kinetic energy. We will next demonstrate this essential difference for our test problem.

To start with, Fig. 3(a) shows results obtained with the dual-sided discretization (79c) which does satisfy the duality relations (58); the whole ξ-family is considered. As predicted by Observation 8, all invariants, including kinetic energy, are preserved discretely.

The situation changes when all four discrete first-order derivative operators $\mathfrak{D}^{(\cdot)}$ are chosen equal to the traditional upwind discretization (79c), as these do not satisfy the duality relations. Fig. 3(b) shows the resulting evolution of the primary

and secondary invariants: most of them are no longer preserved. This is to be expected as it is well-known that upwind discretizations generate artificial diffusion. Only the Feiereisen case $\xi = 1$ conserves mass, which can be understood since then the mass equation is in pure divergence form (7a), for which the upwind derivative gives a conservative discretization.

But when a dual-sided discretization is used, both positive and negative diffusion are generated which neutralize each other. In the mass equation such dual-sided discretizations are not unusual in the study of the reduced Navier–Stokes equations. When accompanied by dual (adjoint) one-sided discretizations of the pressure gradient [54,55], they do not disturb the energy balance. In fact, such dual discretizations are used to avoid the odd-even decoupling of collocated schemes in the pressure Poisson stencil, while at the same time preserving kinetic energy [24].

8. Euler equations

As a 'proof-of-the-pudding', we will now demonstrate the performance of our energy-preserving discretizations when solving the compressible Euler equations under subsonic, shock-free conditions. The combination with shock-capturing discretizations is not yet pursued in this paper, but rather postponed to future publications. The simple test proposed in this section is to be intended as a preliminary application of the present theory to the full system of compressible flow equations. The treatment of the thermodynamic terms will follow the guide lines set out in the *Requirements* in [9] and the *Analytical Relations* in [8]. Further, we adhere to the principle of non-interference [16], with advective terms and thermodynamic terms each having their own energy-preserving role. In particular, we do not consider special combinations of these terms satisfying energy preservation as discussed in Remark 4.1 of [49].

The test case demonstrates that locally conservative and (locally) kinetic energy-preserving discretizations of the compressible flow equations can be obtained in a more general setting by using non skew-symmetric derivative operators. A complete study of the possibilities opened by the general theory developed in the previous sections is out of the scope of the present treatment and will constitute the object of future investigations; see e.g. [10,56]. Also, solution methods for the closely-related shallow-water equations fit in our theoretical framework [57,58].

8.1. Discrete formulation

There exist several ways to formulate the thermodynamic terms in the compressible Euler equations [10,56]; we choose the formulation based on the direct discretization of the internal energy, as used by, e.g., Moin *et al.* [59], Blaisdell *et al.* [60], Veldman [9] and De Michele and Coppola [10]

$$\frac{\partial \rho}{\partial t} = -\nabla \cdot (\rho \boldsymbol{u})\,; \quad \frac{\partial \rho \boldsymbol{u}}{\partial t} = -\nabla \cdot (\rho \boldsymbol{u}\boldsymbol{u}) - \nabla p; \quad \frac{\partial \rho e}{\partial t} = -\nabla \cdot (\rho \boldsymbol{u} e) - p \nabla \cdot \boldsymbol{u}. \tag{81}$$

Here e is the internal energy per unit mass and p is the pressure, obtained from the equation of state $p = (\gamma - 1)\rho e$. Upon semi-discretization, the 1D version of the system (81) can be written as

$$\mathfrak{H}\frac{\mathrm{d}\rho}{\mathrm{d}t} = -\mathfrak{d}; \quad \mathfrak{H}\frac{\mathrm{d}\mathsf{Ru}}{\mathrm{d}t} = -\mathfrak{C}\mathsf{u} - \mathfrak{D}^{\mathsf{p}}\mathsf{p}; \quad \mathfrak{H}\frac{\mathrm{d}\mathsf{Re}}{\mathrm{d}t} = -\mathfrak{C}\mathsf{e} - \mathsf{P}\mathfrak{D}^{\mathsf{e}}\mathsf{u} \tag{82}$$

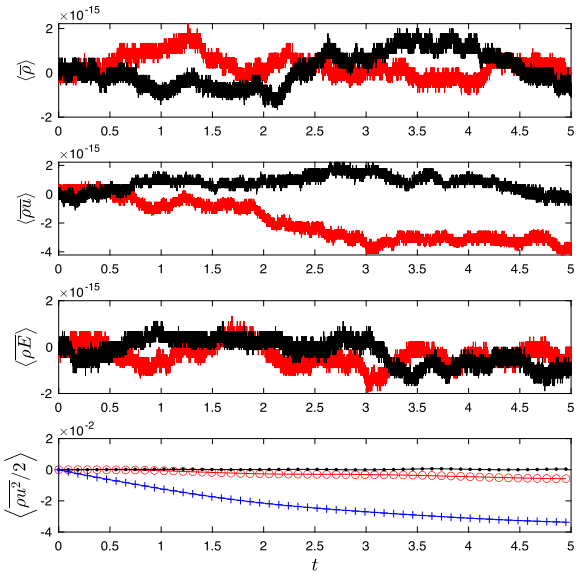
with the usual meaning of the symbols. Although directly discretizing internal energy, this formulation can be easily shown to preserve also total energy, when a kinetic energy-preserving discretization for mass and momentum has been adopted and the additional duality relation for the pressure terms $\mathfrak{D}^{\mathsf{p}} = -(\mathfrak{D}^{\mathsf{e}})^{\mathsf{T}}$ is satisfied [9,10]. In fact, the convective terms are treated by assuming a finite-difference kinetic energy-preserving discretization for $\mathfrak{d}$ and $\mathfrak{C}$:

$$\mathfrak{d} = \xi\, \mathfrak{d}^{D} + (1-\xi)\, \mathfrak{d}^{A}; \qquad \mathfrak{C} = \frac{1}{2}\xi\left(\mathfrak{C}^{D} + \mathfrak{C}^{\phi}\right) + \frac{1}{2}(1-\xi)\left(\mathfrak{C}^{\mathsf{u}} + \mathfrak{C}^{\rho}\right), \tag{83}$$

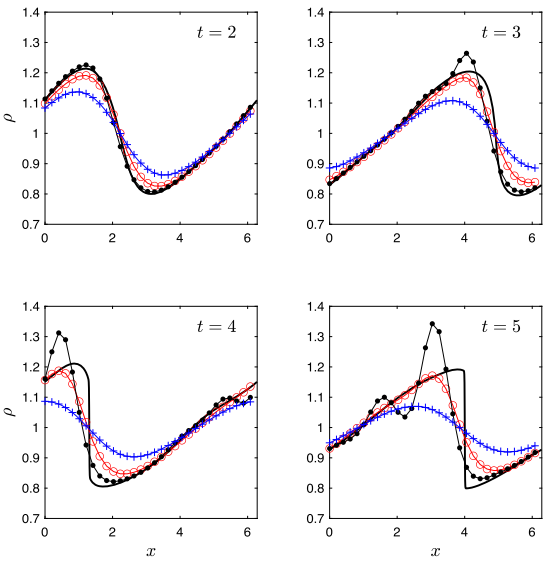
where $\mathfrak{d}^{(\cdot)}$ and $\mathfrak{C}^{(\cdot)}$ are the discretizations of the terms defined in Eqs. (7, 8) (cf. also Eq. (73)) and the value $\xi = 1/2$ is used.

In the test shown below, we use the simple dual-sided discretization obtained by using the derivative matrices $\mathfrak{D}^{\rho} = \mathfrak{D}^{\rho\mathsf{u}} = \mathfrak{D}^{\mathrm{upw}}$, $\mathfrak{D}^{\mathsf{u}} = \mathfrak{D}^{0} = -(\mathfrak{D}^{\mathrm{upw}})^{\mathsf{T}}$ in the formulation defined by Eq. (83), where $\mathfrak{D}^{\mathrm{upw}}$ is the classical first-order upwind derivative matrix $\mathfrak{D}^{\mathrm{upw}} = \mathsf{I} - \mathsf{E}^{-1}$ (see also Eq. (34)). According to the general criteria exposed in Section 3–5, this discretization locally and globally preserves both primary invariants and kinetic energy. To assess the performance of this formulation, we compare our results with two standard discretizations which are analogous to some of the ones considered in the previous sections. In particular, we will use: *(i)* 2nd-order central discretization according to Eq. (79a), and *(ii)* 1st-order upwind discretization in which the upwind derivative matrix $\mathfrak{D}^{\mathrm{upw}}$ is used in all the terms in Eq. (83). The first one is the canonical discretization based on a skew-symmetric derivative matrix, which is known to preserve both linear invariants and kinetic energy, whereas the second one is the standard diffusive upwind method which does not preserve kinetic energy, and is also non conservative of primary invariants when an advective form is used inside the discretization of the convective terms.

To complete the method, the pressure terms have to be discretized. In these preliminary tests we always use the 2nd-order central discretization from Eq. (79a), for both $\mathfrak{D}^{\mathsf{p}}$ and $\mathfrak{D}^{\mathsf{e}}$, for which the duality relation $\mathfrak{D}^{\mathsf{p}} = -(\mathfrak{D}^{\mathsf{e}})^{\mathsf{T}}$ is trivially



(a) Evolution of primary invariants and global kinetic energy in time.



(b) Density profile evolution in time.

Fig. 4. Discretization of the 1D Euler equations for the acoustic wave on uniform mesh with $N = 32$. Dots (black lines): 2nd-order central discretization, circles (red lines): dual-sided discretization, plus symbols (blue lines): 1st-order upwind discretization. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

satisfied. Note that the use of the 2nd-order central discretization in the dual-sided case is coherent with the general guidelines given by the *Requirements* in [9]. According to this theory, in the limit of incompressible flow, the discrete gradient operator acting on p should be equal to minus the transposed divergence in the continuity equation [9, Req. 3.2]. This gives:

$$\mathfrak{D}^{\mathsf{p}} = -\left[\xi\mathfrak{D}^{\rho\mathsf{u}} + (1-\xi)\mathfrak{D}^{\mathsf{u}}\right]^{\mathsf{T}} = \xi\mathfrak{D}^{0} + (1-\xi)\mathfrak{D}^{\rho},$$

As ∇p does not depend on ρ, we can use it also for compressible flow. In the dual-sided case with $\xi = 1/2$ this immediately gives $\mathfrak{D}^{\mathsf{p}} = \mathfrak{D}^{\mathsf{e}} = \mathfrak{D}^{\mathsf{cen}}$.

8.2. Numerical test

The numerical tests are performed on a simple one-dimensional acoustic wave experiencing steepening and nonlinear breakdown. The initial conditions are given by

$$\rho(x,0) = \rho_0(1 + 0.2\sin x), \quad u(x,0) = u_0 + 0.2c_0\sin x, \quad p = p_0 + 0.2\rho_0 c_0^2 \sin x \tag{84}$$

with $u_0 = 1.5$ and $\rho_0 = p_0 = 1$, from which $c_0^2 = \gamma = 1.4$. The equations are spatially discretized on the domain $[0, 2\pi]$ with a uniform mesh and periodic boundary conditions. The semi-discretized equations are integrated in time by using the classical 4th order Runge-Kutta scheme. The Courant number, calculated by using the maximum initial velocity, is set to 10^{-3}, which leads to a time step $\delta t = 1.17 \times 10^{-4}$. This value has been checked to be sufficiently small such that the computations are not affected by temporal errors. The numerical solutions are compared with a high-accuracy reference solution computed with a fifth-order weighted essentially-non-oscillatory (WENO) scheme on a very fine mesh (1024 points).

In Fig. 4 the results of the computations with $N = 32$ grid points are shown. Fig. 4(a) shows the time evolution of the linear global invariants and of the global kinetic energy. The linear invariants mass, momentum and total energy are conserved up to machine precision by both the central and dual-sided discretizations. The conservation properties of the central discretization are well known, the plots in Fig. 4(a) show that also the newly proposed dual-sided discretization exactly conserves the primary invariants.

In Fig. 4 also some results from the standard 1st-order upwind discretization are presented, in which all derivatives are chosen equal to $\mathfrak{D}^{\mathsf{upw}}$ as given in Eq. (34): $\mathfrak{D}^{\rho\mathsf{u}} = \mathfrak{D}^{0} = \mathfrak{D}^{\rho} = \mathfrak{D}^{\mathsf{u}} = \mathfrak{D}^{\mathsf{upw}}$. For the formulation adopted, i.e. $\xi = 1/2$, this discretization neither preserves primary invariants nor global kinetic energy. Therefore the discrete evolution of the invariants corresponding to the upwind discretization are not shown, as their values are out of the scale. At time $t = 5$ the upwind discretization exhibits values of the normalized global mass, momentum and total energy given by -4.5×10^{-3}, -1.2×10^{-2} and -1.5×10^{-2}, respectively. Note that, since the formulation employed directly discretizes the internal energy equation, strictly speaking total energy is not a primary invariant. It is a secondary quantity whose induced discrete

balance equation can be calculated for exact time integration from Eqs. (82) (cf. for example [10]). That the central and the dual-sided discretizations preserve total energy to machine accuracy with a kinetic energy-preserving formulation, as clearly visible in Fig. 4(a), indirectly shows that both discretizations preserve also internal energy. The evolution of global kinetic energy for the proposed schemes is shown in the last panel in Fig. 4(a), where the performance of the dissipative upwind method is also reported.

In Fig. 4(b) the density distribution is shown at selected time instants. After an initial smooth evolution, a discontinuity develops at around $t = 3.5$. The numerical solution computed with the 2nd-order central discretization exhibits large oscillations, as expected, whereas the diffusive 1st-order upwind method smears the discontinuity over a large region. The dual-sided discretization, although being formally 1st order, nicely fits in between the dissipative upwind method and the highly oscillating 2nd-order discretization. Note that the dual-sided discretization illustrated preserves (both locally and globally) primary invariants (mass, momentum and total and internal energies) and kinetic energy, and that in the simulation presented no artificial dissipation has been added. Hence, it could be a good starting point for future research directed to adding special shock operators.

9. Conclusions

In the preceding sections a general framework for supraconservative finite-volume and finite-difference discretizations of transport equations has been presented. The framework is formulated using discrete forms of the equations, in contrast with the usual formulations in analytical split forms. This point of view allows a more abstract and more general study of the (global and local) conservation properties of the equations. In this vein, it generalizes many studies from the literature on conservative split formulations, e.g. [7,8,14,15]. The matrix-vector formulation, enabling decomposition theorems, is a key tool in our analysis. The emphasis is on the (primary and secondary) conservation properties of the advective (transport) terms, and we showed some preliminary results for the full Euler equations. The theoretical study has been carried out under the assumption of exact time integration.

Global versus local With a volume-consistent scaling, the conservation properties can be immediately linked to the discrete matrix operators: global conservation of linear invariants is equivalent with vanishing column sums. Such discrete matrices also can be written in a local flux form by means of a flux-decomposition. This means that every globally conservative discretization is also locally conservative. For split forms to be conservative requires divergence-gradient type duality relations between the discrete differential operators. Skew-symmetric central discretizations satisfy this requirement, and so do dual-sided discretizations. The latter are not to be confused with the traditional upwind discretizations of advection, which do dissipate energy.

Energy preservation As a secondary invariant, to be preserved next to the primary invariants, we have focused on the discrete kinetic energy. Its preservation requires a relation between the discrete advective operator and the discrete equation for mass conservation. In particular, outside its diagonal the advective coefficient matrix should be skew-symmetric, whereas its diagonal defines half the discrete mass transport. This requires that the momentum flux through the faces of the control volume is formulated as the product of the mass flux times the flux of the transported quantity. The latter has to be calculated, independent of the actual geometry, from a $\frac{1}{2}$-$\frac{1}{2}$ arithmetic average of the adjacent nodal values.

In order to achieve energy preservation, the duality relations leading to the preservation of the primary invariants are found necessary. Doing so, the additional conditions for the weights in the split forms (9), derived earlier in [7,15], have been confirmed. Together, these necessary conditions are also sufficient for energy preservation. For the split forms with $\varepsilon = 0$, discrete energy preservation implies discrete conservation of mass and momentum, with the discrete mass operator uniquely defined by the advective operator. Thus, for three-point stencils the discrete advective terms fit in the framework (71). The restrictions put by our supraconservative discretizations are relatively mild. Outside its diagonal the coefficient matrix must be skew-symmetric, while its diagonal should equal half the discrete mass flux which may be completely arbitrary. This leaves a large amount of freedom to obtain additional properties of the discretization, like preserving a dispersion relation [3], and/or achieving accuracy on highly-irregular grids [3,61].

Higher-order Further, it has been demonstrated that extensions to higher-order methods also fit within our theoretical matrix framework; for example the methods obtained from Richardson extrapolation [6]. Moreover, we have shown that a construction of an advective flux as the product of interpolations can at most be second order. To achieve the theoretical order of accuracy, the size of the control volume is found essential: it has to be chosen consistent with the discrete first-order derivative.

Euler equations An application of the proposed theory to the full system of Euler equations, which was the initial motivating reason for this research, is also presented. The transport equations have been extended with thermodynamic effects, by adding a pressure gradient and an equation for the evolution of internal energy; total energy is now a secondary invariant. The discretization of the thermodynamic terms has been chosen to match our supraconservative approach, along the guide lines of [8,9], and directly follows from the discretization of the transport of mass and momentum. Both central and dual-sided discretizations have been explored. We found a correct discrete energy exchange between kinetic and internal energy,

such that total energy is discretely preserved. Moreover, without adding numerical diffusion, the newly developed dual-sided discretizations do a good job in describing the formation of shocks in situations where a central discretization is prone to oscillations.

Generality The described conditions, most of them being necessary and sufficient, have been formulated in matrix properties, and hold irrespective of the discretization method with which the discrete matrices have been created, like finite-differences or finite-elements. As a special consequence, any (linear) conservative discretization of the considered family of split formulations can be re-formulated as a (linear) cell-centered finite-volume discretization. In particular, there exists an equivalence between energy-preserving three-point finite-difference discretizations and three-point finite-volume discretizations with a linear flux: they form a three-parameter family. When solving the Navier–Stokes equations, the energy-preserving methods studied in this paper do not require any numerical diffusion to remain stable. This makes them a popular method in turbulent-flow simulations, where the subtle balance between turbulence production and viscous dissipation is critical [2].

CRediT authorship contribution statement

Gennaro Coppola: Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Arthur E.P. Veldman:** Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

The authors would like to thank the anonymous referees for making several constructive suggestions that helped to improve the presentation of the paper.

Appendix A. Flux decomposition of matrices

In this appendix we will prove some properties of matrices with vanishing column (and row) sums. In particular, it is shown how these matrices can be reformulated in a finite-volume format with a local flux function.

A.1. Local conservation

In a finite-volume method, the discretization is formed by the difference of fluxes through faces halfway between the nodal points, thus it is locally conservative. The next lemma proves that vanishing column sums are a necessary and sufficient condition for local conservation.

Lemma A.1 *(Local conservation). A matrix* $\mathsf{D} = \sum_{k=-L}^{L} \mathrm{diag}(\mathsf{a}_k)\mathsf{E}^k$ *is locally conservative if and only if all column sums vanish. The flux decomposition is given by*

$$\mathsf{D} = (\mathsf{I} - \mathsf{E}^{-1})\mathsf{F} \quad \textit{with} \quad \mathsf{F} = \sum_{k=-L}^{L} \sum_{h=k}^{L} \mathrm{diag}(\mathsf{E}^{k-h}\mathsf{a}_h)\mathsf{E}^k. \tag{A.1}$$

Proof. The proof has two directions:

$\Leftarrow$ When the column sums vanish, we first define new vectors b_k as

$$\mathsf{b}_k = \sum_{h=k}^{L} \mathsf{E}^{k-h}\mathsf{a}_h \quad (k = -L, \ldots, L+1). \tag{A.2}$$

Observe that $\mathsf{b}_{L+1} = 0$ by construction, whereas $\mathsf{b}_{-L} = 0$ because of the vanishing column sums, as $\mathsf{b}_{-L} = -\mathsf{E}^{-L}\sum_{h=-L}^{L}\mathsf{E}^{-h}\mathsf{a}_h$; see (16). These vectors have been constructed to allow discrete summation by parts, as we will see. Now, Eq. (A.2) can be inverted to give

$$\mathsf{a}_k = \left(\mathsf{b}_k - \mathsf{E}^{-1}\mathsf{b}_{k+1}\right) \quad (k = -L, \ldots, L). \tag{A.3}$$

Next, substitution of Eq. (A.3) in the expression for D, and introduction of $\mathsf{B}_k = \mathrm{diag}(\mathsf{b}_k)$, gives

$$\begin{aligned}\mathsf{D} &= \sum_{k=-L}^{L} \left(\mathsf{B}_k - \mathrm{diag}(\mathsf{E}^{-1}\mathsf{b}_{k+1})\right)\mathsf{E}^k = \sum_{k=-L}^{L} \mathsf{B}_k\mathsf{E}^k - \sum_{k=-L}^{L} \mathsf{E}^{-1}\mathsf{B}_{k+1}\mathsf{E}^{k+1} \\ &= \sum_{k=-L}^{L} \mathsf{B}_k\mathsf{E}^k - \mathsf{E}^{-1} \sum_{j=-L+1}^{L+1} \mathsf{B}_j\mathsf{E}^j = (\mathsf{I} - \mathsf{E}^{-1}) \sum_{k=-L}^{L} \mathsf{B}_k\mathsf{E}^k \equiv (\mathsf{I} - \mathsf{E}^{-1})\mathsf{F},\end{aligned} \tag{A.4}$$

where in the last step we used that $\mathsf{B}_{-L} = \mathbf{0} = \mathsf{B}_{L+1}$. Thus D has a locally conservative form with flux given by (A.1).

$\Rightarrow$ When a matrix can be written in flux form (A.1), then its column sums satisfy

$$\mathbf{1}^{\mathrm{T}}\mathsf{D} = \mathbf{1}^{\mathrm{T}}(\mathsf{I} - \mathsf{E}^{-1})\mathsf{F} = \mathbf{0},$$

because E is circulant. □

Remark. Equation (A.3) is a discrete relative of the Main Theorem of calculus: $\frac{\mathrm{d}}{\mathrm{d}x}\int_x^{x_1} a(\eta)\,\mathrm{d}\eta = -a(x)$.

Corollary A.2 *(Circulant matrices). For circulant matrices with vanishing column and row sums, the flux matrix can be written as*

$$\mathsf{F} = \sum_{k=-L}^{L}\sum_{h=k}^{L} a_h\mathsf{E}^k, \quad \textit{leading to} \quad \mathsf{F}\mathbf{1} = \left(\sum_{k=-L}^{L} k a_k\right)\mathbf{1}. \tag{A.5}$$

When the circulant matrix is also skew-symmetric, the flux can be rewritten as

$$\mathsf{F} = \sum_{k=1}^{L} a_k \sum_{h=0}^{k-1} \mathsf{E}^{-h}\left(\mathsf{I} + \mathsf{E}^k\right). \tag{A.6}$$

Proof. For circulant matrices, the flux matrix F given in Eq. (A.1) can be written as

$$\mathsf{F} = \sum_{k=-L}^{L}\sum_{h=k}^{L} a_h\mathsf{E}^k = \sum_{k=-L}^{L} a_k \sum_{h=-L}^{k} \mathsf{E}^h, \tag{A.7}$$

where in the last step the order of the summations has been changed (and the indices renamed). From the middle expression in (A.7) we obtain

$$\mathsf{F}\mathbf{1} = \sum_{k=-L}^{L}\sum_{h=k}^{L} a_h\mathbf{1} \equiv \alpha\mathbf{1}.$$

Written out this equals

$$a_{-L} + 2a_{-L+1} + \ldots + (L+1)a_0 + \ldots + 2La_{L-1} + (2L+1)a_L = \alpha.$$

Because of the vanishing column sums one has $a_0 = -(a_{-L} + a_{-L+1} + \ldots + a_{-1} + a_1 + \ldots + a_L)$. When substituted, this leads to

$$-La_{-L} + (1-L)a_{-L+1} + \ldots + 0a_0 + \ldots + (L-1)a_{L-1} + La_L = \alpha,$$

which is the stated condition (A.5).

When the matrix is also skew-symmetric (i.e. $a_{-k} = -a_k$), by splitting the k-sum in the right-hand side of (A.7) one obtains:

$$\mathsf{F} = \sum_{k=1}^{L} a_k \left(-\sum_{h=-L}^{-k} \mathsf{E}^h + \sum_{h=-L}^{k} \mathsf{E}^h\right) = \sum_{k=1}^{L} a_k \left(\sum_{h=0}^{k-1} \mathsf{E}^{-h}\right)\left(\mathsf{I} + \mathsf{E}^k\right), \tag{A.8}$$

where the identity (for $k > 0$)

$$\sum_{h=-L}^{k} \mathsf{E}^h - \sum_{h=-L}^{-k} \mathsf{E}^h = \sum_{\ell=-k}^{k-1} \mathsf{E}^{-\ell} = \left(\sum_{h=0}^{k-1} \mathsf{E}^{-h}\right)\left(\mathsf{I} + \mathsf{E}^k\right)$$

has been used. □

A.2. Quadratic conservation

We will next derive a discrete product rule, valid for any operator D, not just derivative operators. Also, vanishing column and/or row sums are not required. This product rule can be used to study advective forms of the equations, but it also plays a role in discrete conservation of quadratic forms as we will see below.

Lemma A.3 *(Discrete product rule). For any matrix* $\mathsf{D}=\sum_k \mathsf{A}_k\mathsf{E}^k$ *and for any diagonal matrices* $\Phi\equiv\mathrm{diag}(\phi)$ *and* $\Psi\equiv\mathrm{diag}(\psi)$*, one can write*

$$\Phi\mathsf{D}\psi-\Psi\mathsf{D}^{\mathrm{T}}\phi=(\mathsf{I}-\mathsf{E}^{-1})\mathsf{f}\ \textit{ with flux }\ \mathsf{f}=\sum_{k>0}\left(\sum_{h=0}^{k-1}\mathsf{E}^{-h}\right)\Big(\Phi\mathsf{A}_k\mathsf{E}^k\psi-\Psi\mathsf{E}^k\mathsf{A}_{-k}\phi\Big). \tag{A.9}$$

Proof. Firstly, it is noted that for any diagonal matrices $\mathsf{X}\equiv\mathrm{diag}(\mathsf{x})$ and $\mathsf{Y}\equiv\mathrm{diag}(\mathsf{y})$

$$\mathsf{X}\mathsf{E}^{-k}\mathsf{y}\overset{(14)}{=}\mathsf{E}^{-k}\mathrm{diag}(\mathsf{E}^k\mathsf{x})\mathsf{y}=\mathsf{E}^{-k}\mathsf{Y}\mathsf{E}^k\mathsf{x}. \tag{A.10}$$

Next, we can rewrite the left-hand side of (A.9) by splitting the k-summation in D in a positive and a negative range (note that $k=0$ does not contribute):

$$\begin{aligned}\Phi\mathsf{D}\psi-\Psi\mathsf{D}^{\mathrm{T}}\phi &= \sum_{k>0}\big[\Phi\mathsf{A}_k\mathsf{E}^k\psi-\Psi\mathsf{E}^{-k}\mathsf{A}_k\phi\big]+\sum_{k<0}\big[\Phi\mathsf{A}_k\mathsf{E}^k\psi-\Psi\mathsf{E}^{-k}\mathsf{A}_k\phi\big]\\ &\overset{(A.10)}{=}\sum_{k>0}\big[\Phi\mathsf{A}_k\mathsf{E}^k\psi-\mathsf{E}^{-k}\mathsf{A}_k\Phi\mathsf{E}^k\psi\big]+\sum_{k<0}\big[\mathsf{E}^k\Psi\mathsf{E}^{-k}\mathsf{A}_k\phi-\Psi\mathsf{E}^{-k}\mathsf{A}_k\phi\big]\\ &= \sum_{k>0}(\mathsf{I}-\mathsf{E}^{-k})\mathsf{A}_k\Phi\mathsf{E}^k\psi-\sum_{k<0}(\mathsf{I}-\mathsf{E}^k)\Psi\mathsf{E}^{-k}\mathsf{A}_k\phi\\ &= \sum_{k>0}(\mathsf{I}-\mathsf{E}^{-k})\big[\Phi\mathsf{A}_k\mathsf{E}^k\psi-\Psi\mathsf{E}^k\mathsf{A}_{-k}\phi\big],\end{aligned}$$

where in the last step the summation index has been changed $k\to-k$. Finally, noting the decomposition (valid for $k>0$) $(\mathsf{I}-\mathsf{E}^{-k})=(\mathsf{I}-\mathsf{E}^{-1})\sum_{h=0}^{k-1}\mathsf{E}^{-h}$, the decomposition (A.9) is obtained. □

Next we show that skew-symmetry of a matrix directly leads to local conservation of quadratic forms. Again, there is no need for vanishing column and/or row sums. Note that global conservation of quadratic quantities is immediate: when D is skew-symmetric, then $\phi^{\mathrm{T}}\mathsf{D}\phi=0$ for any ϕ. Local conservation follows from the next corollary.

Corollary A.4 *(Quadratic conservation). For any skew-symmetric matrix* $\mathsf{G}=\sum_{k>0}(\mathsf{A}_k\mathsf{E}^k-\mathsf{E}^{-k}\mathsf{A}_k)$ *and any* $\Phi=\mathrm{diag}(\phi)$ *one can write*

$$\Phi\mathsf{G}\phi=(\mathsf{I}-\mathsf{E}^{-1})\mathsf{f},\quad\textit{with flux }\ \mathsf{f}=\sum_{k>0}\left(\sum_{h=0}^{k-1}\mathsf{E}^{-h}\right)\mathsf{A}_k\Phi\mathsf{E}^k\phi.$$

Proof. We can write the skew-symmetric matrix $\mathsf{G}=\mathsf{D}-\mathsf{D}^{\mathrm{T}}$ with $\mathsf{D}=\sum_{k>0}\mathsf{A}_k\mathsf{E}^k$, where only positive values of the summation index occur. Then Lemma A.3 with $\Psi=\Phi$, and realizing that $\mathsf{A}_k=0$ for $k<0$, concludes the proof. □

References

[1] S. Pirozzoli, Numerical methods for high-speed flows, Annu. Rev. Fluid Mech. 43 (2011) 163–194.
[2] W. Rozema, R.W.C.P. Verstappen, J.C. Kok, A.E.P. Veldman, Low-dissipation simulation methods and models for turbulent subsonic flow, Arch. Comput. Methods Eng. 27 (1) (2020) 299–330, https://doi.org/10.1007/s11831-018-09307-7.
[3] J.C. Kok, A high-order low-dispersion symmetry-preserving finite-volume method for compressible flow on curvilinear grids, J. Comput. Phys. 228 (2009) 6811–6832.
[4] R.W.C.P. Verstappen, A.E.P. Veldman, Direct numerical simulation of turbulence at lesser costs, J. Eng. Math. 32 (1997) 143–159.
[5] R.W.C.P. Verstappen, A.E.P. Veldman, Spectro-consistent discretization: a challenge to RANS and LES, J. Eng. Math. 34 (1998) 163–179.
[6] R.W.C.P. Verstappen, A.E.P. Veldman, Symmetry-preserving discretization of turbulent flow, J. Comput. Phys. 187 (2003) 343–368.
[7] G. Coppola, F. Capuano, L. de Luca, Discrete energy-conservation properties in the numerical simulation of the Navier–Stokes equations, Appl. Mech. Rev. 71 (2019) 010803, https://doi.org/10.1115/1.4042820.
[8] Y. Kuya, K. Totani, S. Kawai, Kinetic energy and entropy preserving schemes for compressible flows by split convective forms, J. Comput. Phys. 375 (2018) 823–853.
[9] A.E.P. Veldman, Supraconservative finite-volume methods for the Euler equations of subsonic compressible flow, SIAM Rev. 63 (2021) 756–779.
[10] C. De Michele, G. Coppola, Numerical treatment of the energy equation in compressible flows simulations, Comput. Fluids 250 (2023) 105709.

[11] W.J. Feiereisen, W.C. Reynolds, J.H. Ferziger, Numerical simulation of a compressible, homogeneous, turbulent shear flow, Report TF-13, Thermosciences Division, Mechanical Engineering, Stanford University, 1981.
[12] A.E. Honein, P. Moin, Higher entropy conservation and numerical stability of compressible turbulence simulations, J. Comput. Phys. 201 (2004) 531–545.
[13] P.K. Subbareddy, G.V. Candler, A fully discrete, kinetic energy consistent finite-volume scheme for compressible flows, J. Comput. Phys. 228 (2009) 1347–1364.
[14] S. Pirozzoli, Generalized conservative approximations of split convective derivative operators, J. Comput. Phys. 229 (19) (2010) 7180–7190.
[15] G. Coppola, F. Capuano, S. Pirozzoli, L. de Luca, Numerically stable formulations of convective terms for turbulent compressible flows, J. Comput. Phys. 382 (2019) 86–104, https://doi.org/10.1016/j.jcp.2019.01.007.
[16] A.E.P. Veldman, A general condition for kinetic-energy preserving discretization of flow transport equations, J. Comput. Phys. 398 (2019) 108894, https://doi.org/10.1016/j.jcp.2019.108894.
[17] P. Chandrashekar, Kinetic energy preserving and entropy stable finite volume schemes for compressible Euler and Navier–Stokes equations, Commun. Comput. Phys. 14 (5) (2013) 1252–1286.
[18] A. Jameson, W. Schmidt, E. Turkel, Numerical solution of the Euler equations by finite volume methods using Runge-Kutta time stepping schemes, AIAA Paper 81-1259, 1981.
[19] K.W. Morton, E. Sūli, Finite volume methods and their analysis, IMA J. Numer. Anal. 11 (2) (1991) 241–260.
[20] E. Süli, The accuracy of cell vertex finite volume methods on quadrilateral meshes, Math. Comput. 59 (200) (1992) 359–382.
[21] B.P. Leonard, Order of accuracy of QUICK and related convection-diffusion schemes, Appl. Math. Model. 19 (11) (1995) 640–653.
[22] H. Nishikawa, The QUICK scheme is a third-order finite-volume scheme with point-valued numerical solutions, Int. J. Numer. Methods Fluids 93 (7) (2021) 2311–2338.
[23] W. Rozema, J.C. Kok, R.W.C.P. Verstappen, A.E.P. Veldman, A symmetry-preserving discretisation and regularisation model for compressible flow with application to turbulent channel flow, J. Turbul. 15 (6) (2014) 386–410.
[24] J. Reiss, A family of energy stable, skew-symmetric finite difference schemes on collocated grids, J. Sci. Comput. 65 (2015) 1–18.
[25] W. Rozema, J.C. Kok, A.E.P. Veldman, R.W.C.P. Verstappen, Numerical simulation with low artificial dissipation of transitional flow over a delta wing, J. Comput. Phys. 405 (2020) 109182, https://doi.org/10.1016/j.jcp.2019.109182.
[26] A. Arakawa, Computational design for long-term numerical integration of the equations of fluid motion: two-dimensional incompressible flow. Part I, J. Comput. Phys. 1 (1966) 119–143.
[27] K. Horiuti, Comparison of conservative and rotational forms in large eddy simulation of turbulent channel flow, J. Comput. Phys. 71 (2) (1987) 343–370.
[28] M.A. Olshanskii, L.G. Rebholz, Velocity–vorticity–helicity formulation and a solver for the Navier–Stokes equations, J. Comput. Phys. 229 (11) (2010) 4291–4303.
[29] S. Charnyi, T. Heister, M.A. Olshanskii, L.G. Rebholz, Efficient discretizations for the EMAC formulation of the incompressible Navier–Stokes equations, Appl. Numer. Math. 141 (2019) 220–233.
[30] H.K. Moffatt, A. Tsinober, Helicity in laminar and turbulent flow, Annu. Rev. Fluid Mech. 24 (1) (1992) 281–312.
[31] I.O. Götze, H. Noguchi, G. Gompper, Relevance of angular momentum conservation in mesoscale hydrodynamics simulations, Phys. Rev. E 76 (4) (2007) 046705.
[32] C.A. Kennedy, A. Gruber, Reduced aliasing formulations of the convective terms within the Navier–Stokes equations for a compressible fluid, J. Comput. Phys. 227 (3) (2008) 1676–1700.
[33] B. Strand, Summation by parts for finite difference approximations for d/dx, J. Comput. Phys. 110 (1994) 47–67.
[34] M. Svärd, J. Nordström, Review of summation-by-parts schemes for initial–boundary-value problems, J. Comput. Phys. 268 (2014) 17–38.
[35] J.B. Perot, Discrete conservation properties of unstructured mesh schemes, Annu. Rev. Fluid Mech. 43 (2011) 299–318.
[36] T.C. Fisher, M.H. Carpenter, J. Nordström, N.K. Yamaleev, C. Swanson, Discretely conservative finite-difference formulations for nonlinear conservation laws in split form: theory and boundary conditions, J. Comput. Phys. 234 (2013) 353–375.
[37] P. Lax, B. Wendroff, Systems of conservation laws, Commun. Pure Appl. Math. 13 (2) (1960) 217–237.
[38] C. Shi, C.-W. Shu, On local conservation of numerical methods for conservation laws, Comput. Fluids 169 (2018) 3–9.
[39] J.E. Castillo, J.M. Hyman, M.J. Shashkov, S. Steinberg, The sensitivity and accuracy of fourth order finite-difference schemes on nonuniform grids in one dimension, Comput. Math. Appl. 30 (8) (1995) 41–55.
[40] A.E.P. Veldman, K.W. Lam, Symmetry-preserving upwind discretization of convection on non-uniform grids, Appl. Numer. Math. 58 (2008) 1881–1891.
[41] G.E. Sharpe, G.P.H. Styan, Circuit duality and the general network inverse, IEEE Trans. Circuit Theory 12 (1) (1965) 22–27.
[42] J.W. Holley, J.P. Guilford, Note on the double centering of dichotomized matrices, Scand. J. Psychol. 7 (1) (1966) 97–101.
[43] G.H. Golub, C.F. Van Loan, Matrix Computations, 3rd edition, JHU Press, 1996.
[44] V. Singh, P. Chandrashekar, On a linear stability issue of split form schemes for compressible flows, arXiv:2104.14941, 2021.
[45] S. Pirozzoli, Stabilized non-dissipative approximations of Euler equations in generalized curvilinear coordinates, J. Comput. Phys. 230 (8) (2011) 2997–3014.
[46] Y. Morinishi, Skew-symmetric form of convective terms and fully conservative finite difference schemes for variable density low-Mach number flows, J. Comput. Phys. 229 (2010) 276–300.
[47] R.A. Remmerswaal, A.E.P. Veldman, Towards a sharp, structure-preserving two-velocity model for two-phase flow: transport of mass and momentum, arXiv:2209.14934 [math.NA], 2022.
[48] A. Jameson, Formulation of kinetic energy preserving conservative schemes for gas dynamics and direct numerical simulation of one-dimensional viscous compressible flow in a shock tube using entropy and kinetic energy preserving schemes, J. Sci. Comput. 34 (2008) 188–208.
[49] H. Ranocha, Comparison of some entropy conservative numerical fluxes for the Euler equations, J. Sci. Comput. 76 (1) (2018) 216–242.
[50] H. Ranocha, G.J. Gassner, Preventing pressure oscillations does not fix local linear stability issues of entropy-based split-form high-order schemes, Commun. Appl. Math. Comput. Sci. 4 (3) (2022) 880–903.
[51] F. Ducros, F. Laporte, T. Souleres, V. Guinot, P. Moinat, B. Caruelle, High-order fluxes for conservative skew-symmetric-like schemes in structured meshes: application to compressible flows, J. Comput. Phys. 161 (1) (2000) 114–139.
[52] Y. Morinishi, T.S. Lund, O.V. Vasilyev, P. Moin, Fully conservative higher order finite difference schemes for incompressible flow, J. Comput. Phys. 143 (1998) 90–124.
[53] T.A. Manteuffel, A.B. White jr., The numerical solution of second-order boundary value problems on nonuniform meshes, Math. Comput. 47 (1986) 511–535.
[54] J.C. Tannehill, D.A. Anderson, R.H. Pletcher, Computational Fluid Mechanics and Heat Transfer, second edition, Taylor and Francis, Washington, 1997.
[55] A.E.P. Veldman, Discretization methods for the subsonic Reduced Navier–Stokes equations, Technical report, University of Groningen, The Netherlands, 2000, www.math.rug.nl/~veldman/preprints/preprints.html.
[56] C. De Michele, G. Coppola, An assessment of various discretizations of the energy equation in compressible flows, in: The 8th European Congress on Computational Methods in Applied Sciences and Engineering. ECCOMAS Congress, 5-9 June 2022, Oslo, Norway, 2022.
[57] B. van't Hof, A.E.P. Veldman, Mass, momentum and energy conserving (MaMEC) discretizations on general grids for the compressible Euler and shallow water equations, J. Comput. Phys. 231 (2012) 4723–4744.

[58] B. van't Hof, M.J. Vuik, Symmetry-preserving finite-difference discretizations of arbitrary order on structured curvilinear staggered grids, J. Comput. Sci. 36 (2019) 101008.
[59] P. Moin, K. Squires, W. Cabot, S. Lee, A dynamic subgrid-scale model for compressible turbulence and scalar transport, Phys. Fluids A, Fluid Dyn. 3 (11) (1991) 2746–2757.
[60] G.A. Blaisdell, E.T. Spyropoulos, J.H. Qin, The effect of the formulation of nonlinear terms on aliasing errors in spectral methods, Appl. Numer. Math. 21 (3) (1996) 207–219.
[61] P.D. Thomas, C.K. Lombard, Geometric conservation law and its application to flow computations on moving grids, AIAA J. 17 (10) (1979) 1030–1037.