SIAM J. Numer. Anal.
Vol. 53, No. 4, pp. 2008–2029

STEADY STATE AND SIGN PRESERVING SEMI-IMPLICIT RUNGE–KUTTA METHODS FOR ODES WITH STIFF DAMPING TERM*

ALINA CHERTOCK†, SHUMO CUI‡, ALEXANDER KURGANOV‡, AND TONG WU‡

Abstract. In this paper, we develop a family of second-order semi-implicit time integration methods for systems of ordinary differential equations (ODEs) with stiff damping term. The important feature of the new methods resides in the fact that they are capable of exactly preserving the steady states as well as maintaining the sign of the computed solution under the time step restriction determined by the nonstiff part of the system only. The new semi-implicit methods are based on the modification of explicit strong stability preserving Runge–Kutta (SSP-RK) methods and are proven to have a formal second order of accuracy, $A(\alpha)$-stability, and stiff decay. We illustrate the performance of the proposed SSP-RK based semi-implicit methods on both a scalar ODE example and a system of ODEs arising from the semi-discretization of the shallow water equations with stiff friction term. The obtained numerical results clearly demonstrate that the ability of the introduced ODE solver to exactly preserve equilibria plays an important role in achieving high resolution when a coarse grid is used.

Key words. ordinary differential equations with stiff damping terms, semi-implicit methods, strong stability preserving Runge–Kutta methods, implicit-explicit methods, shallow water equations with friction terms

AMS subject classifications. 65L04, 65L06, 65L07, 65L20, 65M22, 86-08

DOI. 10.1137/151005798

1. Introduction.

In this paper, we consider the numerical integration of ordinary differential equations (ODEs) of the form

$$\boldsymbol{u}' = \boldsymbol{f}(\boldsymbol{u}, t) + G(\boldsymbol{u}, t)\boldsymbol{u}, \tag{1.1}$$

where $\boldsymbol{u} = \boldsymbol{u}(t) \in \mathbb{R}^N$ is an unknown vector function, $\boldsymbol{f} : \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}^N$ is a given vector field, and $G : \mathbb{R}^{N\times N} \times \mathbb{R} \to \mathbb{R}^{N\times N}$ is a diagonal nonpositive definite matrix representing a (stiff) damping term. We assume that $\boldsymbol{f}$ and G are Lipchitz continuous with respect to $\boldsymbol{u}$. Systems like (1.1) may arise from semi-discretizations of time-dependent partial differential equations (PDEs) by the method of lines. In such case, the vector $\boldsymbol{u}$ usually corresponds to the spatial discretization of an unknown quantity and functions $\boldsymbol{f}$ and G correspond to the spatial discretization of the terms of different types in the given PDE (e.g., nonlinear hyperbolic fluxes, friction/source terms).

Development of numerical methods for the system (1.1) is a challenging task due to the presence of the stiff damping term, which can lead to a great loss in accuracy and efficiency of the method. In many applications, the explicit treatment of the

*Received by the editors January 27, 2015; accepted for publication (in revised form) May 28, 2015; published electronically August 18, 2015.
http://www.siam.org/journals/sinum/53-4/100579.html

†Department of Mathematics, North Carolina State University, Raleigh, NC 27695 (chertock@math.ncsu.edu). The work this author was supported in part by NSF grants DMS-1115682 and DMS-1216974 and ONR grant N00014-12-1-0832.

‡Mathematics Department, Tulane University, New Orleans, LA 70118 (scui2@tulane.edu, kurganov@math.tulane.edu, twu2@tulane.edu). The work of the second and fourth authors was supported in part by NSF grant DMS-1216957. The work of the third author was supported in part by NSF grants DMS-1115718 and DMS-1216957 and ONR grant N00014-12-1-0833.

stiff damping term imposes a severe time step restriction which is several orders of magnitude smaller than a typical time step used for the corresponding damping-free version of the studied system.

An attractive alternative to explicit methods is implicit-explicit (IMEX) Runge–Kutta (RK) schemes, which treat the stiff part of (1.1) implicitly and thus typically have the stability domains based on the nonstiff term only; see, e.g., [2, 14, 15, 16, 20, 21]. Modern IMEX-RK methods are based on a combination of explicit strong-stability preserving Runge–Kutta (SSP-RK) methods [11, 12, 25] for the nonstiff terms and L-stable implicit RK methods [1, 27] for the stiff terms. SSP methods are popular since when solving an ODE system obtained as a semi-discretization of a nonlinear time-dependent PDE, a stronger (nonlinear) stability is required to resolve discontinuous solutions of the underlying PDE in a nonoscillatory manner; see, e.g., [22, 23].

Another remedy to avoid severe time step and accuracy limitations is to use a semi-implicit method, in which the stiff term is discretized in a semi-implicit approach, that is, only a portion of the stiff term is implicitly treated. Such semi-implicit methods have been widely used in shallow water equations (see, e.g., [3, 4]) and other applications (see, e.g., [24, 26]). One of the strategies is to let the first factor of the stiff term in (1.1), $G(\boldsymbol{u}, t)$, be treated explicitly and the second factor $\boldsymbol{u}$ be treated implicitly. For such methods, the extra time step restriction introduced by the stiff terms is much milder than their fully explicit counterpart and thus semi-implicit methods are typically much more efficient. In addition, compared to the time integration methods with a fully implicit treatment of the damping terms, the evolution equation for the semi-implicit treatment is very easy to solve and implement.

In many practical applications, some special properties of the ODE solver are demanded to reflect the critical characters of the solution of the system (1.1) that represent significant physical features of the underlying model. In this case, in order to preserve the physical meaning of the numerical solution, it is important to maintain these properties with both the spatial discretization and time integration. In particular, we are interested in problems whose solutions are small perturbations of steady states,

$$\boldsymbol{u}(t) \equiv \widehat{\boldsymbol{u}} \quad \text{s.t.} \quad \boldsymbol{f}(\widehat{\boldsymbol{u}}, t) \equiv -G(\widehat{\boldsymbol{u}}, t)\widehat{\boldsymbol{u}} \quad \forall t, \tag{1.2}$$

and thus it is very important to derive a numerical method that preserves such steady state solutions exactly. Most of the IMEX-RK methods, however, do not satisfy this requirement. Another key property of a numerical scheme for (1.1) is to preserve the sign of the numerical solution when the exact solution is either positive or negative. When, for instance, the initial condition $\boldsymbol{u}(0)$ and function $\boldsymbol{f}$ satisfy

$$\{\boldsymbol{u}(0) \geq 0, \quad \boldsymbol{f} \geq 0\} \quad \text{or} \quad \{\boldsymbol{u}(0) \leq 0, \quad \boldsymbol{f} \leq 0\}, \tag{1.3}$$

the exact solution of (1.1), $\boldsymbol{u}(t)$ maintains the same sign as $\boldsymbol{u}(0)$. Though the nonpositive stiff damping term $G(\boldsymbol{u}, t)\boldsymbol{u}$ may be dominating, the numerical discretization of this term should not alter the sign of the numerical solution. The violation of such requirement may result in unphysical solutions. Examples of a positivity preserving IMEX-RK method can be found in [8, 15], but the use of a more restrictive than usual time step may be required to maintain the sign preserving property of these schemes.

In this paper, we propose a new class of second-order semi-implicit time integration methods for the system (1.1), which are capable of preserving the steady state (1.2) as well as maintaining the sign of solution under the condition (1.3). Our semi-implicit methods are based on the modification of explicit SSP-RK methods as shown in section 2, where we also prove their formal second order of accuracy, $A(\alpha)$-stability with $\alpha = \pi/4$ and stiff decay, steady state, and sign preserving properties under the time step restriction determined by the nonstiff part of the system (1.1) only. The rest of the paper is organized as follows. In section 3, we study stability of the semi-implicit methods that are based on two popular SSP-RK explicit solvers. In section 4, we illustrate the performance of the proposed SSP-RK based semi-implicit methods on both a simple scalar ODE example and a system of ODEs arising from the spatial discretization of the shallow water equations with (stiff) friction term.

2. New semi-implicit methods.

In this section, we develop a new class of second-order semi-implicit RK (SI-RK) methods for the system (1.1). A unique feature of the developed methods is their ability to exactly preserve the steady states (1.2) and maintain the sign of the computed solution under the condition (1.3).

For the simplicity of presentation, we consider here a scalar ODE with a nonstiff term $f(u,t)$ and stiff damping term $g(u,t)u$ such that $g(u,t) \le 0$:

$$u' = f(u,t) + g(u,t)u. \tag{2.1}$$

A general explicit m-stage RK method for (2.1) reads (see, e.g., [12])

$$\begin{aligned} &u^{(0)} = u^n, \\ &u^{(i)} = \sum_{k=0}^{i-1} \alpha_{i,k} \left[u^{(k)} + \beta_{i,k}\Delta t (f^{(k)} + g^{(k)}u^{(k)}) \right], \quad i = 1, \dots, m, \\ &u^{n+1} = u^{(m)}, \end{aligned} \tag{2.2}$$

where $f^{(k)} := f(u^{(k)}, t^{(k)})$ and $g^{(k)} := g(u^{(k)}, t^{(k)})$. Here, $t^{(k)} := t^n + D_k \Delta t$, where D_k is given by

$$D_0 = 0, \quad D_i = \sum_{k=0}^{i-1} \alpha_{i,k}(D_k + \beta_{i,k}), \quad i = 1, \dots, m. \tag{2.3}$$

The RK method defined in (2.2) is fully determined by its coefficients $\{\alpha_{i,k}, \beta_{i,k}\}$, which can be used to describe its properties. In particular, the consistency requirement of the ith intermediate solution $u^{(i)}$ can be written as

$$\sum_{k=0}^{i-1} \alpha_{i,k} = 1, \quad i = 1, \dots, m, \tag{2.4}$$

and for the final solution u^{n+1} as

$$D_m = 1. \tag{2.5}$$

Note that the RK method (2.2) is in fact a linear combination of the first-order forward Euler (FE) steps, namely, we can rewrite

$$u^{(i)} = \sum_{k=0}^{i-1} \alpha_{i,k} u_{i,k}^{\mathrm{FE}}, \tag{2.6}$$

where

$$u_{i,k}^{\mathrm{FE}} := u^{(k)} + \beta_{i,k}\Delta t (f^{(k)} + g^{(k)}u^{(k)}) \tag{2.7}$$

is, in fact, one-step FE evolution with the time step equal to $\beta_{i,k}\Delta t$. According to [11, 12], the RK method is SSP provided the linear combination in (2.6) is a convex combination, that is,

$$\alpha_{i,k} \geq 0 \quad \forall i, k,$$

and an appropriate time step restriction (based on the time step restriction for the FE method) is imposed. Also notice that negative time increments (which are undesirable when time irreversible equations are solved) are avoided if $\beta_{i,k} \geq 0$ for all i, k.

Unfortunately, when the SSP-RK or any other explicit RK methods are applied to ODEs of the form (2.1), the stiffness of damping term will impose a severe time step restriction which will greatly impair the efficiency of the method. The aim of this work is to derive a class of efficient semi-implicit methods with a time step requirement, which depends on the nonstiff part only. To this end, we first replace the FE evolution steps (2.7), which are used as components in the RK method (2.2), by the semi-implicit ones:

$$u_{i,k}^{\mathrm{SI}} := u^{(k)} + \beta_{i,k}\Delta t(f^{(k)} + g^{(k)}u_{i,k}^{\mathrm{SI}}) \quad \Longleftrightarrow \quad u_{i,k}^{\mathrm{SI}} = \frac{u^{(k)} + \beta_{i,k}\Delta t f^{(k)}}{1 - \beta_{i,k}\Delta t g^{(k)}}. \tag{2.8}$$

This leads to the following semi-implicit scheme:

$$\begin{aligned} u^{(0)} &= u^n, \\ u^{(i)} &= \sum_{k=0}^{i-1} \alpha_{i,k}\left(\frac{u^{(k)} + \beta_{i,k}\Delta t f^{(k)}}{1 - \beta_{i,k}\Delta t g^{(k)}}\right), \quad i = 1, \ldots, m, \\ u^{n+1} &= u^{(m)}. \end{aligned} \tag{2.9}$$

However, the scheme (2.9) with the coefficients $\{\alpha_{i,k}, \beta_{i,k}\}$ directly borrowed from (2.2) is at most first-order accurate (see Remark 2). We, therefore, propose a correction step which rectifies the final stage solution $u^{(m)}$:

$$u^{n+1} = \frac{u^{(m)} - C_m(\Delta t)^2 f^{(m)} g^{(m)}}{1 + C_m(\Delta t g^{(m)})^2}, \tag{2.10}$$

where the constant C_m can be recursively computed by

$$C_0 = 0, \quad C_i = \sum_{k=0}^{i-1} \alpha_{i,k}(C_k + \beta_{i,k}^2), \quad i = 1, \ldots, m. \tag{2.11}$$

Combining the semi-implicit evolution formula (2.9) with the correction step (2.10), we introduce a new class of second-order SI-RK methods for (2.1):

$$\begin{aligned} u^{(0)} &= u^n, \\ u^{(i)} &= \sum_{k=0}^{i-1} \alpha_{i,k}\left(\frac{u^{(k)} + \beta_{i,k}\Delta t f^{(k)}}{1 - \beta_{i,k}\Delta t g^{(k)}}\right), \quad i = 1, \ldots, m, \\ u^{n+1} &= \frac{u^{(m)} - C_m(\Delta t)^2 f^{(m)} g^{(m)}}{1 + C_m(\Delta t g^{(m)})^2}, \end{aligned} \tag{2.12}$$

where the set of coefficients $\{\alpha_{i,k},\beta_{i,k}\}$ is taken directly from the explicit SSP-RK method of an appropriate order.

Remark 1. Note that in the degenerate case of $g\equiv 0$, the SI-RK method (2.12) is identical to the corresponding explicit RK method (2.2).

In the rest of this section, we provide proofs of the second-order accuracy, $A(\alpha)$-stability with $\alpha=\pi/4$ and stiff decay, steady state, and sign preserving properties of the new SI-RK methods.

We begin with proving the following lemma, where we measure the difference between the intermediate solutions computed by the RK method (2.2) and the SI-RK method (2.12).

LEMMA 2.1. *Let us assume that the RK* (2.2) *and SI-RK* (2.12) *methods with* $\beta_{i,k}\ge 0$ *for all* i,k *are applied to* (2.1) *to evolve the solution* u^n *for one time step from* t^n *to* $t^{n+1}=t^n+\Delta t$*. We denote the obtained intermediate solutions by* $u^{(i)}_{\rm RK}$ *and* $u^{(i)}_{\text{SI-RK}}$*, respectively. Then,*

$$u^{(i)}_{\text{SI-RK}}-u^{(i)}_{\rm RK}=C_i(\Delta t)^2g^n\left(f^n+g^nu^n\right)+\mathcal{O}((\Delta t)^3),\quad i=0,\ldots,m, \tag{2.13}$$

where C_i *are defined in* (2.11)*,* $f^n:=f(u^n,t^n)$ *and* $g^n:=g(u^n,t^n)$*.*

Proof. For the sake of simplicity, we define $f^{(i)}_{\text{SI-RK}}:=f(u^{(i)}_{\text{SI-RK}},t^{(i)})$, $g^{(i)}_{\text{SI-RK}}:=g(u^{(i)}_{\text{SI-RK}},t^{(i)})$, $f^{(i)}_{\rm RK}:=f(u^{(i)}_{\rm RK},t^{(i)})$, and $g^{(i)}_{\rm RK}:=g(u^{(i)}_{\rm RK},t^{(i)})$ for $i=0,\ldots,m$. We will prove the lemma by induction.

First, for $i=0$ (2.13) is true because

$$u^{(0)}_{\text{SI-RK}}-u^{(0)}_{\rm RK}=u^n-u^n=0,$$

and $C_0=0$ as defined in (2.11).

For $i=\ell$, $0<\ell\le m$, let us assume that (2.13) is true for all $i\le \ell-1$. The ℓth intermediate solutions in (2.2) and (2.12) are

$$u^{(\ell)}_{\text{SI-RK}}=\sum_{k=0}^{\ell-1}\alpha_{\ell,k}\left(\frac{u^{(k)}_{\text{SI-RK}}+\beta_{\ell,k}\Delta tf^{(k)}_{\text{SI-RK}}}{1-\beta_{\ell,k}\Delta tg^{(k)}_{\text{SI-RK}}}\right), \tag{2.14}$$

$$u^{(\ell)}_{\rm RK}=\sum_{k=0}^{\ell-1}\alpha_{\ell,k}\left[u^{(k)}_{\rm RK}+\beta_{\ell,k}\Delta t\left(f^{(k)}_{\rm RK}+g^{(k)}_{\rm RK}u^{(k)}_{\rm RK}\right)\right]. \tag{2.15}$$

The Taylor expansion of (2.14) with respect to Δt gives

$$\begin{aligned}u^{(\ell)}_{\text{SI-RK}}=\sum_{k=0}^{\ell-1}\alpha_{\ell,k}\Big[&u^{(k)}_{\text{SI-RK}}+\beta_{\ell,k}\Delta t\big(f^{(k)}_{\text{SI-RK}}+g^{(k)}_{\text{SI-RK}}u^{(k)}_{\text{SI-RK}}\big)\\&+\beta^2_{\ell,k}(\Delta t)^2g^{(k)}_{\text{SI-RK}}\big(f^{(k)}_{\text{SI-RK}}+g^{(k)}_{\text{SI-RK}}u^{(k)}_{\text{SI-RK}}\big)\Big]+\mathcal{O}((\Delta t)^3).\end{aligned} \tag{2.16}$$

By the induction assumption

$$u^{(k)}_{\text{SI-RK}}=u^{(k)}_{\rm RK}+C_k(\Delta t)^2g^n(f^n+g^nu^n))+\mathcal{O}((\Delta t)^3),\quad 0\le k\le \ell-1,$$

and thus the second term in the summation on the right-hand side (RHS) of (2.16) can be written as follows:

$$\beta_{\ell,k}\Delta t\big(f^{(k)}_{\text{SI-RK}}+g^{(k)}_{\text{SI-RK}}u^{(k)}_{\text{SI-RK}}\big)=\beta_{\ell,k}\Delta t\big(f^{(k)}_{\rm RK}+g^{(k)}_{\rm RK}u^{(k)}_{\rm RK}\big)+\mathcal{O}((\Delta t)^3). \tag{2.17}$$

By consistency of the SI-RK method (2.12)

$$u^{(k)}_{\text{SI-RK}} = u^n + \mathcal{O}(\Delta t), \quad 0 \le k \le \ell - 1, \tag{2.18}$$

and hence the third term in the summation on the RHS of (2.16) is

$$\beta^2_{\ell,k}(\Delta t)^2 g^{(k)}_{\text{SI-RK}}\big(f^{(k)}_{\text{SI-RK}} + g^{(k)}_{\text{SI-RK}} u^{(k)}_{\text{SI-RK}}\big) = \beta^2_{\ell,k}(\Delta t)^2 g^n\big(f^n + g^n u^n\big) + \mathcal{O}((\Delta t)^3). \tag{2.19}$$

Finally, substituting (2.17) and (2.19) into (2.16) and subtracting (2.15) from (2.16), we obtain the desired estimate:

$$\begin{aligned} u^{(\ell)}_{\text{SI-RK}} - u^{(\ell)}_{\text{RK}} &= \sum_{k=0}^{\ell-1} \alpha_{\ell,k}\big(C_k(\Delta t)^2 g^n(f^n + g^n u^n) \\ &\quad + \beta^2_{\ell,k}(\Delta t)^2 g^n(f^n + g^n u^n)\big) + \mathcal{O}((\Delta t)^3) \\ &= \bigg(\sum_{k=0}^{\ell-1} \alpha_{\ell,k}(C_k + \beta^2_{\ell,k})\bigg)(\Delta t)^2 g^n(f^n + g^n u^n) + \mathcal{O}((\Delta t)^3) \\ &\overset{(2.11)}{=} C_\ell(\Delta t)^2 g^n(f^n + g^n u^n) + \mathcal{O}((\Delta t)^3). \qquad \square \end{aligned}$$

Remark 2. As it immediately follows from Lemma 2.1, the scheme (2.9) is at most first-order accurate. The correction step introduced in (2.10) is needed to increase the order to the second one, as shown in the following theorem.

THEOREM 2.2 (second-order accuracy). *Let us assume that the RK* (2.2) *and SI-RK* (2.12) *methods are applied to* (2.1)*. If the RK method* (2.2) *is at least second-order accurate, then the corresponding SI-RK method* (2.12) *with the same set of coefficients* $\alpha_{i,k}, \beta_{i,k} \ge 0$ *is second order provided the coefficient* C_m *is calculated by* (2.11)*.*

Proof. First, we use the Taylor expansion of (2.10) with respect to Δt to obtain

$$u^{n+1}_{\text{SI-RK}} = u^{(m)}_{\text{SI-RK}} - C_m(\Delta t)^2 g^{(m)}_{\text{SI-RK}}(f^{(m)}_{\text{SI-RK}} + g^{(m)}_{\text{SI-RK}} u^{(m)}_{\text{SI-RK}}) + \mathcal{O}((\Delta t)^3),$$

which using the consistency condition (2.18) can be rewritten as

$$u^{n+1}_{\text{SI-RK}} = u^{(m)}_{\text{SI-RK}} - C_m(\Delta t)^2 g^{n}_{\text{SI-RK}}(f^{n}_{\text{SI-RK}} + g^{n}_{\text{SI-RK}} u^{n}_{\text{SI-RK}}) + \mathcal{O}((\Delta t)^3).$$

It then follows from Lemma 2.1 that

$$u^{n+1}_{\text{SI-RK}} = u^{n+1}_{\text{RK}} + \mathcal{O}((\Delta t)^3).$$

Finally, our accuracy assumption on the RK method (2.2) implies that its truncation error is

$$u^{n+1}_{\text{RK}} - u(t^{n+1}) = \mathcal{O}((\Delta t)^\ell), \quad \ell \ge 3,$$

which, in turn, implies that the truncation error of the SI-RK method (2.12) is

$$u^{n+1}_{\text{SI-RK}} - u(t^{n+1}) = \mathcal{O}((\Delta t)^3).$$

We have thus proved that the SI-RK method is second-order accurate. $\square$

Next, we prove that the SI-RK methods are $A(\alpha)$-stable with $\alpha = \pi/4$ and have stiff decay for the equation $u' = g(u,t)u$.

THEOREM 2.3 ($A(\alpha)$-stability and stiff decay). *Let us assume that the SI-RK methods* (2.12) *are applied to* (2.1) *with* $f(u,t) \equiv 0$ *and* $g(u,t) \equiv \lambda$*, where* $\lambda \in \mathbb{C}$ *is a constant with* $\operatorname{Re}\lambda < 0$*. Then, the resulting methods, which can be written as*

$$u^{n+1} = R(z)u^n, \quad z = \lambda\Delta t,$$

satisfy the following two requirements:

$$|R(z)| \le 1 \ \forall z \in \mathbb{C} \textit{ such that } \operatorname{Re} z \le -|\operatorname{Im} z| \qquad \Big(A(\alpha)\text{-stability with } \alpha = \frac{\pi}{4}\Big) \tag{2.20}$$

and

$$R(z) \to 0 \text{ as } \operatorname{Re} z \to -\infty, \tag{2.21}$$

provided $\alpha_{i,k} \ge 0$ *and* $\beta_{i,k} \ge 0$ *for all* i,k*.*

Proof. We first write the function $R(z)$ using the following recursive relationship for $R^{(i)}(z) := u^{(i)}/u^n$:

$$R^{(0)}(z) = 1, \tag{2.22}$$

$$R^{(i)}(z) = \sum_{k=0}^{i-1} \alpha_{i,k}\left(\frac{R^{(k)}(z)}{1-\beta_{i,k}z}\right), \quad i = 1,\dots,m, \tag{2.23}$$

$$R(z) = \frac{R^{(m)}(z)}{1 + C_m z^2}. \tag{2.24}$$

It then follows from (2.23), (2.4), and positivity of $\beta_{i,k}$ that for $z \in \mathbb{C}$, $\operatorname{Re} z \le 0$,

$$|R^{(i)}(z)| \le \sum_{k=0}^{i-1} \alpha_{i,k}|R^{(k)}(z)| \le \max_{0\le k\le i-1} |R^{(k)}(z)|, \quad i = 1,\dots,m,$$

which together with (2.22) implies that

$$|R^{(m)}(z)| \le |R^{(0)}(z)| = 1. \tag{2.25}$$

The $A(\alpha)$-stability with $\alpha = \pi/4$ is then obtained from the inequality

$$|1 + C_m z^2| \ge 1,$$

which is clearly true as long as $\operatorname{Re} z \le -|\operatorname{Im} z|$ and $C_m \ge 0$ (note that the latter is ensured by (2.11) and positivity of $\alpha_{i,k} \ge 0$ for all i,k).

In fact, $C_m > 0$ since C_m may be equal to zero only if $\alpha_{i,k}\beta_{i,k} \equiv 0$, which contradicts the consistency requirement (2.5). Therefore, (2.24) together with (2.25) implies (2.21), which completes the proof of the theorem. □

We next prove the steady state preserving property of the proposed SI-RK methods.

THEOREM 2.4 (steady state preserving property). *Let us assume that the SI-RK methods* (2.12) *with* $\beta_{i,k} \ge 0$ *for all* i,k *are applied to* (2.1)*. Then, if the computed solution is at a steady state at time* t^n*, that is,* $u^n = \widehat{u}$ *such that*

$$f(\widehat{u},t) \equiv -g(\widehat{u},t)\widehat{u} \quad \forall t,$$

it will remain at the same steady state, namely,

$$u^{n+1} = \widehat{u}.$$

Proof. We first prove by induction that $u^{(m)} = \widehat{u}$. Indeed, if $u^{(k)} = \widehat{u}$ for all $k \le i-1$, then

$$u^{(i)} = \sum_{k=0}^{i-1} \alpha_{i,k} \left(\frac{u^{(k)} + \beta_{i,k} \Delta t f^{(k)}}{1 - \beta_{i,k} \Delta t g^{(k)}} \right) = \sum_{k=0}^{i-1} \alpha_{i,k} \left(\frac{\widehat{u} - \beta_{i,k} \Delta t g(\widehat{u}, t^{(k)}) \widehat{u}}{1 - \beta_{i,k} \Delta t g(\widehat{u}, t^{(k)})} \right) = \sum_{k=0}^{i-1} \alpha_{i,k} \widehat{u} = \widehat{u},$$

where the last equality is obtained using the consistency requirement (2.4). We then substitute $u^{(m)} = \widehat{u}$ into the correction step of the SI-RK methods (2.12) to end up with

$$u^{n+1} = \frac{u^{(m)} - C_m (\Delta t)^2 f^{(m)} g^{(m)}}{1 + C_m (\Delta t g^{(m)})^2} = \frac{\widehat{u} + C_m (\Delta t g(\widehat{u}, t^{(m)}))^2 \widehat{u}}{1 + C_m (\Delta t g(\widehat{u}, t^{(m)}))^2} = \widehat{u}. \qquad \square$$

At the end of this section, we prove the sign preserving property of the proposed SI-RK methods.

THEOREM 2.5 (sign preserving property). *Let us assume that the SI-RK methods* (2.12) *are applied to* (2.1) *within a time step* $[t^n, t^{n+1}]$. *If* u^n *and* $f^{(i)}, i = 0, \dots, m$, *are of the same sign, then*

$$\operatorname{sgn}(u^{n+1}) \equiv \operatorname{sgn}(u^n), \tag{2.26}$$

provided $\alpha_{i,k} \ge 0$ *and* $\beta_{i,k} \ge 0$ *for all* i, k.

Proof. First, assume that $u^n = u^{(0)} \ge 0$ and $f^{(0)} \ge 0$. It will be enough to show that $u^{(1)} \ge 0$ ((2.26) will then follow by induction). The positivity of $u^{(1)}$ immediately follows from (2.12) by taking into account that $f^{(i)} \ge 0$, $g^{(i)} \le 0$, $\alpha_{i,k} \ge 0$, $\beta_{i,k} \ge 0$, and $C_m \ge 0$. The case of $u^n \le 0$ is completely analogous and thus the proof of the theorem is complete. $\square$

Remark 3. It follows from the proof of Theorem 2.5 that the damping term $g(u,t)u$ cannot alter the sign of the numerical solution even when it is very stiff and dominating. Therefore, as long as $f(u,t)$ does not change its sign during one step of the computation, the sign of the solution u remains unchanged during this time step. Moreover, if f does not change sign at all, the local result from Theorem 2.5 extends to the global one.

COROLLARY 2.6. *Let us assume that the SI-RK methods* (2.12) *are applied to* (2.1) *with the initial condition* u^0 *and function* f *satisfying either*

$$\{u^0 \ge 0, \quad f \ge 0\} \quad \text{or} \quad \{u^0 \le 0, \quad f \le 0\}$$

Then,

$$\operatorname{sgn}(u^n) \equiv \operatorname{sgn}(u^0) \tag{2.27}$$

for all n *provided* $\alpha_{i,k} \ge 0$ *and* $\beta_{i,k} \ge 0$ *for all* i, k.

3. Absolute stability of two SSP-based SI-RK methods. In this section, we study the absolute stability of two SI-RK methods, which are particular cases of the general SI-RK methods (2.12). The first SI-RK method, based on the second-order SSP-RK solver [11, 12] also known as the Heun method [13], reads

$$\begin{aligned} u^{(1)} &= \frac{u^n + \Delta t f^n}{1 - \Delta t g^n}, \\ u^{(2)} &= \frac{1}{2} u^n + \frac{1}{2} \cdot \frac{u^{(1)} + \Delta t f^{(1)}}{1 - \Delta t g^{(1)}}, \\ u^{n+1} &= \frac{u^{(2)} - (\Delta t)^2 f^{(2)} g^{(2)}}{1 + (\Delta t g^{(2)})^2}, \end{aligned} \tag{3.1}$$

and the second one, based on the third-order SSP-RK method, can be written as

$$\begin{aligned} u^{(1)} &= \frac{u^n + \Delta t f^n}{1 - \Delta t g^n}, \\ u^{(2)} &= \frac{3}{4} u^n + \frac{1}{4} \cdot \frac{u^{(1)} + \Delta t f^{(1)}}{1 - \Delta t g^{(1)}}, \\ u^{(3)} &= \frac{1}{3} u^n + \frac{2}{3} \cdot \frac{u^{(2)} + \Delta t f^{(2)}}{1 - \Delta t g^{(2)}}, \\ u^{n+1} &= \frac{u^{(3)} - (\Delta t)^2 f^{(3)} g^{(3)}}{1 + (\Delta t g^{(3)})^2}. \end{aligned} \tag{3.2}$$

In what follows, the SI-RK method (3.1) will be referred to as the SI-RK2 method, while the SI-RK method (3.2) will be referred to as the SI-RK3.

To analyze the absolute stability, we consider the following test problem:

$$y' = \lambda_1 y + \lambda_2 y, \quad \lambda_1 \in \mathbb{C}, \ \mathrm{Re}(\lambda_1) \le 0, \quad \lambda_2 \in \mathbb{R}, \ \lambda_2 \le 0, \tag{3.3}$$

where $\lambda_1 y$ and $\lambda_2 y$ are the nonstiff and stiff parts, respectively. We denote $z_1 := \lambda_1 \Delta t$ and $z_2 := \lambda_2 \Delta t$.

Applying the schemes (3.1) and (3.2) to (3.3), that is, substituting $f(u,t) = \lambda_1 u$ and $g(u,t) = \lambda_2$ into (3.1) and (3.2), results in

$$\begin{aligned} u^{(1)} &= \frac{1 + z_1}{1 - z_2}\, u^n, \\ u^{(2)} &= \frac{1}{2} u^n + \frac{1}{2} \cdot \frac{1 + z_1}{1 - z_2}\, u^{(1)}, \\ u^{n+1} &= \frac{1 - z_1 z_2}{1 + z_2^2}\, u^{(2)}, \end{aligned} \tag{3.4}$$

and

$$\begin{aligned} u^{(1)} &= \frac{1 + z_1}{1 - z_2}\, u^n, \\ u^{(2)} &= \frac{3}{4} u^n + \frac{1}{4} \cdot \frac{1 + z_1}{1 - z_2}\, u^{(1)}, \\ u^{(3)} &= \frac{1}{3} u^n + \frac{2}{3} \cdot \frac{1 + z_1}{1 - z_2}\, u^{(2)}, \\ u^{n+1} &= \frac{1 - z_1 z_2}{1 + z_2^2}\, u^{(3)}, \end{aligned} \tag{3.5}$$

respectively.

It should be observed that if $g \equiv 0$, the underlying ODE (2.1) is nonstiff and thus $\lambda_2 = 0$ in the test problem (3.3). In this case, the SI-RK2 and SI-RK3 methods reduce to the second-order and third-order SSP-RK methods, respectively, and their stability regions are known; see, e.g., [11, 12]. Let us denote these stability regions by $\mathcal{D}_{\text{SSP2}}$ and $\mathcal{D}_{\text{SSP3}}$, respectively, and the corresponding time step restrictions by $\Delta t \le \Delta t_{\text{SSP2}}$ and $\Delta t \le \Delta t_{\text{SSP3}}$, given parameter λ_1.

Theorem 3.1 (absolute stability of the SI-RK2 method). *The region of absolute stability of the SI-RK2 method* (3.1) *contains* $\mathcal{D}_{\text{SSP2}}$*, that is, for any* $z_2 \le 0$*, the solution of* (3.4) *satisfies* $|u^{n+1}| \le |u^n|$ *provided* $\Delta t \le \Delta t_{\text{SSP2}}$.

Proof. First, we note that the stability functions for the SI-RK2 and the corresponding second-order SSP-RK methods are

$$R_{\text{SI-RK2}}(z_1, z_2) = \frac{1 - z_1 z_2}{1 + z_2^2} \cdot \left[\frac{1}{2} + \frac{1}{2}\left(\frac{1 + z_1}{1 - z_2}\right)^2\right] \quad \text{and} \quad R_{\text{SSP2}}(z_1) = \frac{1}{2} + \frac{1}{2}(1 + z_1)^2,$$

respectively. To prove the theorem, it will be enough to show that both

$$\left|\frac{1}{2} + \frac{1}{2}\left(\frac{1 + z_1}{1 - z_2}\right)^2\right| \le 1 \tag{3.6}$$

and

$$\left|\frac{1 - z_1 z_2}{1 + z_2^2}\right| \le 1 \tag{3.7}$$

for all z_1, z_2 such that $|R_{\text{SSP2}}(z_1)| \le 1$ and $z_2 \le 0$.

It follows from the definition of $R_{\text{SSP2}}(z_1)$ that $|\frac{1}{2} + \frac{1}{2}z^2| \le 1$ for all $z \in \mathcal{S}$, where $\mathcal{S} = \{z \in \mathbb{C} \mid z - 1 \in \mathcal{D}_{\text{SSP2}}\}$. Note that $\mathcal{S}$ is a convex region that encloses the origin. Hence, for any $z \in \mathcal{S}$ and any $z_2 \le 0$, we have $\frac{z}{1 - z_2} \in \mathcal{S}$, which implies that (3.6) for all $z_1 \in \mathcal{D}$.

We now turn to the proof of (3.7), which is obviously true for $z_2 = 0$. We therefore consider $z_2 < 0$, for which (3.7) is equivalent to

$$\left|z_1 + \frac{1}{|z_2|}\right| \le |z_2| + \frac{1}{|z_2|}.$$

For a fixed $z_2 < 0$, this inequality is satisfied in a disk with radius $|z_2| + \frac{1}{|z_2|}$ centered at $z_1 = -\frac{1}{|z_2|}$. Denoting $z_1 := x + iy$, we can write this domain as

$$\mathcal{C}(z_2) := \left\{z_1 \mid \left|\frac{1 - z_1 z_2}{1 + z_2^2}\right| \le 1\right\} = \left\{x + iy \mid y^2 \le \left(z_2 + \frac{1}{z_2}\right)^2 - \left(x - \frac{1}{z_2}\right)^2\right\} \quad \forall z_2 < 0. \tag{3.8}$$

We thus need to show that $\mathcal{D}_{\text{SSP2}} \subset \mathcal{C} := \bigcap\limits_{z_2 < 0} \mathcal{C}(z_2)$. Finding the intersection of $\mathcal{C}(z_2)$'s is equivalent to minimizing the set in (3.8) over $z_2 < 0$, which results in

$$\begin{aligned} \mathcal{C} &= \left\{x + iy \mid y^2 \le \min_{z_2 < 0}\left[\left(z_2 + \frac{1}{z_2}\right)^2 - \left(x - \frac{1}{z_2}\right)^2\right]\right\} \\ &= \left\{x + yi \mid y^2 \le 2 + 3x^{2/3} - x^2,\ x \in \left[-2\sqrt{2}, 0\right]\right\}. \end{aligned}$$

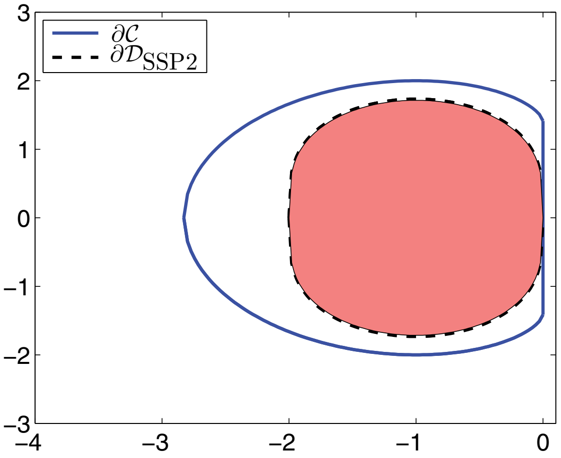

FIG. 1. *Stability region $\mathcal{D}_{\text{SSP2}}$ contained in $\mathcal{C}$.*

Therefore, the boundary of $\mathcal{C}$ consists of the curve $y^2 = 2+3x^{2/3}-x^2$, $-2\sqrt{2} \le x \le 0$ and a part of the y-axis from $(0,-\sqrt{2})$ to $(0,\sqrt{2})$. The boundaries of $\mathcal{D}_{\text{SSP2}}$ and $\mathcal{C}$ are shown in Figure 1, which clearly indicates that $\mathcal{D}_{\text{SSP2}} \subset \mathcal{C}$, and thus the proof of the theorem is complete. □

Remark 4. Our numerical experiments clearly indicate that a similar result holds for the SI-RK3 method as well. However, no rigorous proof of that fact is available and therefore we formulate it as a conjecture and provide supporting numerical evidences.

CONJECTURE 3.1 (absolute stability of the SI-RK3 method). *The region of absolute stability of the SI-RK*3 *method* (3.2) *contains $\mathcal{D}_{\text{SSP3}}$, that is, for any $z_2 \le 0$, the solution of* (3.5) *satisfies $|u^{n+1}| \le |u^n|$ provided $\Delta t \le \Delta t_{\text{SSP3}}$.*

Discussion. The stability functions for the SI-RK3 and the corresponding third-order SSP-RK methods are

$$R_{\text{SI-RK3}}(z_1,z_2) = \frac{1-z_1z_2}{1+z_2^2}\cdot\left[\frac{1}{3}+\frac{1}{2}\left(\frac{1+z_1}{1-z_2}\right)+\frac{1}{6}\left(\frac{1+z_1}{1-z_2}\right)^3\right]$$

and

$$R_{\text{SSP3}}(z_1) = \frac{1}{3}+\frac{1}{2}(1+z_1)+\frac{1}{6}(1+z_1)^3,$$

respectively. The statement of the conjecture would be true if one could show that

$$(3.9)\qquad |R_{\text{SI-RK3}}(z_1,z_2)| \le 1 \quad \forall z_1 \text{ such that } |R_{\text{SSP3}}(z_1)| \le 1 \quad\text{and}\quad \forall z_2 \le 0.$$

In order to verify (3.9), we first observe that $R_{\text{SI-RK3}}(z_1,0) = R_{\text{SSP3}}(z_1)$ and first consider the case $z_2 \le -3$. It follows from the definition of $R_{\text{SSP3}}(z_1)$ that $|\frac{1}{3}+\frac{1}{2}z+\frac{1}{6}z^3| \le 1$ for all $z \in \mathcal{S}$, where $\mathcal{S} = \{z \in \mathbb{C} \mid z-1 \in \mathcal{D}_{\text{SSP3}}\}$. Note that $\mathcal{S}$ is a region that encloses the origin O and contains the segments Oz for all $z \in \mathcal{S}$. Hence, for any $z \in \mathcal{S}$ and any $z_2 \le 0$, we have $\frac{z}{1-z_2} \in \mathcal{S}$, which implies that $|\frac{1}{3}+\frac{1}{2}(\frac{1+z_1}{1-z_2})+\frac{1}{6}(\frac{1+z_1}{1-z_2})^3| \le 1$. Notice that $\mathcal{D}_{\text{SSP3}}$ is contained in $\mathcal{B}_3(O)$, which is the disk of radius 3 centered at the origin, and therefore we have $|\frac{1-z_1z_2}{1+z_2^2}| \le \frac{1+3|z_2|}{1+|z_2|^2} \le 1$. Hence, we obtain that $|R_{\text{SI-RK3}}(z_1,z_2)| \le 1$ in the case $z_2 \le -3$.

To study the case $z_2 \in (-3, 0)$, we introduce a polynomial $P(x, y) := |R_{\text{SSP3}}(x + iy)|^2 - 1$ and a rational function $Q(x, y, z_2) := |R_{\text{SI-RK3}}(x + iy, z_2)|^2 - 1$. For fixed values of z_2, the curves given by $P(x, y) = 0$ and $Q(x, y, z_2) = 0$ are boundaries of the domains $\mathcal{D}_{\text{SSP3}}$ and $\mathcal{D}_{\text{SI-RK3}}(z_2)$, respectively (we denote by $\mathcal{D}_{\text{SI-RK3}}(z_2)$ the stability domain for the SI-RK3 method for a fixed value of z_2). To determine whether $\mathcal{D}_{\text{SSP3}} \subset \mathcal{D}_{\text{SI-RK3}}(z_2)$, we only need to show that $\partial\mathcal{D}_{\text{SSP3}}$ is enclosed by $\partial\mathcal{D}_{\text{SI-RK3}}(z_2)$ for all $z_2 \in (-3, 0)$.

To this end, we consider $P(x, y)$ and $Q(x, y, z_2)$ as polynomials of a single variable x and compute their resultant

$$K(y, z_2) := \text{res}(P, Q) = \frac{\widetilde{K}(y, z_2)}{6140942214464815497216(z_2 - 1)^{36}(z_2^2 + 1)^{12}},$$

where $\widetilde{K}(y, z_2)$ is a specific function, whose explicit expression is quite complicated and not instructive and we thus omit it for the sake of brevity. Instead, we use graphic software to visualize K. In Figure 2 (left), we plot $\log_{10}(\widetilde{K}(y, z_2) + 1)$, which clearly indicates that $K(y, z_2) > 0$ for all $(y, z_2) \in [-2.4, 2.4] \times (-3, 0)$ (note that we take the y-bounds to be $[-2.4, 2.4]$ in order to include the entire $\mathcal{D}_{\text{SSP3}}$ in the studied domain in the (y, z_2)-plane). This implies that $\partial\mathcal{D}_{\text{SSP3}}$ and $\partial\mathcal{D}_{\text{SI-RK3}}(z_2)$ have no intersections when $z_2 \in (-3, 0)$. To cite an example, we take $z_2 = -1$ and illustrate in Figure 2 (right) that $\mathcal{D}_{\text{SSP3}} \subset \mathcal{D}_{\text{SI-RK3}}(-1)$. Since $K(y, z_2)$ is continuous, we conclude that $\mathcal{D}_{\text{SSP3}} \subset \mathcal{D}_{\text{SI-RK3}}(z_2)$ for all $z_2 \in (-3, 0)$.

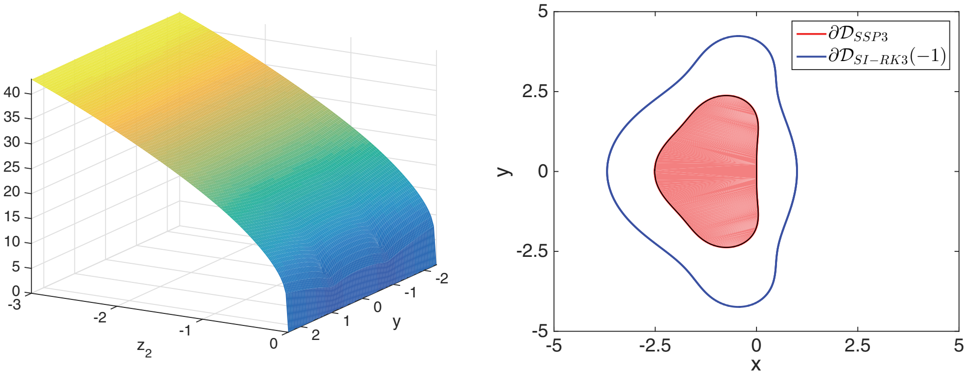

FIG. 2. *Surface plot of* $\log_{10}(\widetilde{K}(y, z_2)+1)$ *(left); the stability domains* $\mathcal{D}_{\text{SSP3}}$ *and* $\mathcal{D}_{\text{SI-RK3}}(-1)$ *(right).*

To further verify the relationship between the stability functions $R_{\text{SI-RK3}}$ and R_{SSP3}, we numerically evaluate $\max_{z_1,z_2} |R_{\text{SI-RK3}}(z_1, z_2)|$ under a stronger restriction on $|R_{\text{SSP3}}(z_1)|$:

$$|R_{\text{SSP3}}(z_1)| \le \alpha < 1. \tag{3.10}$$

The obtained results, presented in Figure 3, indicate even stronger dependence of $R_{\text{SI-RK3}}$ and R_{SSP3}, namely,

$$|R_{\text{SI-RK3}}(z_1, z_2)| \le \alpha \quad \forall z_1 \text{ such that } |R_{\text{SSP3}}(z_1)| \le \alpha \quad \text{and} \quad \forall z_2 \le 0.$$

Remark 5. Since the statement of Conjecture 3.1 has not been proved, we suggest the following strategy, which should guarantee the stability of the SI-RK3 method in

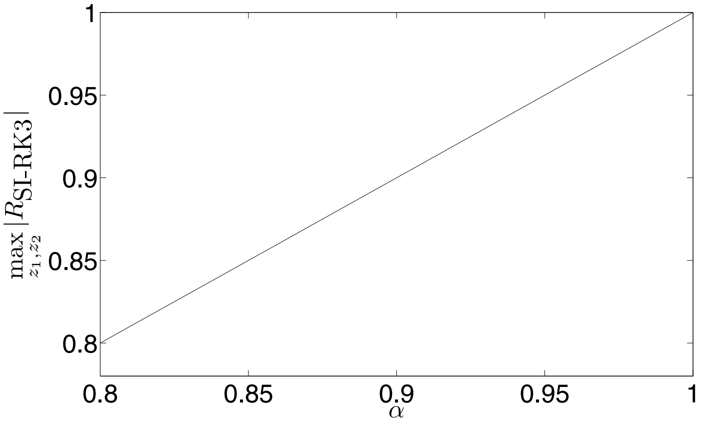

FIG. 3. *Conjecture* 3.1: $\max|R_{\text{SI-RK3}}(z_1,z_2)|$ *over* z_1 *such that* $|R_{\text{SSP3}}(z_1)| \le \alpha$ *and* $z_2 \le 0$ *as a function of* α.

practice. The time step Δt should be chosen such that (3.10) is satisfied with α being slightly smaller than 1, say, $\alpha = 0.98$, which will ensure that $|R_{\text{SI-RK3}}(z_1,z_2)| < 1$ as desired for the stability.

Remark 6. The results in Figure 3 are obtained in the most straightforward manner by sampling the variable domains with very fine meshes and then computing the discrete maxima. We use a mesh grid with space 0.01 to sample z_1 in the region $\mathcal{D}_{\text{SSP3}}$. To sample $z_2 \in [0,-\infty)$, we employ 1000 sample points $(z_2)_i = (1-\xi_i^{-1})^{-1}$, where $\xi_i = i/1000$, $i = 0,\ldots,999$. Note that the points ξ_i are uniformly spaced in $[0,1)$.

4. Numerical examples.

In this section, we test the second-order SI-RK3 method (3.2) on several examples including both scalar ODEs (section 4.1) and systems of ODEs arising from semi-discretizations of PDEs (section 4.2). We compare the results with the ones obtained using the second-order IMEX-SSP3(3,3,2) method described in [21, Table 5]. The obtained results clearly demonstrate that the new SI-RK3 method outperforms the IMEX-SSP3(3,3,2) when a large time step and/or coarse grid are used.

4.1. Scalar ODEs.

We consider the following scalar ODE:

$$u' = 1 - k|u|u, \tag{4.1}$$

where k is a positive real number. Equation (4.1) has one equilibrium point $u^* = 1/\sqrt{k}$, and its exact solution is given by

$$u_{\text{ex}}(t) = \begin{cases} \dfrac{1}{\sqrt{k}}\coth\left(\sqrt{k}\,t + \coth^{-1}(\sqrt{k}\,u(0))\right), & u(0) \ge u^*, \\ \dfrac{1}{\sqrt{k}}\tan\left(\sqrt{k}\,t + \tan^{-1}(\sqrt{k}\,u(0))\right), & u(0) < 0 \text{ and } t < -\dfrac{\tan^{-1}\left(\sqrt{k}\,u(0)\right)}{\sqrt{k}}, \\ \dfrac{1}{\sqrt{k}}\tanh\left(\sqrt{k}\,t + \tanh^{-1}(\sqrt{k}\,u(0))\right) & \text{otherwise.} \end{cases}$$

Example 1. *Accuracy test.* In this example, we apply the SI-RK3 and IMEX-SSP3(3,3,2) methods to (4.1) subject to the initial condition $u(0) = 0.2$. We compute the numerical solution (denoted by $\widehat{u}(t)$) until the final time $T = 0.1$ using N uniform time steps. In order to compare the performances of both methods, we apply them to problems with various levels of stiffness by taking different values of k ranging from 10^2 to 10^{14}.

In Figure 4, we show in a logarithmic scale the absolute value of the error $|\widehat{u}(T) - u_{\rm ex}(T)|$ as a function of the number of time steps N. As one can see, the results obtained by both methods demonstrate second-order convergence rates in the nonstiff regime ($k = 10^2$). However, in the stiff regimes ($k = 10^6$, 10^{10}, and 10^{14}) the convergence rates of the IMEX-SSP(3,3,2) method are effected by the increasing level of stiffness and an order reduction can be noticed, while the accuracy achieved by the SI-RK3 method in the stiff regime is in fact higher than its accuracy in the nonstiff regime.

Moreover, a machine accuracy is achieved by the SI-RK3 method for the stiff problems when N is chosen to be between 20 and 45, depending on the value of k; see Figure 4(a). For smaller N the error, however, seems to be insensitive to the number of steps. Such a phenomenon occurs since the final time solution $\widehat{u}(T)$ is very close to zero and the magnitude of $u_{\rm ex}(T) \approx 1/\sqrt{k}$ computed by the SI-RK3 method dominates in the error $|u_{\rm ex}(T) - \widehat{u}(T)|$ for small values of N.

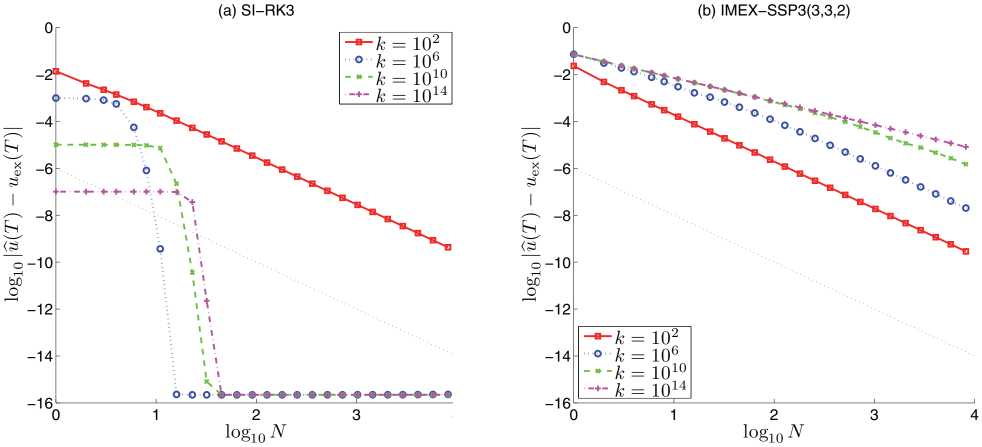

FIG. 4. *Example* 1: *The logarithm of absolute error at the final time T as a function of* $\log_{10} N$ *computed by the SI-RK*3 *(left) and IMEX-SSP*3(3, 3, 2) *(right) methods. The dotted straight lines in both figures have the slope* -2.

Example 2. *Steady state preserving test.* In this example, we take $k = 10,000$, which corresponds to the equilibrium point $u^* = 0.01$. We consider three different initial values,

$$\text{(a) } u(0) = 0.9u^*, \quad \text{(b) } u(0) = u^*, \quad \text{(c) } u(0) = 1.1u^*,$$

and solve (4.1) using both the SI-RK3 and IMEX-SSP3(3,3,2) methods. The numerical results computed with $\Delta t = 1/100, 1/200, 1/400, 1/800$, and $1/1600$ are plotted in Figure 5. As one can see, when u is initially at the equilibrium (case (b)), the SI-RK3 method preserves the steady state exactly as expected (see Theorem 2.4), while the difference between the numerical steady states obtained by the IMEX-SSP3(3,3,2) method and the exact steady state is of order $\mathcal{O}((\Delta t)^2)$. Similarly, in cases (a) and (c), the SI-RK3 method accurately captures and preserves the exact equilibrium, while the IMEX-SSP3(3,3,2) method does not.

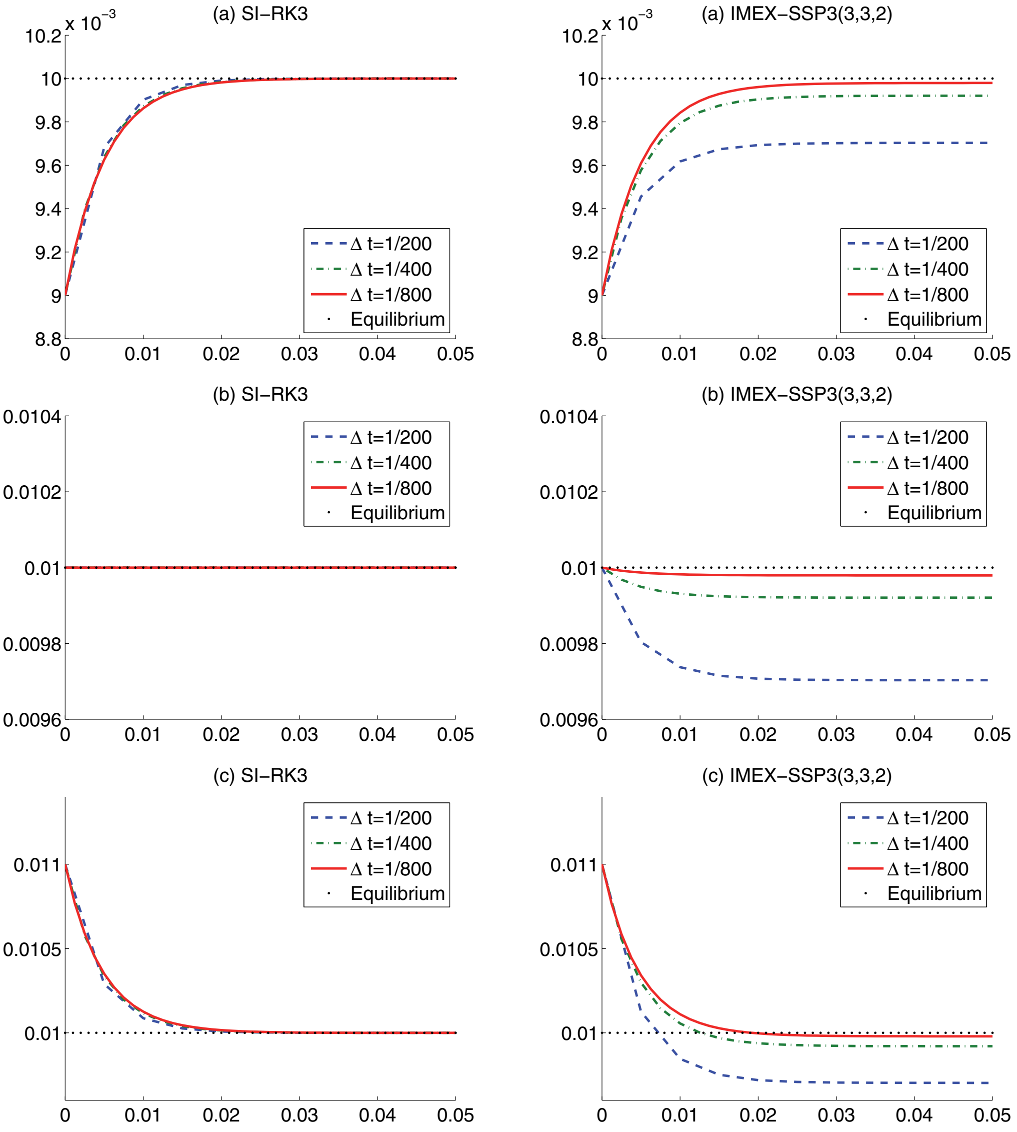

FIG. 5. *Example* 2: *Convergence toward the equilibrium for the SI-RK*3 *(left column) and IMEX-SSP*3*(*3, 3, 2*) (right column) methods.*

Example 3. *Sign preserving test.* As in Example 2, we choose $k = 10,000$ and solve (4.1) subject to the initial condition $u(0) = 1$, for which the exact solution must remain positive for all t. Once again, we implement both the SI-RK3 and IMEX-SSP3(3,3,2) methods with $\Delta t = 1/200, 1/400, 1/800, 1/1600$ and plot the obtained results in Figure 6. As one can see, the SI-RK3 method preserves the positive sign of the computed solution (as proved in Theorem 2.5), while the IMEX-SSP3(3,3,2) produces negative values of u.

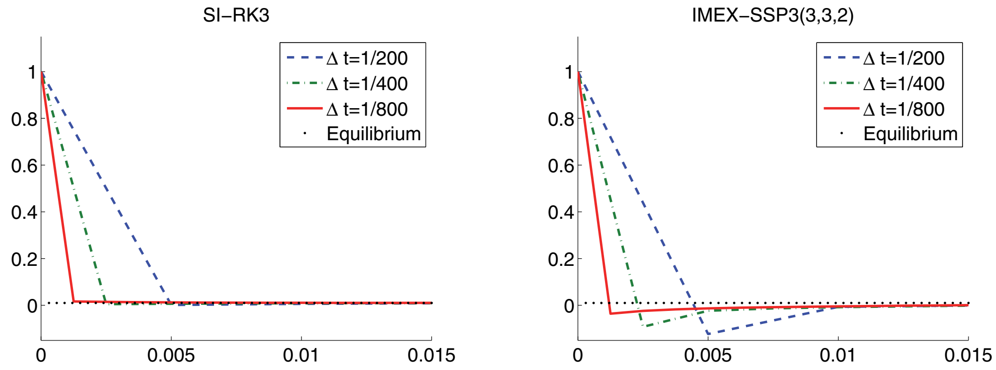

FIG. 6. *Example* 3*: Solutions computed by the SI-RK*3 *(left) and IMEX-SSP*3*(*3, 3, 2*) (right) methods.*

4.2. Systems of ODEs arising from semi-discretizations of PDEs. In this section, we consider the Saint-Venant system [7] of shallow water equations with the friction term in classical Manning formulation [6, 9, 10, 19]. In the one-dimensional case, the system reads

$$\begin{cases} h_t + q_x = 0, \\ q_t + \left(hv^2 + \dfrac{\mathfrak{g}}{2}h^2\right)_x = -\mathfrak{g}hB_x - \mathfrak{g}\dfrac{n^2}{h^{7/3}}|q|q, \end{cases} \tag{4.2}$$

where $h(x,t)$ denotes the water depth, $v(x,t)$ is the velocity, $q(x,t) := h(x,t)v(x,t)$ is the discharge, and $\mathfrak{g}$ is the gravitational constant. The first term on the RHS of (4.2) is the geometric source with $B(x)$ representing the bottom topography function. The second term on the RHS of (4.2) models the bottom friction with n being the Manning coefficient.

Solving the system (4.2) numerically is a challenging task due to the following reasons. First, the system admits several physically relevant steady states and many practically important solutions are in fact small perturbations of these steady states. A good numerical method should be well-balanced in the sense that it should be able to exactly preserve discrete versions of the relevant steady states since otherwise the numerical error can become larger than the waves to be captured. Second, the water depth h may be very small and even zero (in the islands and shore areas) and therefore it is crucial to design a numerical method that is capable of preserving the positivity of the computed water depth. Finally, when h is small in certain parts of the domain, the friction term in (4.2) becomes stiff and thus the use of an implicit or a semi-implicit discretization of this term is required for designing an efficient numerical method.

We solve the system (4.2) using the second-order semi-discrete central-upwind scheme from [5]; see also [17, 18]. For simplicity, we consider the semi-discretization framework, in which the computational domain is divided into N uniform cells $C_j :=$

$[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$ with $\Delta x \equiv x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$ and thus the system of PDEs (4.2) reduces to an ODE system consisting of $2N$ coupled equations:

$$\begin{cases} \dfrac{d\overline{h}_j(t)}{dt} = f_j^{(1)}\Big(\{\overline{h}_j(t)\}, \{\overline{q}_j(t)\}, B\Big), \\ \dfrac{d\overline{q}_j(t)}{dt} = f_j^{(2)}\Big(\{\overline{h}_j(t)\}, \{\overline{q}_j(t)\}, B\Big) - g_j\Big(\overline{h}_j(t), \overline{q}_j(t)\Big)\,\overline{q}_j(t), \end{cases}$$

where

$$\overline{h}_j(t) := \frac{1}{\Delta x}\int\limits_{C_j} h(x,t)dx \quad \text{and} \quad \overline{q}_j(t) := \frac{1}{\Delta x}\int\limits_{C_j} q(x,t)dx$$

are the average values of water height h and discharge q in the cell C_j, and $f_j^{(1)}$, $f_j^{(2)}$, g_j are obtained using the central-upwind discretization from [5].

As was shown in [18], the overall method is both well-balanced and positivity preserving provided the system of ODEs is integrated using an explicit SSP ODE solver. To design a method which is also efficient in the stiff regime, one can use the proposed SI-RK methods (2.12), which, as we have proved in section 2, are both steady state and sign preserving.

We would like to point out that the positivity preserving property of the central-upwind scheme from [5, 17, 18] guarantees the positivity of the water height h and is achieved by using a series of special techniques in spatial discretization, which are not the focus of this paper. By adopting the proposed SI-RK schemes, we take advantage of their sign preserving property, which is expected to be reflected in the evolution of the discharge q, that is, using the SI-RK schemes guarantees the sign of the computed velocity will not be changed due to the bottom friction terms.

We consider a nonsmooth periodic flow over the slanted surface. In this setting, physically relevant steady states satisfy $h \equiv \text{Const}, q \equiv \text{Const}, B_x \equiv \text{Const}$. In the numerical examples below, we implement the SI-RK3 method (3.2), which exactly preserves the above steady states and the sign of the velocity. We again compare the performance of the SI-RK3 method with the IMEX-SSP3(3,3,2) method, which is also efficient but is neither steady state or sign preserving. These features play an important role in the ability of the numerical scheme to capture the correct numerical solution of the shallow water system even when it is still far from the steady state as demonstrated below.

We numerically solve the system (4.2) with $B_x \equiv -0.2$, $n = 0.09$ and subject to the following initial conditions:

$$h(x,0) = \begin{cases} 0.02, & x < 50, \\ 0.01, & x > 50, \end{cases} \qquad q(x,0) = \begin{cases} 0, & x < 50, \\ 0.04, & x > 50. \end{cases} \tag{4.3}$$

We restrict the computational domain to $[0, 100]$, which is divided into $N = 100$ uniform cells, and impose the periodic boundary conditions.

Example 4. *Time steps restricted by the CFL condition.* In this example, we conduct a number of numerical experiments in which the time step Δt is adaptively determined based on the CFL condition with the CFL number 0.3. The final times T in different simulations slightly vary because of the different time steps chosen.

We first implement the proposed SI-RK3 method and plot the solution (h and v) computed at final time T in Figures 7(a) and (b). As one can see, the results obtained on coarse ($N = 100$) and fine ($N = 1000$) grids are in good agreement (obviously, the shock is sharper resolved on a fine grid). In addition, in Figure 7(c), we plot the value of the velocity v at $x = 50$ as a function of time. The figure clearly demonstrates that both the velocity v and the speed of the shock are captured quite accurately even with $N = 100$. Also note that the time steps shown in Figure 7(d) slightly increase in time, since they are adaptively selected using the CFL condition.

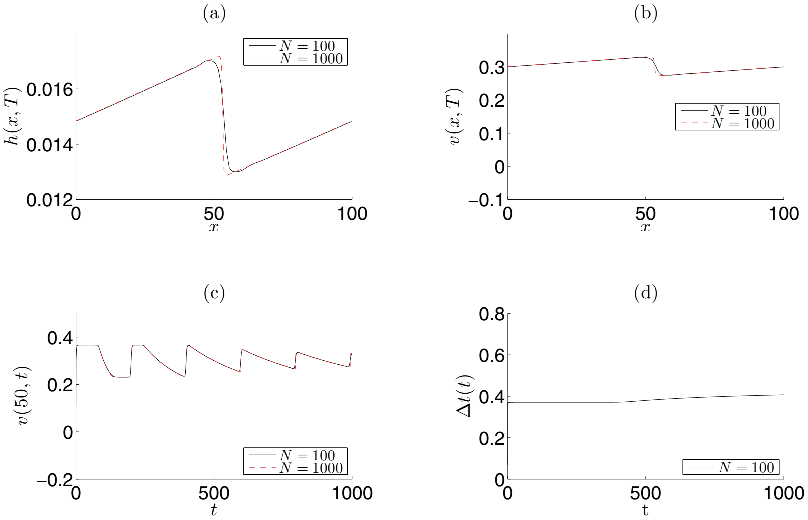

FIG. 7. *Example* 4: *Solution obtained using the SI-RK*3 *method: the water depth* (a) *and velocity* (b) *at time* $T = 1000.37$*; the velocity* (c) *at* $x = 50$ *as a function of time. The size of time steps* (d) *as a function of time.*

The solution computed on a course grid ($N = 100$) using the IMEX-SSP3(3,3,2) time integration is shown in Figure 8 and compared with the reference solution obtained by the same IMEX-SSP3(3,3,2) method with $N = 1000$. As one can see, the results show considerable disagreements between the fine- and coarse-grid solutions. In Figure 8(a), a phase error in h can be noticed. Such a delay in shock propagation is due to the large numerical error in velocity magnitude; see Figure 8(b). In Figure 8(c), one can also notice that the error in velocity magnitude arises initially and gradually increases in time (here again we show the values of the computed v at $x = 50$ as a function of time). Notice that at large times, $v(50,t)$ often admits negative values, which are unphysical and trigger oscillations that develop in both space and time, as can be clearly seen in Figures 8(b)–(d).

It is instructive to check whether the appearance of negative velocities is related to the lack of sign preserving property for the IMEX-SSP3(3,3,2) method. To this end, we compute $f^{(2)}_{\max} := \max_j \left\{f^{(2)}_j(t)\right\}$, $f^{(2)}_{\min} := \min_j \left\{f^{(2)}_j(t)\right\}$ and plot the obtained results in Figure 9. As one can see, at very small times the sign of f changes, but then it remains positive. Yet, the sign of $\overline{q}^{\,n}_j$ computed by the IMEX-SSP3(3,3,2)

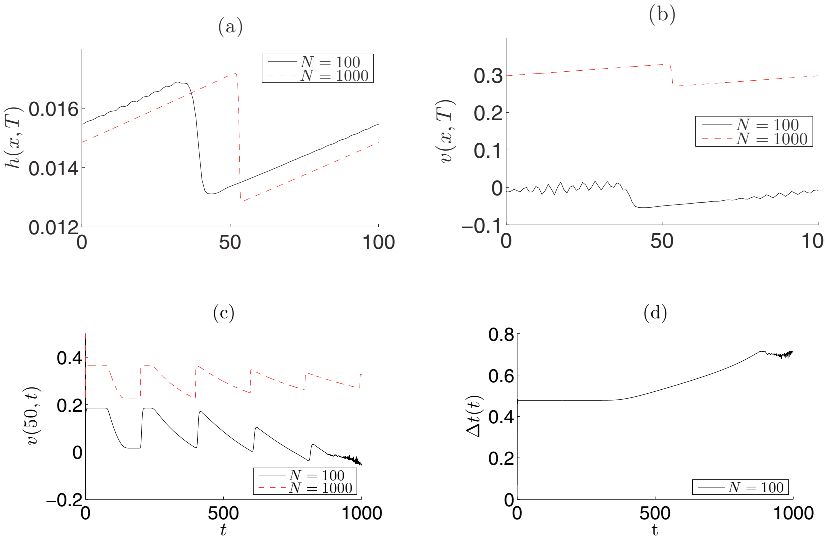

FIG. 8. *Example* 4*: Solution obtained using the IMEX-SSP*3*(*3, 3, 2*) method: the water depth* (a) *and velocity* (b) *at time* $T = 1000.25$*; the velocity* (c) *at* $x = 50$ *as a function of time. The size of time steps* (d) *as a function of time.*

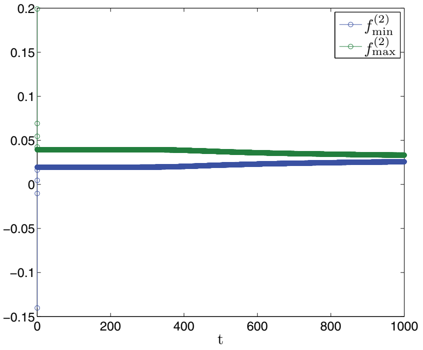

FIG. 9. *Example* 4*: The values of* $f^{(2)}_{\rm min}$ *and* $f^{(2)}_{\rm max}$ *computed by the SI-RK*3 *method with* $N = 100$ *and CFL number* 0.3 *as functions of time.*

method changes, as one can clearly see from Figure 8(c). Therefore, having a sign preserving method is advantageous for computing solutions of shallow water equations with friction terms.

Example 5. *Fixed time step restriction.* We have demonstrated in Example 2 that the SI-RK3 method converges to the exact equilibrium as long as the stability restriction is satisfied while the IMEX-SSP3(3,3,2) method delivers different numerical equilibria depending on the time step adopted within the stability region. When both ODE solvers are applied to the system of ODEs arising from the central-upwind semi-discretization of the shallow water system, the computed water depth h and velocity v are affected by the choice of Δt in a similar manner. To illustrate this, we use a combination of the CFL condition and fixed time restriction,

$$\Delta t = \min\{\Delta t_{\mathrm{CFL}}, \Delta t_{\max}\},$$

and implement the SI-RK3 and IMEX-SSP3(3,3,2) with $\Delta t_{\max} = 0.3, 0.15$ and 0.01.

In Figures 10(a)–(c), we show $h(x,T)$, $v(x,T)$, and $v(50,t)$, computed using the SI-RK3 time integration method. As one can observe, the results obtained with different $\Delta t_{\max}$'s are visually indistinguishable. In addition, a satisfactory accuracy of the obtained solution is confirmed by comparing it with the reference ones computed by the same SI-RK3 method with $N = 1000$ and Δt determined by the CFL condition. The Δt profile depicted in Figure 10(d) indicates that Δt is effectively controlled by $\Delta t_{\max}$ and is thus constant for almost all t.

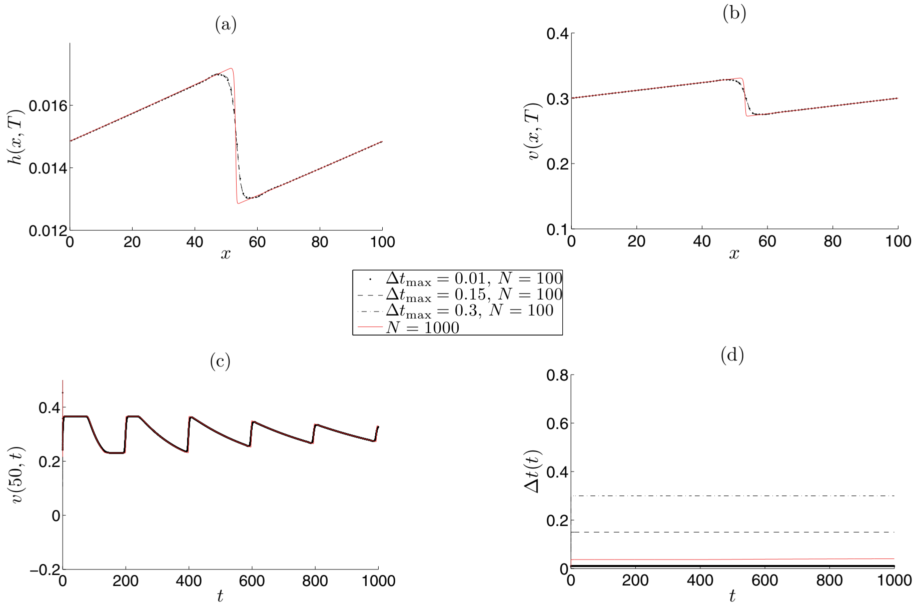

FIG. 10. *Example* 5: *Solutions obtained using the SI-RK*3 *method with fixed time step restrictions: the water depth* (a) *and velocity* (b) *at time $T = 1000$; the velocity* (c) *at $x = 50$ as a function of time. The size of time steps* (d) *as a function of time.*

The results obtained using the IMEX-SSP3(3,3,2) time integration are shown in Figure 11 and are free of oscillations in both time and space due to the fixed time step employed. However, a great disagreement can be observed when a large value of $\Delta t_{\max}$ is used. When the time step decreases, one can still notice a decreasing error (both in velocity magnitude and shock speed), which indicates that the dominating error is introduced by the time integration method. Of course, once a very small time step ($\Delta t_{\max} = 0.01$) or a very fine mesh is used, the results obtained become comparable with those obtained by the SI-RK3 method.

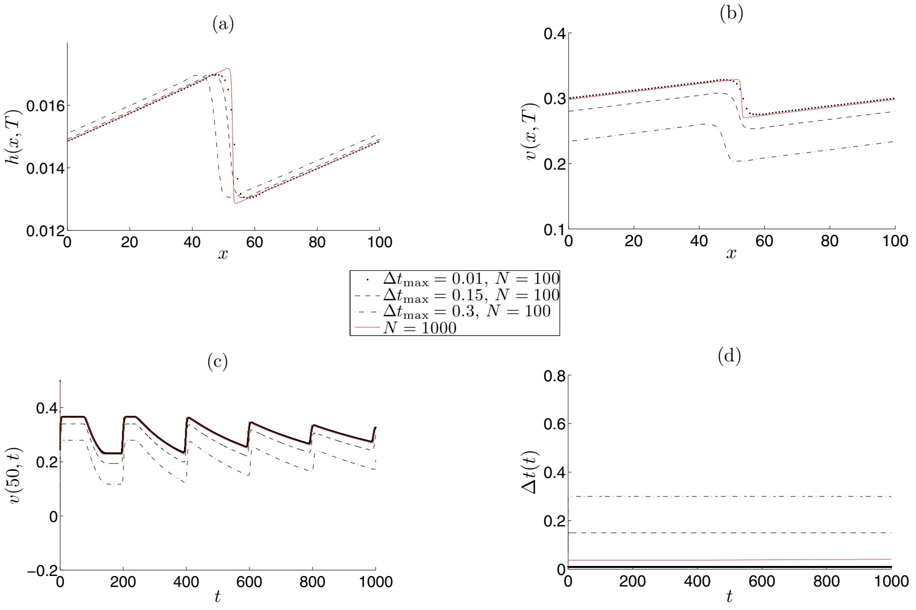

FIG. 11. *Example* 5: *Solutions obtained using the IMEX-SSP*3*(*3, 3, 2*) method with fixed time step restrictions: the water depth* (a) *and velocity* (b) *at time* $T = 1000$*; the velocity* (c) *at* $x = 50$ *as a function of time. The size of time steps* (d) *as a function of time.*

Acknowledgment. The authors thank Dr. Tewodros Amdeberhan for fruitful discussions.

REFERENCES

[1] R. Alexander, *Diagonally implicit Runge-Kutta methods for stiff ODE's*, SIAM J. Numer. Anal., 14 (1977), pp. 1006–1021.

[2] U. M. Ascher, S. J. Ruuth, and R. J. Spiteri, *Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations*, Appl. Numer. Math., 25 (1997), pp. 151–167.

[3] V. Casulli, *Semi-implicit finite difference methods for the two-dimensional shallow water equations*, J. Comput. Phys., 86 (1990), pp. 56–74.

[4] L. Cea and M. E. Vázquez-Cendón, *Unstructured finite volume discretisation of bed friction and convective flux in solute transport models linked to the shallow water equations*, J. Comput. Phys., 231 (2012), pp. 3317–3339.

[5] A. Chertock, S. Cui, A. Kurganov, and T. Wu, *Well-balanced positivity preserving central-upwind scheme for the shallow water system with friction terms*, Internat. J. Numer. Methods Fluids, 78 (2015), pp. 355–383.

[6] H. Darcy, *Recherches expérimentales relatives au mouvement de l'eau dans les tuyaux*, Vol. 1, Mallet-Bachelier, 1857.

[7] A. J. C. de Saint-Venant, *Thèorie du mouvement non-permanent des eaux, avec application aux crues des rivière at à l'introduction des marèes dans leur lit.*, C.R. Acad. Sci. Paris, 73 (1871), pp. 147–154.

[8] G. Dimarco and L. Pareschi, *Asymptotic preserving implicit-explicit Runge-Kutta methods for nonlinear kinetic equations*, SIAM J. Numer. Anal., 51 (2013), pp. 1064–1087.

[9] A. Flamant, *Mécanique appliquée: Hydraulique*, Baudry éditeur, Paris, 1891.

[10] Ph. Gauckler, *Etudes Théoriques et Pratiques sur l'Ecoulement et le Mouvement des Eaux*, Gauthier-Villars, 1867.

[11] S. Gottlieb, D. I. Ketcheson, and C.-W. Shu, *Strong Stability Preserving Runge-Kutta and Multistep Time Discretizations*, World Scientific, Hackensack, NJ, 2011.

[12] S. Gottlieb, C.-W. Shu, and E. Tadmor, *Strong stability-preserving high-order time discretization methods*, SIAM Rev., 43 (2001), pp. 89–112.

[13] K. Heun, *Neue Methoden zur approximativen Integration der Differentialgleichungen einer unabhängigen Veränderlichen*, Z. Math. Phys, 45 (1900), pp. 23–38.

[14] I. Higueras, N. Happenhofer, O. Koch, and F. Kupka, *Optimized strong stability preserving IMEX Runge–Kutta methods*, J. Comput. Appl. Math., 272 (2014), pp. 116–140.

[15] I. Higueras and T. Roldán, *Positivity-preserving and entropy-decaying IMEX methods*, in Ninth International Conference Zaragoza-Pau on Applied Mathematics and Statistics, Monogr. Semin. Mat. García Galdeano 33, Prensas University Zaragoza, Zaragoza, Spain, 2006, pp. 129–136.

[16] W. Hundsdorfer and S. J. Ruuth, *IMEX extensions of linear multistep methods with general monotonicity and boundedness properties*, J. Comput. Phys., 225 (2007), pp. 2016–2042.

[17] A. Kurganov and D. Levy, *Central-upwind schemes for the Saint-Venant system*, M2AN Math. Model. Numer. Anal., 36 (2002), pp. 397–425.

[18] A. Kurganov and G. Petrova, *A second-order well-balanced positivity preserving central-upwind scheme for the Saint-Venant system*, Commun. Math. Sci., 5 (2007), pp. 133–160.

[19] R. Manning, *On the flow of water in open channel and pipes*, Trans. Inst. Civil Engineers of Ireland, 20 (1891), pp. 161–207.

[20] L. Pareschi and G. Russo, *Implicit-explicit Runge-Kutta schemes for stiff systems of differential equations*, in Recent Trends in Numerical Analysis, Adv. Theory Comput. Math. 3, Nova Science Publishers, Huntington, NY, 2001, pp. 269–288.

[21] L. Pareschi and G. Russo, *Implicit-Explicit Runge-Kutta schemes and applications to hyperbolic systems with relaxation*, J. Sci. Comput., 25 (2005), pp. 129–155.

[22] C.-W. Shu, *Total-variation-diminishing time discretizations*, SIAM J. Sci. Comput., 6 (1988), pp. 1073–1084.

[23] C.-W. Shu and S. Osher, *Efficient implementation of essentially non-oscillatory shock-capturing schemes*, J. Comput. Phys., 77 (1988), pp. 439–471.

[24] P. Song, J.-S. Pang, and V. Kumar, *A semi-implicit time-stepping model for frictional compliant contact problems*, Internat. J. Numer. Methods Engrg., 60 (2004), pp. 2231–2261.

[25] R. J. Spiteri and S. J. Ruuth, *Non-linear evolution using optimal fourth-order strong-stability-preserving Runge-Kutta methods*, Math. Comput. Simulation, 62 (2003), pp. 125–135.

[26] X. Zhong, *Additive semi-implicit Runge-Kutta methods for computing high-speed nonequilibrium reactive flows*, J. Comput. Phys., 128 (1996), pp. 19–31.

[27] Z. Zlatev, *Modified diagonally implicit Runge-Kutta methods*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 321–334.