Journal of Computational Physics 444 (2021) 110543

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

High order sign-preserving and well-balanced exponential Runge-Kutta discontinuous Galerkin methods for the shallow water equations with friction

Ruize Yang [a], Yang Yang [b,1], Yulong Xing [a,*,2]

[a] Department of Mathematics, The Ohio State University, Columbus, OH 43210, USA
[b] Department of Mathematical Sciences, Michigan Technological University, Houghton, MI 49931, USA

ARTICLE INFO

Article history:
Available online 7 July 2021

Keywords:
Shallow water equations
Stiff friction terms
Discontinuous Galerkin method
Sign-preserving
Well-balanced
High order accuracy

ABSTRACT

In this paper, we propose a family of second and third order temporal integration methods for systems of stiff ordinary differential equations, and explore their application in solving the shallow water equations with friction. The new temporal discretization methods come from a combination of the traditional Runge-Kutta method (for non-stiff equation) and exponential Runge-Kutta method (for stiff equation), and are shown to have both the sign-preserving and steady-state-preserving properties. They are combined with the well-balanced discontinuous Galerkin spatial discretization to solve the nonlinear shallow water equations with non-flat bottom topography and (stiff) friction terms. We have demonstrated that the fully discrete schemes satisfy the well-balanced, positivity-preserving and sign-preserving properties simultaneously. The proposed methods have been tested and validated on several one- and two-dimensional test cases, and good numerical results have been observed.

1. Introduction

In this paper, we design sign-preserving and well-balanced exponential Runge-Kutta discontinuous Galerkin (DG) schemes for the system of nonlinear shallow water equations (SWEs) with a non-flat bottom topography and a Manning friction term. The SWEs, derived from the Navier-Stokes equation describing the motion of fluids, are a system of hyperbolic PDEs governing fluid flow in the oceans, coastal regions, estuaries, rivers and channels. They can be used to predict tides, storm surge levels and coastline changes from hurricanes, ocean currents, and also arise in atmospheric flows and debris flows. The two-dimensional shallow water equations take the form

* Corresponding author.
E-mail addresses: yang.4097@osu.edu (R. Yang), yyang7@mtu.edu (Y. Yang), xing.205@osu.edu (Y. Xing).
[1] The work of this author was partially supported by the NSF grant DMS-1818467.
[2] The work of this author was partially supported by the NSF grant DMS-1753581.

https://doi.org/10.1016/j.jcp.2021.110543
0021-9991/

$$\begin{cases} h_t + q_x + p_y = 0, \\ q_t + \left(hu^2 + \frac{1}{2}gh^2\right)_x + (huv)_y = -ghb_x - gn^2 \frac{q\sqrt{q^2+p^2}}{h^\eta}, \\ p_t + (huv)_x + \left(hv^2 + \frac{1}{2}gh^2\right)_y = -ghb_y - gn^2 \frac{p\sqrt{q^2+p^2}}{h^\eta}, \end{cases} \tag{1.1}$$

where h denotes the water depth, $(u, v)^T$ is the velocity vector, $q := hu$ and $p := hv$ are the discharges and g is the gravitational acceleration constant. In one-dimensional case, the system is reduced to

$$\begin{cases} h_t + q_x = 0, \\ q_t + \left(hu^2 + \frac{1}{2}gh^2\right)_x = -ghb_x - gn^2 \frac{|q|q}{h^\eta}. \end{cases} \tag{1.2}$$

The discharge equation contains two source terms on the right-hand side. The first term is the geometric source with b representing the bottom topography, and the second term models the bottom friction with n being the Manning coefficient and the parameter η chosen as $7/3$ in this paper.

The nonlinear SWEs belong to the family of hyperbolic conservation laws with source term, also referred as hyperbolic balance laws, which have gained growing attention in the last few decades. The one-dimensional hyperbolic balance law is given by

$$U_t + F(U)_x = S(U),$$

and introduce new computational challenges beyond the existing challenges of hyperbolic conservation laws, due to the existence of the source term $S(U)$. They often admit non-trivial steady state solutions in which the source term balances the effect of the flux gradients. Such balance may not be well captured by standard numerical methods and introduce spurious oscillation, making it challenging to simulate steady state solutions or their small perturbations unless a much refined mesh is used. The well-balanced methods are introduced to exactly preserve equilibrium solutions at the discrete level and resolve small perturbations to steady state solutions accurately on a relatively coarse mesh. The still-water steady state of the SWEs is given by

$$h(x,t) \equiv C - b(x), \quad q(x,t) \equiv 0, \tag{1.3}$$

which represents a still flat water surface. Many interesting physical phenomena are small perturbations of this steady state. The well-balanced methods were first designed for the SWEs by Bermudez and Vazquez [4] in 1994. Since then, many well-balanced methods [1,7,8,14,19,20,22,23,25,27,28] have been studied, and we refer to the survey papers [18,32] and the references therein for more works on this topic.

Another well-known challenge in numerically solving the SWEs appears at the wetting-drying front. Physically, the water height h should be non-negative, however, standard numerical methods may produce unacceptable negative water height in dry or nearly dry regions. In [34], high order positivity-preserving DG methods were designed for the SWEs by introducing a positivity-preserving limiter, which also preserves the higher order accuracy without losing local conservation. The well-balanced property of the resulting methods was also investigated in [34], and the extension to high order finite volume weighted essentially non-oscillatory methods was studied in [31]. We refer to [1,3,13,17,19] for more existing numerical methods which maintain both well-balanced and positivity-preserving properties at the same time.

The SWEs without the friction term were considered in [34], when the positivity-preserving high order well-balanced DG methods were designed. When the friction term is included, one could simply treat it explicitly in the existing framework. However, when the region is nearly dry, i.e., h is small, the friction term in the discharge equation becomes a stiff source term, and a tiny time step size is needed for the standard explicit numerical methods to be stable. One approach is to utilize the implicit-explicit Runge-Kutta scheme and treat the friction term implicitly, which may lead to solving a nonlinear system and could be time-consuming. Some tricks could be used to save the computational cost, for example, finite volume methods with semi-implicit time integration are developed to handle such issue in [5], [6] and [26]. Second order well-balanced finite element method for the SWEs with friction is studied in [14], where a regularization term is added to the explicit approximation of the friction term. Other well-balanced scheme for the SWEs with Manning friction can be found in [9,17,21].

Our main focus in this work is to present efficient high order well-balanced and positivity-preserving DG methods for the SWEs with non-flat bottom and friction term. We will propose a new temporal discretization for the stiff or partially stiff system of ordinary differential equations, which can take larger time step size, and at the same time won't affect the steady-state-preserving property. This is achieved by a combination of Runge-Kutta (RK) method for the non-stiff part and exponential RK methods for the stiff component. We will provide the rigorous analysis to show that the new method maintains the same order of accuracy as the underlying RK method. The novel temporal integration is then combined with the well-balanced and positivity-preserving DG methods in [34] to provide an efficient solver for the SWEs with friction

(1.2). The new time discretization is also shown to be sign-preserving, which means that the sign of the computed solution is determined by the non-stiff part of the system only. In other words, if the non-stiff part (i.e., without friction term) is used to update the numerical velocity and it stays positive (or negative), adding the friction term to the system will not change this sign. As illustrated in [5], maintaining the sign-preserving property is crucial, otherwise, large numerical error may appear in the simulation. Both one- and two-dimensional numerical results demonstrate that the proposed method yields desirable results even on coarse grids.

The rest of the paper is organized as follows. We start with a quick introduction to exponential RK method and then propose our new sign-preserving and steady-state-preserving time integrations in Section 2. In Section 3, we start by reviewing the semi-discrete well-balanced DG methods for the SWEs and the construction of positivity-preserving limiter, and then explain the application of the exponential RK method to obtain the fully discrete methods for the SWEs, which are well-balanced, positivity-preserving and sign-preserving simultaneously. Numerical examples are shown in Section 4 to verify the accuracy of our new scheme and demonstrate the behavior of the proposed exponential RK and DG methods for SWEs. Conclusion remarks are given in Section 5.

2. New sign-preserving time integration

In this section, a novel sign-preserving time integration is discussed and analyzed for a system of stiff or partially stiff ordinary differential equations (ODEs).

As a prototype example, we consider the following ODE system

$$\mathbf{w}_t = \mathbf{L}(\mathbf{w}) + \mathbf{s}(\mathbf{w}), \tag{2.1}$$

where $\mathbf{w} = (w_1, ..., w_l)^T$ is the unknown variable, $\mathbf{L} = (L_1, ..., L_l)$ is a linear or nonlinear operator (which could come from the spatial discretization of the flux term) and $\mathbf{s} = (s_1, ..., s_l)^T$ is the source term, which might be stiff. For simplicity, we call a vector $\mathbf{v}$ to be nonnegative ($\mathbf{v} \geqslant 0$) if each component in this vector is nonnegative. In the following discussion, we let $0 = t_0 \leq t_1 \leq \cdots \leq t_N = T$ be a partition of the entire time interval $[0, T]$ with the time step $\Delta t^n = t_n - t_{n-1}$, $n = 1, 2, \cdots, N$. We use the notation $\mathbf{w}^n$ to represent the numerical solution at the n-th time step t^n. In many applications, the solutions to (2.1) reflect significant physical meaning of the underlying model. In particular, steady state and sign of solutions are of interest. Our goal is to construct a suitable high order scheme that enjoys the following properties.

1. Steady-state-preserving: If $\mathbf{L}(\mathbf{w}^n) + \mathbf{s}(\mathbf{w}^n) = 0$, then $\mathbf{w}^{n+1} = \mathbf{w}^n$;
2. Sign-preserving: Suppose $\mathbf{w}^n \geqslant 0$, we have $\mathbf{w}^{n+1} \geqslant 0$;
3. Time step size: Even in the presence of stiff term $\mathbf{s}(\mathbf{w})$, small time step is not required.

The sign-preserving property refers to the ability of the numerical method to preserve the sign of the numerical solution when the exact solution is always non-negative (or non-positive). We refer to [5] for the detailed explanation of this property. Below, we will first review the exponential RK method introduced in [11], and then discuss our new time integration by combining it with traditional RK method.

2.1. Exponential sign-preserving discretization

We start by presenting the exponential RK method for the ODE system (2.1). By introducing the exponential factor, we can derive a new ODE equation

$$(e^{\mu t}\mathbf{w})_t = e^{\mu t}(\mathbf{L}(\mathbf{w}) + \mathbf{s}(\mathbf{w}) + \mu\mathbf{w}),$$

and the general framework of exponential RK scheme is given by [16] (three-stage RK method is presented below as an example)

$$\begin{aligned}
\mathbf{w}^{(1)} =& e^{-\beta_{10}\mu k}\left[\alpha_{10}\mathbf{w}^n + \beta_{10}k(\mathbf{L}(\mathbf{w}^n) + \mathbf{s}(\mathbf{w}^n) + \mu\mathbf{w}^n)\right],\\
\mathbf{w}^{(2)} =& e^{-A\mu k}\left[\alpha_{20}\mathbf{w}^n + \beta_{20}k(\mathbf{L}(\mathbf{w}^n) + \mathbf{s}(\mathbf{w}^n) + \mu\mathbf{w}^n)\right]\\
&+ e^{(\beta_{10}-A)\mu k}\left[\alpha_{21}\mathbf{w}^{(1)} + \beta_{21}k(\mathbf{L}(\mathbf{w}^{(1)}) + \mathbf{s}(\mathbf{w}^{(1)}) + \mu\mathbf{w}^{(1)})\right],\\
\mathbf{w}^{n+1} =& e^{-\mu k}\left[\alpha_{30}\mathbf{w}^n + \beta_{30}k(\mathbf{L}(\mathbf{w}^n) + \mathbf{s}(\mathbf{w}^n) + \mu\mathbf{w}^n)\right]\\
&+ e^{(\beta_{10}-1)\mu k}\left[\alpha_{31}\mathbf{w}^{(1)} + \beta_{31}k(\mathbf{L}(\mathbf{w}^{(1)}) + \mathbf{s}(\mathbf{w}^{(1)}) + \mu\mathbf{w}^{(1)})\right]\\
&+ e^{(A-1)\mu k}\left[\alpha_{32}\mathbf{w}^{(2)} + \beta_{32}k(\mathbf{L}(\mathbf{w}^{(2)}) + \mathbf{s}(\mathbf{w}^{(2)}) + \mu\mathbf{w}^{(2)})\right],
\end{aligned} \tag{2.2}$$

where $A = \beta_{20} + \alpha_{21}\beta_{10} + \beta_{21}$ and $k = \Delta t^n$ denotes the time step. All of the coefficients, α_{ij} and β_{ij}, are positive constants to be determined by the order conditions, and μ is a nonnegative constant to be determined by the sign-preserving property.

It is easy to see that steady-state-preserving property does not hold for this scheme. Therefore, one can modify (2.2) and construct

$$\begin{aligned}
\mathbf{w}^{(1)} &= \left[\alpha_{10}\mathbf{w}^n + \beta_{10}k(\mathbf{L}(\mathbf{w}^n) + \mathbf{s}(\mathbf{w}^n) + \mu\mathbf{w}^n)\right]/A_1,\\
\mathbf{w}^{(2)} &= \left[\alpha_{20}\mathbf{w}^n + \beta_{20}k(\mathbf{L}(\mathbf{w}^n) + \mathbf{s}(\mathbf{w}^n) + \mu\mathbf{w}^n)\right]/A_2\\
&\quad + e^{\beta_{10}\mu k}\left[\alpha_{21}\mathbf{w}^{(1)} + \beta_{21}k(\mathbf{L}(\mathbf{w}^{(1)}) + \mathbf{s}(\mathbf{w}^{(1)}) + \mu\mathbf{w}^{(1)})\right]/A_2,\\
\mathbf{w}^{n+1} &= \left[\alpha_{30}\mathbf{w}^n + \beta_{30}k(\mathbf{L}(\mathbf{w}^n) + \mathbf{s}(\mathbf{w}^n) + \mu\mathbf{w}^n)\right]/A_3\\
&\quad + e^{\beta_{10}\mu k}\left[\alpha_{31}\mathbf{w}^{(1)} + \beta_{31}k(\mathbf{L}(\mathbf{w}^{(1)}) + \mathbf{s}(\mathbf{w}^{(1)}) + \mu\mathbf{w}^{(1)})\right]/A_3\\
&\quad + e^{A\mu k}\left[\alpha_{32}\mathbf{w}^{(2)} + \beta_{32}k(\mathbf{L}(\mathbf{w}^{(2)}) + \mathbf{s}(\mathbf{w}^{(2)}) + \mu\mathbf{w}^{(2)})\right]/A_3,
\end{aligned} \tag{2.3}$$

where

$$\begin{aligned}
&A_1 = \alpha_{10} + \beta_{10}\mu k, \quad A_2 = [\alpha_{20} + \beta_{20}\mu k] + e^{\beta_{10}\mu k}[\alpha_{21} + \beta_{21}\mu k],\\
&A_3 = [\alpha_{30} + \beta_{30}\mu k] + e^{\beta_{10}\mu k}[\alpha_{31} + \beta_{31}\mu k] + e^{A\mu k}[\alpha_{32} + \beta_{32}\mu k].
\end{aligned}$$

For this scheme, we have the following property:

Proposition 2.1. *The exponential RK scheme of the form (2.3) is steady-state-preserving.*

Proof. When the steady state is reached, i.e., $\mathbf{L}(\mathbf{w}^n) + \mathbf{s}(\mathbf{w}^n) = 0$, we have

$$\begin{aligned}
\mathbf{w}^{(1)} &= (\alpha_{10}\mathbf{w}^n + \beta_{10}k\mu\mathbf{w}^n)/A_1 = \mathbf{w}^n\\
\mathbf{w}^{(2)} &= (\alpha_{20}\mathbf{w}^n + \beta_{20}k\mu\mathbf{w}^n)/A_2 + e^{\beta_{10}\mu k}(\alpha_{21}\mathbf{w}^{(1)} + \beta_{21}k\mu\mathbf{w}^{(1)})/A_2 = \mathbf{w}^n\\
\mathbf{w}^{n+1} &= (\alpha_{30}\mathbf{w}^n + \beta_{30}k\mu\mathbf{w}^n)/A_3 + e^{\beta_{10}\mu k}(\alpha_{31}\mathbf{w}^{(1)} + \beta_{31}k\mu\mathbf{w}^{(1)})/A_3\\
&\quad + e^{A\mu k}(\alpha_{32}\mathbf{w}^{(2)} + \beta_{32}k\mu\mathbf{w}^{(2)})/A_3 = \mathbf{w}^n,
\end{aligned}$$

as desired, which finishes the proof. □

The sufficient and necessary conditions to obtain third-order accuracy were studied in [11] following the idea in [24]. We first make the following positivity assumption:

There exists a sufficiently small k_E such that if $\boldsymbol{w} \geqslant 0$ and $k \leqslant k_E$, one has

$$\boldsymbol{w} + k\boldsymbol{L}(\boldsymbol{w}) \geqslant 0. \tag{2.4}$$

The following result in [11] demonstrates the sign-preserving property of the proposed exponential RK scheme.

Theorem 2.2. *Consider the ODE system (2.1) with the flux $\mathbf{L}$ satisfying (2.4). The scheme (2.3) is sign-preserving: if $\mathbf{w}^n \geqslant 0$, we can conclude that $\mathbf{w}^{n+1} \geqslant 0$ under the conditions*

$$\mu \geqslant \max_{1\leqslant i\leqslant l}\left\{-\frac{s_i(\boldsymbol{w}^n)}{w_i^n} - \frac{s_i(\boldsymbol{w}^{(1)})}{w_i^{(1)}}, -\frac{s_i(\boldsymbol{w}^{(2)})}{w_i^{(2)}}, 0\right\} \quad \text{and} \quad k \leqslant \zeta k_E, \tag{2.5}$$

where

$$\zeta = \min\left\{\frac{\alpha_{10}}{\beta_{10}}, \frac{\alpha_{20}}{\beta_{20}}, \frac{\alpha_{21}}{\beta_{21}}, \frac{\alpha_{30}}{\beta_{30}}, \frac{\alpha_{31}}{\beta_{31}}, \frac{\alpha_{32}}{\beta_{32}}\right\}.$$

Based on the above theorem, we would like the value of ζ to be as large as possible. As studied in [11], the optimal coefficients are

$$\begin{aligned}
&\alpha_{10} = 1, \quad \beta_{10} = 0.7071933376925014,\\
&\alpha_{20} = 0.6686892933074404, \quad \beta_{20} = 0,\\
&\alpha_{21} = 0.3313107066925596, \quad \beta_{21} = 0.4178047564915065,\\
&\alpha_{30} = 0.3487419430256090, \quad \beta_{30} = 0,\\
&\alpha_{31} = 0.2039576138780898, \quad \beta_{31} = 0,\\
&\alpha_{32} = 0.4473004430963011, \quad \beta_{32} = 0.5640754637100439,
\end{aligned} \tag{2.6}$$

with $\zeta = 0.7929797388491311$. We observe that with this collection of coefficients (2.6), the numerical scheme (2.3) becomes

$$\begin{aligned}
\mathbf{w}^{(1)} =& \mathbf{w}^n + \beta_{10} k(\mathbf{L}(\mathbf{w}^n) + \mathbf{s}(\mathbf{w}^n))/A_1,\\
\mathbf{w}^{(2)} =& \alpha_{20}\mathbf{w}^n/A_2 + e^{\beta_{10}\mu k}\left[\alpha_{21}\mathbf{w}^{(1)} + \beta_{21} k(\mathbf{L}(\mathbf{w}^{(1)}) + \mathbf{s}(\mathbf{w}^{(1)}) + \mu\mathbf{w}^{(1)})\right]/A_2,\\
\mathbf{w}^{n+1} =& \alpha_{30}\mathbf{w}^n/A_3 + e^{\beta_{10}\mu k}\alpha_{31}\mathbf{w}^{(1)}/A_3\\
&+ e^{A\mu k}\left[\alpha_{32}\mathbf{w}^{(2)} + \beta_{32} k(\mathbf{L}(\mathbf{w}^{(2)}) + \mathbf{s}(\mathbf{w}^{(2)}) + \mu\mathbf{w}^{(2)})\right]/A_3,
\end{aligned} \tag{2.7}$$

after dropping the terms with zero coefficients. These exponential functions could be extremely large for large μ, and in practical implementation we often use the following approximation

$$e^x \approx \left(1 - x + \frac{1}{2}x^2 - \frac{1}{6}x^3 + \frac{1}{24}x^4\right)^{-1}, \quad x = \beta_{10}\mu k \text{ or } A\mu k.$$

If $\mu = 0$, this exponential RK method reduces to the standard third order RK method:

$$\begin{aligned}
\mathbf{w}^{(1)} =& w^n + \beta_{10} k(\mathbf{L}(\mathbf{w}^n) + \mathbf{s}(\mathbf{w}^n)),\\
\mathbf{w}^{(2)} =& \alpha_{20}\mathbf{w}^n + \alpha_{21}\mathbf{w}^{(1)} + \beta_{21} k(\mathbf{L}(\mathbf{w}^{(1)}) + \mathbf{s}(\mathbf{w}^{(1)})),\\
\mathbf{w}^{n+1} =& \alpha_{30}\mathbf{w}^n + \alpha_{31}\mathbf{w}^{(1)} + \alpha_{32}\mathbf{w}^{(2)} + \beta_{32} k(\mathbf{L}(\mathbf{w}^{(2)}) + \mathbf{s}(\mathbf{w}^{(2)})).
\end{aligned} \tag{2.8}$$

As $\beta_{ik} \geqslant 0$, this RK method is strong stability preserving, according to [15, Lemma 2.1].

For completeness, we recall the second-order version of the scheme given in [12]:

$$\begin{aligned}
\mathbf{w}^{(1)} =& B_1^1\left[\mathbf{w}^n + k(\mathbf{L}(\mathbf{w}^n) + \mathbf{s}(\mathbf{w}^n) + \mu\mathbf{w}^n)\right],\\
\mathbf{w}^{n+1} =& B_2^1\mathbf{w}^n + B_2^2\left[\mathbf{w}^{(1)} + k(\mathbf{L}(\mathbf{w}^{(1)}) + \mathbf{s}(\mathbf{w}^{(1)}) + \mu\mathbf{w}^{(1)})\right],
\end{aligned} \tag{2.9}$$

where

$$B_1^1 = \frac{1 - \mu k + \frac{1}{2}(\mu k)^2}{1 - \frac{1}{2}(\mu k)^2 + \frac{1}{2}(\mu k)^3}, \quad B_2^1 = \frac{1}{2}\frac{1 - \mu k + \frac{1}{2}(\mu k)^2}{1 + \frac{1}{4}(\mu k)^2}, \quad B_2^2 = \frac{1}{2}\frac{1}{1 + \frac{1}{4}(\mu k)^2}.$$

One can easily verify that the scheme is steady-state-preserving, sign-preserving and of order $O(k^2)$, and we refer to [12, Theorem 5.1] for the detailed proof.

2.2. Novel sign-preserving discretization for system of equations

The exponential RK method in the previous subsection is appropriate for stiff ODEs. When extending such method to system of ODEs, we may encounter the case when different equations have different stiffness, or some equations are stiff and the others are not. If the same coefficient μ is used for all the equations, the non-stiff or less-stiff equations will also be approximated by the same exponential RK method, and this may lead to large computational error. One numerical example will be provided in Section 4 to illustrate such large errors.

This motivates us to develop a new scheme for this kind of system, such that exponential RK method is applied to the stiff equation and standard RK method is applied to the non-stiff equation. If the equations with different stiffness are encountered, we may apply the exponential RK method with different μ (which could be 0 when non-stiff equation is considered). Let us present the method using the simple example of two sets of equations. Suppose $\mathbf{w} = \begin{pmatrix}\mathbf{u}\\ \mathbf{v}\end{pmatrix}$ in (2.1), and the system of ODEs becomes

$$\begin{cases}
\mathbf{u}'(t) = \mathbf{L}_1(\mathbf{u}, \mathbf{v}) + \mathbf{s}_1(\mathbf{u}, \mathbf{v}),\\
\mathbf{v}'(t) = \mathbf{L}_2(\mathbf{u}, \mathbf{v}) + \mathbf{s}_2(\mathbf{u}, \mathbf{v}),
\end{cases} \tag{2.10}$$

where $\mathbf{u} = (u_1, ..., u_{l_1})^T$, $\mathbf{v} = (v_1, ..., v_{l_2})^T$ are the unknown variables. As the concrete forms of $\mathbf{L}_i$ and $\mathbf{s}_i$ $(i = 1, 2)$ are of no importance in this subsection, we simplify the notation further, by denoting their sums by $\mathbf{F}$ and $\mathbf{G}$. Now we consider the system of ODEs given by

$$\begin{cases}
\mathbf{u}'(t) = \mathbf{F}(\mathbf{u}, \mathbf{v}),\\
\mathbf{v}'(t) = \mathbf{G}(\mathbf{u}, \mathbf{v}),
\end{cases} \tag{2.11}$$

where the terms $\mathbf{F} = (f_1, ..., f_{l_1})$ and $\mathbf{G} = (g_1, ..., g_{l_2})$ may contain stiff term of different magnitude.

We start by presenting the second-order scheme. If the first equation is non-stiff, i.e., $\mathbf{F}$ does not contain stiff term, we can apply the Heun's method to the first equation and the exponential RK method (2.9) to the second (stiff) equation in (2.11), which leads to the method of the form

$$\begin{aligned}
\mathbf{u}^{(1)} =& \mathbf{u}^n + k\mathbf{F}(\mathbf{u}^n, \mathbf{v}^n),\\
\mathbf{v}^{(1)} =& \mathbf{v}^n + \frac{k}{1+a}\mathbf{G}(\mathbf{u}^n, \mathbf{v}^n),\\
\mathbf{u}^{n+1} =& \frac{1}{2}\mathbf{u}^n + \frac{1}{2}\left[\mathbf{u}^{(1)} + k\mathbf{F}(\mathbf{u}^{(1)}, \mathbf{v}^{(1)})\right],\\
\mathbf{v}^{n+1} =& \frac{1}{2}\frac{1-a+\frac{a^2}{2}}{1+\frac{a^2}{4}}\mathbf{v}^n + \frac{1}{2\left(1+\frac{a^2}{4}\right)}\left[(1+a)\mathbf{v}^{(1)} + k\mathbf{G}(\mathbf{u}^{(1)}, \mathbf{v}^{(1)})\right],
\end{aligned} \tag{2.12}$$

with $a = \mu k$ with μ specified by the condition (2.5) to ensure the sign-preserving property. When both equations contain stiff term while with different magnitude, we can compute μ_1 and μ_2 for each equation. Let $a_1 = \mu_1 k$ and $a_2 = \mu_2 k$, and we have the following numerical scheme

$$\begin{aligned}
\mathbf{u}^{(1)} =& \mathbf{u}^n + \frac{k}{1+a_1}\mathbf{F}(\mathbf{u}^n, \mathbf{v}^n),\\
\mathbf{v}^{(1)} =& \mathbf{v}^n + \frac{k}{1+a_2}\mathbf{G}(\mathbf{u}^n, \mathbf{v}^n),\\
\mathbf{u}^{n+1} =& \frac{1}{2}\frac{1-a_1+\frac{a_1^2}{2}}{1+\frac{a_1^2}{4}}\mathbf{v}^n + \frac{1}{2\left(1+\frac{a_1^2}{4}\right)}\left[(1+a_1)\mathbf{v}^{(1)} + k\mathbf{F}(\mathbf{u}^{(1)}, \mathbf{v}^{(1)})\right],\\
\mathbf{v}^{n+1} =& \frac{1}{2}\frac{1-a_2+\frac{a_2^2}{2}}{1+\frac{a_2^2}{4}}\mathbf{v}^n + \frac{1}{2\left(1+\frac{a_2^2}{4}\right)}\left[(1+a_2)\mathbf{v}^{(1)} + k\mathbf{G}(\mathbf{u}^{(1)}, \mathbf{v}^{(1)})\right],
\end{aligned} \tag{2.13}$$

for the ODE system (2.11). As the Heun's method and the exponential RK method (2.9) are both second-order accurate, we expect that the combination of them has the same order of accuracy. Indeed, we have the following theorem on its accuracy.

Theorem 2.3. *The numerical schemes* (2.13) *for the ODE system* (2.11) *is second-order accurate. In particular, the scheme* (2.12)*, as a special case when* $\mu_1 = 0$*, is also second-order accurate.*

Proof. For ease of presentation, we omit the superscript n unless otherwise listed. It follows that

$$\begin{aligned}
\mathbf{v}^{n+1} =& \frac{1}{2}\frac{1-a_2+\frac{a_2^2}{2}}{1+\frac{a_2^2}{4}}\mathbf{v} + \frac{1}{2\left(1+\frac{a_2^2}{4}\right)}\left[(1+a_2)(\mathbf{v} + \frac{k}{1+a_2}\mathbf{G}) + k\mathbf{G}(\mathbf{u} + \frac{k}{1+a_1}\mathbf{F}, \mathbf{v} + \frac{k}{1+a_2}\mathbf{G})\right]\\
=& \mathbf{v} + \frac{k}{2\left(1+\frac{a_2^2}{4}\right)}\left[\mathbf{G} + \mathbf{G} + \frac{k\mathbf{F}}{1+a_1}\mathbf{G}_\mathbf{x} + \frac{k\mathbf{G}}{1+a_2}\mathbf{G}_\mathbf{y} + O(k^2)\right]\\
=& \mathbf{v} + k\left(1+O(a_2^2)\right)\mathbf{G} + \frac{k^2}{2}\left(1+O(a_2^2)\right)\left[(1+O(a_1))\mathbf{F}\mathbf{G}_\mathbf{x} + (1+O(a_2))\mathbf{G}\mathbf{G}_\mathbf{y}\right] + O(k^3)\\
=& \mathbf{v} + k\mathbf{G} + \frac{k^2}{2}\left(\mathbf{F}\mathbf{G}_\mathbf{x} + \mathbf{G}\mathbf{G}_\mathbf{y}\right) + O\left(a_2^2 k, a_1 k^2, a_2 k^2, k^3\right).
\end{aligned}$$

Here we mimic the notations in the Taylor series in one variable, for example,

$$\mathbf{F}\mathbf{G}_\mathbf{x} := (\mathbf{F}\cdot g_{1\mathbf{x}}, ..., \mathbf{F}\cdot g_{l_2\mathbf{x}})^T,$$

and $g_{i\mathbf{x}}$ is a vector consisting of partial derivatives of g_i with respect to u_1 through u_{l_1}. By symmetry, one can compute

$$\mathbf{u}^{n+1} = \mathbf{u} + k\mathbf{F} + \frac{k^2}{2}\left(\mathbf{F}\mathbf{F}_\mathbf{x} + \mathbf{G}\mathbf{F}_\mathbf{y}\right) + O\left(a_1^2 k, a_1 k^2, a_2 k^2, k^3\right).$$

Recall that $a_i = \mu_i k$ $(i = 1, 2)$, hence $O(a_2^2 k) = O(a_1 k^2) = O(a_2 k^2) = O(k^3)$. Therefore, we have shown that the local truncation error is of order $O(k^2)$ for both $\mathbf{u}$ and $\mathbf{v}$. When $\mu_1 = 0$, the scheme reduces to (2.12), which is also second-order accurate. □

In the same fashion, we can design the combination of two third order methods, given by

$$
\begin{aligned}
\mathbf{u}^{(1)} =& \mathbf{u}^n + \beta_{10} k \mathbf{F}(\mathbf{u}^n, \mathbf{v}^n)/A_{11},\\
\mathbf{u}^{(2)} =& \left(\alpha_{20}\mathbf{u}^n + \beta_{20} k(\mathbf{F}(\mathbf{u}^n, \mathbf{v}^n) + \mu_1 \mathbf{u}^n)\right)/A_{21}\\
&+ e^{\beta_{10}\mu_1 k}\left(\alpha_{21}\mathbf{u}^{(1)} + \beta_{21} k(\mathbf{F}(\mathbf{u}^{(1)}, \mathbf{v}^{(1)}) + \mu_1 \mathbf{u}^{(1)})\right)/A_{21},\\
\mathbf{u}^{n+1} =& \left(\alpha_{30}\mathbf{u}^n + \beta_{30} k(\mathbf{F}(\mathbf{u}^n, \mathbf{v}^n) + \mu_1 \mathbf{u}^n)\right)/A_{31}\\
&+ e^{\beta_{10}\mu_1 k}\left(\alpha_{31}\mathbf{u}^{(1)} + \beta_{31} k(\mathbf{F}(\mathbf{u}^{(1)}, \mathbf{v}^{(1)}) + \mu_1 \mathbf{u}^{(1)})\right)/A_{31}\\
&+ e^{A\mu_1 k}\left(\alpha_{32}\mathbf{u}^{(2)} + \beta_{32} k(\mathbf{F}(\mathbf{u}^{(2)}, \mathbf{v}^{(2)}) + \mu_1 \mathbf{u}^{(2)})\right)/A_{31},\\
\mathbf{v}^{(1)} =& \mathbf{v}^n + \beta_{10} k \mathbf{G}(\mathbf{u}^n, \mathbf{v}^n)/A_{12},\\
\mathbf{v}^{(2)} =& \left(\alpha_{20}\mathbf{v}^n + \beta_{20} k(\mathbf{G}(\mathbf{u}^n, \mathbf{v}^n) + \mu_2 \mathbf{v}^n)\right)/A_{22}\\
&+ e^{\beta_{10}\mu_2 k}\left(\alpha_{21}\mathbf{v}^{(1)} + \beta_{21} k(\mathbf{G}(\mathbf{u}^{(1)}, \mathbf{v}^{(1)}) + \mu_2 \mathbf{v}^{(1)})\right)/A_{22},\\
\mathbf{v}^{n+1} =& \left(\alpha_{30}\mathbf{v}^n + \beta_{30} k(\mathbf{G}(\mathbf{u}^n, \mathbf{v}^n) + \mu_2 \mathbf{v}^n)\right)/A_{32}\\
&+ e^{\beta_{10}\mu_2 k}\left(\alpha_{31}\mathbf{v}^{(1)} + \beta_{31} k(\mathbf{G}(\mathbf{u}^{(1)}, \mathbf{v}^{(1)}) + \mu_2 \mathbf{v}^{(1)})\right)/A_{32}\\
&+ e^{A\mu_2 k}\left(\alpha_{32}\mathbf{v}^{(2)} + \beta_{32} k(\mathbf{G}(\mathbf{u}^{(2)}, \mathbf{v}^{(2)}) + \mu_2 \mathbf{v}^{(2)})\right)/A_{32},
\end{aligned}
\tag{2.14}
$$

where A is defined as before and

$$
\begin{aligned}
&A_{1i} = \alpha_{10} + \beta_{10}\mu_i k, \quad A_{2i} = [\alpha_{20} + \beta_{20}\mu_i k] + e^{\beta_{10}\mu_i k}[\alpha_{21} + \beta_{21}\mu_i k],\\
&A_{3i} = [\alpha_{30} + \beta_{30}\mu_i k] + e^{\beta_{10}\mu_i k}[\alpha_{31} + \beta_{31}\mu_i k] + e^{A\mu_i k}[\alpha_{32} + \beta_{32}\mu_i k],
\end{aligned}
$$

with $i = 1, 2$. In the case when the first equation is non-stiff, this reduces to the following temporal discretization method

$$
\begin{aligned}
\mathbf{u}^{(1)} =& \alpha_{10}\mathbf{u}^n + \beta_{10} k \mathbf{F}(\mathbf{u}^n, \mathbf{v}^n),\\
\mathbf{u}^{(2)} =& \alpha_{20}\mathbf{u}^n + \beta_{20} k \mathbf{F}(\mathbf{u}^n, \mathbf{v}^n) + \alpha_{21}\mathbf{u}^{(1)} + \beta_{21} k \mathbf{F}(\mathbf{u}^{(1)}, \mathbf{v}^{(1)}),\\
\mathbf{u}^{n+1} =& \alpha_{30}\mathbf{u}^n + \beta_{30} k \mathbf{F}(\mathbf{u}^n, \mathbf{v}^n) + \alpha_{31}\mathbf{u}^{(1)} + \beta_{31} k \mathbf{F}(\mathbf{u}^{(1)}, \mathbf{v}^{(1)}) + \alpha_{32}\mathbf{u}^{(2)} + \beta_{32} k \mathbf{F}(\mathbf{u}^{(2)}, \mathbf{v}^{(2)}),\\
\mathbf{v}^{(1)} =& \mathbf{v}^n + \beta_{10} k \mathbf{G}(\mathbf{u}^n, \mathbf{v}^n)/A_1,\\
\mathbf{v}^{(2)} =& \left(\alpha_{20}\mathbf{v}^n + \beta_{20} k(\mathbf{G}(\mathbf{u}^n, \mathbf{v}^n) + \mu \mathbf{v}^n)\right)/A_2\\
&+ e^{\beta_{10}\mu k}\left(\alpha_{21}\mathbf{v}^{(1)} + \beta_{21} k(\mathbf{G}(\mathbf{u}^{(1)}, \mathbf{v}^{(1)}) + \mu \mathbf{v}^{(1)})\right)/A_2,\\
\mathbf{v}^{n+1} =& \left(\alpha_{30}\mathbf{v}^n + \beta_{30} k(\mathbf{G}(\mathbf{u}^n, \mathbf{v}^n) + \mu \mathbf{v}^n)\right)/A_3\\
&+ e^{\beta_{10}\mu k}\left(\alpha_{31}\mathbf{v}^{(1)} + \beta_{31} k(\mathbf{G}(\mathbf{u}^{(1)}, \mathbf{v}^{(1)}) + \mu \mathbf{v}^{(1)})\right)/A_3\\
&+ e^{A\mu k}\left(\alpha_{32}\mathbf{v}^{(2)} + \beta_{32} k(\mathbf{G}(\mathbf{u}^{(2)}, \mathbf{v}^{(2)}) + \mu \mathbf{v}^{(2)})\right)/A_3,
\end{aligned}
\tag{2.15}
$$

which is a combination of the RK method (2.8) and the exponential RK method (2.7) and we denote it by the RK-ERK method. For these types of equations, we can prove the following result on their accuracy.

Theorem 2.4. *With the set of coefficients (e.g. (2.6)) which leads to a third order method (2.3), the new scheme (2.14) is third-order accurate when applied to the ODE system (2.11). In particular, the RK-ERK scheme (2.15) is also a third-order numerical scheme.*

Proof. The strategy in the proof of Theorem 2.3 becomes too cumbersome for this third order method, and will not be adopted here. Instead we compare the new method (2.14) to (2.7), which is proven to be third-order accurate. For simplicity, we omit the superscript n, and only consider the case when $\mathbf{u}, \mathbf{v}$ are scalars. Since the proof is rather long, we separate it into three steps.

Step 1: One can apply the third order method (2.7) with $\mu = \mu_2$ to both u and v, which leads to the method

$$\begin{aligned}
\tilde{u}^{(1)} =& u^n + \beta_{10} k F(u^n, v^n)/A_{12},\\
\tilde{u}^{(2)} =& (\alpha_{20} u^n + \beta_{20} k (F(u^n, v^n) + \mu_2 u^n))/A_{22}\\
&+ e^{\beta_{10}\mu_2 k}\left(\alpha_{21}\tilde{u}^{(1)} + \beta_{21} k (F(\tilde{u}^{(1)}, \tilde{v}^{(1)}) + \mu_2 \tilde{u}^{(1)})\right)/A_{22},\\
\tilde{u}^{n+1} =& (\alpha_{30} u^n + \beta_{30} k (F(u^n, v^n) + \mu_2 u^n))/A_{32}\\
&+ e^{\beta_{10}\mu_2 k}\left(\alpha_{31}\tilde{u}^{(1)} + \beta_{31} k (F(\tilde{u}^{(1)}, \tilde{v}^{(1)}) + \mu_2 \tilde{u}^{(1)})\right)/A_{32}\\
&+ e^{A\mu_2 k}\left(\alpha_{32}\tilde{u}^{(2)} + \beta_{32} k (F(\tilde{u}^{(2)}, \tilde{v}^{(2)}) + \mu_2 \tilde{u}^{(2)})\right)/A_{32},\\
\tilde{v}^{(1)} =& v^n + \beta_{10} k G(u^n, v^n)/A_{12},\\
\tilde{v}^{(2)} =& (\alpha_{20} v^n + \beta_{20} k (G(u^n, v^n) + \mu_2 v^n))/A_{22}\\
&+ e^{\beta_{10}\mu_2 k}\left(\alpha_{21}\tilde{v}^{(1)} + \beta_{21} k (G(\tilde{u}^{(1)}, \tilde{v}^{(1)}) + \mu_2 \tilde{v}^{(1)})\right)/A_{22},\\
\tilde{v}^{n+1} =& (\alpha_{30} v^n + \beta_{30} k (G(u^n, v^n) + \mu_2 v^n))/A_{32}\\
&+ e^{\beta_{10}\mu_2 k}\left(\alpha_{31}\tilde{v}^{(1)} + \beta_{31} k (G(\tilde{u}^{(1)}, \tilde{v}^{(1)}) + \mu_2 \tilde{v}^{(1)})\right)/A_{32}\\
&+ e^{A\mu_2 k}\left(\alpha_{32}\tilde{v}^{(2)} + \beta_{32} k (G(\tilde{u}^{(2)}, \tilde{v}^{(2)}) + \mu_2 \tilde{v}^{(2)})\right)/A_{32},
\end{aligned} \tag{2.16}$$

where $\tilde{u}$ and $\tilde{v}$ are used to represent the inner stages of (2.7), to differentiate from the targeting method (2.14). With the notation of $a_i = \mu_i k$ $(i = 1, 2)$, we have $O(a_1) = O(a_2) = O(k)$ and

$$\begin{aligned}
A_{1i} &= 1 + \beta_{10} a_i,\\
A_{2i} &= [\alpha_{20} + \beta_{20} a_i] + e^{\beta_{10} a_i}[\alpha_{21} + \beta_{21} a_i] = 1 + (\beta_{20} + \alpha_{21}\beta_{10} + \beta_{21}) a_i + O(a_i^2),\\
A_{3i} &= [\alpha_{30} + \beta_{30} a_i] + e^{\beta_{10} a_i}[\alpha_{31} + \beta_{31} a_i] + e^{A a_i}[\alpha_{32} + \beta_{32} a_i] = 1 + O(a_i).
\end{aligned}$$

Note that $\tilde{v}^{(1)} = v^{(1)}$, however $\tilde{v}^{(2)} \neq v^{(2)}$ due to the coupling of u and v. We take the differences between (2.14) and (2.16) to obtain

$$\begin{aligned}
\tilde{u}^{(1)} - u^{(1)} =& (1/A_{12} - 1/A_{11})\beta_{10} k F = \beta_{10}^2 k (a_1 - a_2)\mu F / A_{11} A_{12} = O(k^2),\\
\tilde{v}^{(2)} - v^{(2)} =& e^{\beta_{10} a_2}\beta_{21} k\left(G(\tilde{u}^{(1)}, \tilde{v}^{(1)}) - G(u^{(1)}, v^{(1)})\right)/A_{22}\\
=& \beta_{21} k\left(G(\tilde{u}^{(1)}, \tilde{v}^{(1)}) - G(u^{(1)}, v^{(1)})\right) + O(k^4) = O(k^3),\\
\tilde{v}^{n+1} - v^{n+1} =& e^{\beta_{10} a_2}\beta_{31} k\left(G(\tilde{u}^{(1)}, \tilde{v}^{(1)}) - G(u^{(1)}, v^{(1)})\right)/A_{32}\\
&+ e^{A a_2}\left[(\alpha_{32} + \beta_{32} a_2)(\tilde{v}^{(2)} - v^{(2)}) + \beta_{32} k\left(G(\tilde{u}^{(2)}, \tilde{v}^{(2)}) - G(u^{(2)}, v^{(2)})\right)\right]/A_{32}\\
=& \beta_{31} k\left(G(\tilde{u}^{(1)}, \tilde{v}^{(1)}) - G(u^{(1)}, v^{(1)})\right) + \alpha_{32}(\tilde{v}^{(2)} - v^{(2)}) + \beta_{32} k(\tilde{u}^{(2)} - u^{(2)}) G_1(u^{(2)}, v^{(2)}) + O(k^4)\\
=& (\beta_{31} + \alpha_{32}\beta_{21}) k(\tilde{u}^{(1)} - u^{(1)}) G_1(u^{(1)}, v^{(1)}) + \beta_{32} k(\tilde{u}^{(2)} - u^{(2)}) G_1(u^{(2)}, v^{(2)}) + O(k^4),
\end{aligned}$$

by repeatedly using the Taylor expansion. Here G_1 denotes the partial derivative of G with respect to the first variable, and we denote $F(u^n, v^n)$ by F for ease of presentation. It is easy to see that

$$u^{(2)} - u^{(1)} = O(k) \quad \text{and} \quad v^{(2)} - v^{(1)} = O(k),$$

from which we conclude that

$$G_1(u^{(2)}, v^{(2)}) = G_1(u^{(1)}, v^{(1)}) + O(k).$$

If the following estimates

$$\tilde{u}^{(2)} - u^{(2)} = O(k^2), \tag{2.17}$$

$$(\beta_{31} + \alpha_{32}\beta_{21})(\tilde{u}^{(1)} - u^{(1)}) + \beta_{32}(\tilde{u}^{(2)} - u^{(2)}) = O(k^3), \tag{2.18}$$

hold (which will be proved in step 2), we have

$$\tilde{v}^{n+1} - v^{n+1} = kG_1(u^{(1)}, v^{(1)})\left[(\beta_{31} + \alpha_{32}\beta_{21})(\tilde{u}^{(1)} - u^{(1)}) + \beta_{32}(\tilde{u}^{(2)} - u^{(2)})\right] + O(k^4) = O(k^4),$$

from which we can conclude that the numerical scheme of v is third-order accurate.

Step 2: Next we provide the proof of (2.17) and (2.18). From (2.16), one can obtain

$$\begin{aligned}\tilde{u}^{(2)} &= (\alpha_{20}u + \beta_{20}kF + \beta_{20}a_2u)/A_{22} + e^{\beta_{10}a}\left(\alpha_{21}\tilde{u}^{(1)} + \beta_{21}kF(\tilde{u}^{(1)}, \tilde{v}^{(1)}) + \beta_{21}a_2\tilde{u}^{(1)}\right)/A_{22}\\ &= u + \beta_{20}kF/A_{22} + e^{\beta_{10}a_2}(\alpha_{21} + \beta_{21}a_2)(\tilde{u}^{(1)} - u)/A_{22} + e^{\beta_{10}a_2}\beta_{21}kF(\tilde{u}^{(1)}, \tilde{v}^{(1)})/A_{22}\\ &= u + \beta_{20}kF/A_{22} + e^{\beta_{10}a_2}(\alpha_{21} + \beta_{21}a_2)\beta_{10}kF/A_{22}A_{12} + e^{\beta_{10}a_2}\beta_{21}kF(\tilde{u}^{(1)}, \tilde{v}^{(1)})/A_{22}.\end{aligned}$$

For $i = 1, 2$, we introduce the notations

$$\begin{aligned}M_i &= (1 + \beta_{10}a_i)[1 - (\beta_{20} + \alpha_{21}\beta_{10} + \beta_{21})a_i](\alpha_{21} + \beta_{21}a_i)\\ N_i &= (1 + \beta_{10}a_i)[1 - (\beta_{20} + \alpha_{21}\beta_{10} + \beta_{21})a_i].\end{aligned}$$

This leads to

$$\begin{aligned}\tilde{u}^{(2)} - u^{(2)} &= (1/A_{22} - 1/A_{21})\beta_{20}kF + (M_2/A_{12} - M_1/A_{11})\beta_{10}kF\\ &\quad + \beta_{21}k(N_2F(\tilde{u}^{(1)}, \tilde{v}^{(1)}) - N_1F(u^{(1)}, v^{(1)})) + O(k^3).\end{aligned}$$

Since $\tilde{v}^{(1)} - v^{(1)} = 0$ and $\tilde{u}^{(1)} - u^{(1)} = O(k^2)$, we have

$$F(\tilde{u}^{(1)}, \tilde{v}^{(1)}) - F(u^{(1)}, v^{(1)}) = O(k^2),$$

and therefore

$$\begin{aligned}\tilde{u}^{(2)} - u^{(2)} &= (\beta_{20} + \alpha_{21}\beta_{10} + \beta_{21})(a_1 - a_2)\beta_{20}kF + (M_2/A_{12} - M_1/A_{11})\beta_{10}kF\\ &\quad + (\alpha_{20}\beta_{10} - \beta_{20} - \beta_{21})(a_2 - a_1)\beta_{21}kF(\tilde{u}^{(1)}, \tilde{v}^{(1)}) + O(k^3).\end{aligned}$$

Using $\alpha_{20} + \alpha_{21} = 1$, we can rewrite the coefficient of $\beta_{10}kF$ as

$$\begin{aligned}\frac{M_2}{A_{12}} - \frac{M_1}{A_{11}} &= [1 + (\alpha_{20}\beta_{10} - \beta_{20} - \beta_{21})a_2](\alpha_{21} + \beta_{21}a_2)/A_{12} + O(a_2^2)\\ &\quad - [1 + (\alpha_{20}\beta_{10} - \beta_{20} - \beta_{21})a_1](\alpha_{21} + \beta_{21}a_1)/A_{11} + O(a_1^2)\\ &= [\alpha_{21} + (\beta_{21} + \alpha_{20}\alpha_{21}\beta_{10} - \alpha_{21}\beta_{20} - \alpha_{21}\beta_{21})a_2 + O(a_2^2)](1 - \beta_{10}a_2 + O(a_2^2)) + O(a_2^2)\\ &\quad - [\alpha_{21} + (\beta_{21} + \alpha_{20}\alpha_{21}\beta_{10} - \alpha_{21}\beta_{20} - \alpha_{21}\beta_{21})a_1 + O(a_1^2)](1 - \beta_{10}a_1 + O(a_1^2)) + O(a_1^2)\\ &= (\alpha_{20}\beta_{21} + \alpha_{20}\alpha_{21}\beta_{10} - \alpha_{21}\beta_{20} - \alpha_{21}\beta_{10})(a_2 - a_1) + O(a_1^2, a_2^2)\\ &= (\alpha_{20}\beta_{21} - \alpha_{21}^2\beta_{10} - \alpha_{21}\beta_{20})(a_2 - a_1) + O(k^2).\end{aligned}$$

Since $F(\tilde{u}^{(1)}, \tilde{v}^{(1)}) - F = O(k)$, it follows that

$$\begin{aligned}\tilde{u}^{(2)} - u^{(2)} &= [-\beta_{20}(\beta_{20} + \alpha_{21}\beta_{10} + \beta_{21}) + \beta_{10}(\alpha_{20}\beta_{21} - \alpha_{21}^2\beta_{10} - \alpha_{21}\beta_{20})\\ &\quad + \beta_{21}(\alpha_{20}\beta_{10} - \beta_{20} - \beta_{21})](a_2 - a_1)kF + O(k^3)\\ &= (2\beta_{10}\beta_{21} - A^2)(a_2 - a_1)kF + O(k^3)\\ &= O((a_2 - a_1)k) + O(k^3) = O(k^2),\end{aligned}$$

which provides the estimate (2.17). In addition,

$$\begin{aligned}&(\beta_{31} + \alpha_{32}\beta_{21})(\tilde{u}^{(1)} - u^{(1)}) + \beta_{32}(\tilde{u}^{(2)} - u^{(2)})\\ &= (\beta_{31} + \alpha_{32}\beta_{21})\beta_{10}^2(a_1 - a_2)kF(1 + O(a)) + \beta_{32}[(2\beta_{10}\beta_{21} - A^2)(a_2 - a_1)kF + O(k^3)]\\ &= \left\{2\beta_{10}\beta_{21}\beta_{32} - [\beta_{10}^2(\beta_{31} + \alpha_{32}\beta_{21}) + \beta_{32}A^2]\right\}(a_2 - a_1)kF + O(k^3)\\ &= O(k^3),\end{aligned}$$

where the last equality follows from the fact that $\beta_{10}\beta_{21}\beta_{32} = 1/6$ and $\beta_{10}^2(\beta_{31} + \alpha_{32}\beta_{21}) + \beta_{32}A^2 = 1/3$, as the condition to ensure the third order accuracy shown in [11, Eq. (2.19)]. This finishes the proof of the estimates (2.17) and (2.18).

Step 3: The error estimate of $\tilde{u}^{n+1} - u^{n+1}$ can be done symmetrically following the same approach. To save space, the detailed analysis is ignored. □

Remark 2.1. In this subsection, we consider the combination of two exponential RK methods with any nonnegative μ_1 and μ_2 for equations with different stiffness. We can also extend them to a more general case of N exponential RK methods with a set of μ_i, $i = 1, 2, \cdots, N$.

Remark 2.2. The RK-ERK method (2.15) will be studied in the next section. The constant μ in this method is determined dynamically by applying the conditions (2.5) in Theorem 2.2 on the equation containing stiff term, which is a sufficient condition to guarantee the sign-preserving property.

3. Positivity-preserving well-balanced DG spatial discretization

In this section, we start with a quick review of the high-order positivity-preserving well-balanced semi-discrete DG scheme in [34] for the one-dimensional SWEs (1.2) without a friction term. The RK-ERK method will be applied to the resulting equations, leading to a fully discrete method that is well-balanced, positivity-preserving and sign-preserving simultaneously.

3.1. Notations and conventional DG methods

We discretize the one-dimensional computational domain I into cells $I_j = \left[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}\right]$, and denote by Δx_j the size of the j-th cell and by $\Delta x = \max_j \Delta x_j$ the maximum mesh size. For simplicity, we rewrite (1.2) as

$$U_t + f(U)_x = s(U, b),$$

where $U = (h, q)^T$, $f(U)$ is the flux and $s(U, b)$ is the source term. In a high order DG method, we seek an approximation solution, still denoted by U for the ease of notation, which belongs to the finite dimensional space

$$V_{\Delta x} = V_{\Delta x}^d = \left\{ w : \text{each component of } w|_{I_j} \in P^d(I_j), \;\; j = 1, ..., J \right\}, \tag{3.1}$$

where $P^d(I_j)$ is the space of polynomials in I_j of degree at most d and J is the total number of computational cells. We project the bottom function b onto the same space $V_{\Delta x}$, to obtain an approximation which is still denoted by b.

The conventional DG method in each cell I_j can be formulated as follows: find $U \in V_{\Delta x}$, such that

$$\int_{I_j} \partial_t U v dx - \int_{I_j} f(U) \partial_x v dx + \hat{f}_{j+\frac{1}{2}} v(x^-_{j+\frac{1}{2}}) - \hat{f}_{j-\frac{1}{2}} v(x^+_{j-\frac{1}{2}}) = \int_{I_j} s(U, b) v dx, \tag{3.2}$$

where v is a test function in $V_{\Delta x}$ and

$$\hat{f}_{j+\frac{1}{2}} = F\left(U(x^-_{j+\frac{1}{2}}, t), U(x^+_{j+\frac{1}{2}}, t) \right), \tag{3.3}$$

with $F(a_1, a_2)$ being a numerical flux. In our numerical examples we will use the Lax-Friedrichs flux

$$F(a_1, a_2) = \frac{1}{2}(f(a_1) + f(a_2) - \alpha(a_2 - a_1)), \tag{3.4}$$

with $\alpha = \max(|u| + \sqrt{gh})$ and the maximum is taken over the whole computational domain.

3.2. Well-balanced DG methods

In order to preserve the still water stationary solution (1.3) exactly, the modified well-balanced scheme [30] has the form

$$\begin{aligned} &\int_{I_j} \partial_t U v dx - \int_{I_j} f(U) \partial_x v dx + \hat{f}_{j+\frac{1}{2}} v(x^-_{j+\frac{1}{2}}) - \hat{f}_{j-\frac{1}{2}} v(x^+_{j-\frac{1}{2}}) \\ &\qquad = \int_{I_j} s(U, b) v dx + (\hat{f}_{j+\frac{1}{2}} - \hat{f}^l_{j+\frac{1}{2}}) v(x^-_{j+\frac{1}{2}}) - (\hat{f}_{j-\frac{1}{2}} - \hat{f}^r_{j-\frac{1}{2}}) v(x^+_{j-\frac{1}{2}}). \end{aligned} \tag{3.5}$$

The left and right fluxes are given by

$$\begin{aligned} \hat{f}^l_{j+\frac{1}{2}} &= F\left(U^{*,-}_{j+\frac{1}{2}}, U^{*,+}_{j+\frac{1}{2}} \right) + \begin{pmatrix} 0 \\ \frac{g}{2}(h^-_{j+\frac{1}{2}})^2 - \frac{g}{2}(h^{*,-}_{j+\frac{1}{2}})^2 \end{pmatrix}, \\ \hat{f}^r_{j-\frac{1}{2}} &= F\left(U^{*,-}_{j-\frac{1}{2}}, U^{*,+}_{j-\frac{1}{2}} \right) + \begin{pmatrix} 0 \\ \frac{g}{2}(h^+_{j-\frac{1}{2}})^2 - \frac{g}{2}(h^{*,+}_{j-\frac{1}{2}})^2 \end{pmatrix}, \end{aligned} \tag{3.6}$$

respectively, where the left and right values of U are redefined as

$$U^{*,\pm}_{j+\frac{1}{2}} = \begin{pmatrix} h^{*,\pm}_{j+\frac{1}{2}} \\ h^{*,\pm}_{j+\frac{1}{2}} u^{\pm}_{j+\frac{1}{2}} \end{pmatrix}, \quad h^{*,\pm}_{j+\frac{1}{2}} = \max\left(0, h^{\pm}_{j+\frac{1}{2}} + b^{\pm}_{j+\frac{1}{2}} - \max(b^{+}_{j+\frac{1}{2}}, b^{-}_{j+\frac{1}{2}})\right), \tag{3.7}$$

following the hydrostatic reconstruction idea that was first introduced in [1]. Note that if the piecewise polynomial approximation of b is continuous, we have $U^{*,\pm}_{j+\frac{1}{2}} = U^{\pm}_{j+\frac{1}{2}}$, and the well-balanced method (3.5) coincides with the conventional DG scheme (3.2). We also point out here that $\hat{f}_{j+\frac{1}{2}} - \hat{f}^{l}_{j+\frac{1}{2}}$ and $\hat{f}_{j-\frac{1}{2}} - \hat{f}^{r}_{j-\frac{1}{2}}$ are high order correction terms at the level of $O(\Delta x^{k+1})$ regardless of the smoothness of the solution U.

Another important ingredient in DG methods is the slope limiter which might be needed if the solution contains discontinuities. We use the total variation bounded (TVB) limiter, with a corrected minmod function defined by

$$\overline{m}(a_1, ..., a_n) = \begin{cases} a_1, & |a_1| \leqslant M\Delta x^2, \\ m(a_1, ..., a_n), & \text{otherwise}, \end{cases} \tag{3.8}$$

where M is the TVB parameter to be chosen adequately [10] and the minmod function m is given by

$$m(a_1, ..., a_n) = \begin{cases} s \min |a_i|, & s = \text{sgn}(a_1) = \cdots = \text{sgn}(a_n), \\ 0, & \text{otherwise}. \end{cases}$$

This limiter procedure might destroy the preservation of the steady state $h + b = C$. Therefore we apply the limiter procedure on the function $(h + b, q)^T$ instead. The modified solution is then defined by $h^{mod} = (h + b)^{mod} - b$. Since the average of h in cell I_j, denoted by $\overline{h_j}$, satisfies $\overline{h^{mod}_j} = \overline{(h+b)^{mod}_j} - \overline{b_j} = \overline{(h+b)_j} - \overline{b_j} = \overline{h_j}$, this limiting procedure will not destroy the conservativity of the water height h.

3.3. Positivity-preserving limiter

The positivity-preserving limiter in [34] to ensure the non-negativity of the water height h in the numerical simulation will be discussed in this subsection. Before presenting the main result, we first introduce the N-point Gauss-Lobatto quadrature rule on the interval $I_j = \left[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}\right]$, which is exact for the integral of polynomials of degree up to $2N - 3$, with N chosen such that $2N - 3 \geqslant k$. We denote these quadrature points on I_j as

$$S_j = \left\{x_{j-\frac{1}{2}} = x^1_j, x^2_j, ..., x^{N-1}_j, x^N_j = x_{j+\frac{1}{2}}\right\}.$$

Let w_r be the corresponding quadrature weights on the interval $[-1/2, 1/2]$ such that $\sum_{r=1}^{N} w_r = 1$. Next, let us consider the update of cell averages of h in the well-balanced DG methods with a forward Euler time discretization, given by

$$\overline{h}^{n+1}_j = \overline{h}^{n}_j + \frac{\Delta t}{\Delta x}\left[\hat{F}\left(h^{*,-}_{j-\frac{1}{2}}, u^{-}_{j-\frac{1}{2}}; h^{*,+}_{j-\frac{1}{2}}, u^{+}_{j-\frac{1}{2}}\right) - \hat{F}\left(h^{*,-}_{j+\frac{1}{2}}, u^{-}_{j+\frac{1}{2}}; h^{*,+}_{j+\frac{1}{2}}, u^{+}_{j+\frac{1}{2}}\right)\right], \tag{3.9}$$

where

$$\hat{F}\left(h^{*,-}_{j+\frac{1}{2}}, u^{-}_{j+\frac{1}{2}}; h^{*,+}_{j+\frac{1}{2}}, u^{+}_{j+\frac{1}{2}}\right) = \frac{1}{2}\left[h^{*,-}_{j+\frac{1}{2}} u^{-}_{j+\frac{1}{2}} + h^{*,+}_{j+\frac{1}{2}} u^{+}_{j+\frac{1}{2}} - \alpha\left(h^{*,+}_{j+\frac{1}{2}} - h^{*,-}_{j+\frac{1}{2}}\right)\right]. \tag{3.10}$$

Now we are ready to state the main result in [34].

Proposition 3.1. *Consider the scheme (3.9) satisfied by the cell averages of the water height. Let $h^n_j(x)$ be the DG polynomial for the water height in the cell I_j. If $h^{-}_{j-\frac{1}{2}}$, $h^{+}_{j+\frac{1}{2}}$ and $h^n_j(x^r_j)$ $(r = 1, ..., N)$ are all nonnegative, then $\overline{h}^{n+1}_j$ is also nonnegative under the CFL condition $\alpha\frac{\Delta t}{\Delta x} \leqslant w_1$.*

The proposition gives us an image of how the time step is chosen. Following the proposition, the constant k_E in the assumption (2.4) can be taken as $w_1\Delta x/\alpha$. To enforce the conditions of the proposition, we need to modify $h^n_j(x)$ such that it is non-negative at all $x \in S_j$. Given $\overline{h}^{n}_j \geqslant 0$, we introduce the following limiter on the DG polynomial $U^n_j(x) = \left(h^n_j(x), q^n_j(x)\right)^T$, which is a linear scaling around its cell average:

$$\widetilde{U}^n_j(x) = \theta(U^n_j(x) - \overline{U}^n_j) + \overline{U}^n_j, \quad \theta = \min\left(1, \frac{\overline{h}^n_j}{\overline{h}^n_j - m_j}\right), \tag{3.11}$$

with

$$m_j = \min_{r=1\ldots,N} h_j^n(x_j^r). \tag{3.12}$$

It is easy to observe $\widetilde{h}_j^n(x_j^r) \geqslant 0$ for $r = 1, \ldots, N$. Then, we use the modified polynomial $\widetilde{U}_j^n(x)$ instead of $U_j^n(x)$ in the scheme (3.5). It follows by the proposition that $\overline{h}_j^{n+1}$ is also non-negative and therefore (3.11) is indeed a positivity-preserving limiter.

Note that this modification does not change the averages of $U_j^n(x)$, namely $\overline{\widetilde{U}_j^n(x)} = \overline{U}_j^n$. Also, this limiter does not destroy the high order accuracy, and we refer to [35] for the detailed proof. Special attention should be paid in practical implementation when the water height is close to zero. In these nearly dry regions, a small numerical error in q can induce large values of the velocity $u = q/h$, and in turn leads to very small time steps. There have been many attempts to address this challenge, which is beyond the focus of this paper. Since the velocity in these nearly dry regions should be at the same magnitude as the maximum of the velocity in wet regions, we simply set $q = 0$ if $h \leqslant 10^{-6}$ in the numerical tests of this paper.

3.4. Applications of the new sign-preserving time integration

We rewrite the well-balanced DG scheme (3.5) for SWEs as

$$\int_{I_j} \partial_t U v dx = \int_{I_j} f(U)\partial_x v dx - \hat{f}^l_{j+\frac{1}{2}} v(x^-_{j+\frac{1}{2}}) + \hat{f}^r_{j-\frac{1}{2}} v(x^+_{j-\frac{1}{2}}) + \int_{I_j} s(U,b) v dx. \tag{3.13}$$

Choosing the test function v as the basis function of $V_{\Delta x}$ and also representing U as a linear combination of these basis lead to an ODE system. Following the notations in (2.1), we have

$$L(U) = \int_{I_j} f(U)\partial_x v dx - \hat{f}^l_{j+\frac{1}{2}} v(x^-_{j+\frac{1}{2}}) + \hat{f}^r_{j-\frac{1}{2}} v(x^+_{j-\frac{1}{2}}) + \int_{I_j} s_{\mathrm{b}}(U,b) v dx,$$

in which $f(U) = \left(q,\ \frac{q^2}{h} + \frac{1}{2}gh^2\right)^T$, $s_{\mathrm{b}}(U,b) = (0,\ -ghb_x)^T$, and

$$s(U) = \int_{I_j} s_{\mathrm{st}}(U) v dx,$$

with $s_{\mathrm{st}}(U) = \left(0,\ -gn^2\frac{|q|q}{h^{7/3}}\right)^T$ representing the stiff term. As the first component of s_{st} is zero, we can apply the standard RK method (2.8) to discretize the first equation in the SWEs. The second component of s_{st} could be stiff when h is small, therefore we apply exponential RK method (2.7) to the second equation. In other words, the RK-ERK method (2.15) or (2.13) is chosen as the temporal discretization. According to Theorem 2.2, the parameter μ is dynamically computed by $\mu = gn^2 \max(|q|/h^{7/3})$.

For the proposed RK-ERK DG methods, we have following properties.

Proposition 3.2. *The fully discrete scheme obtained by applying RK-ERK method* (2.15) *to the semi-discrete DG method* (3.13) *preserves the still water steady state solution* (1.3).

The proof of this proposition is straightforward. The well-balanced property of the semi-discrete DG method is analyzed in [34]. Since the proposed RK-ERK temporal discretization is also steady-state preserving (similar to Proposition 2.1), the well-balanced property of the fully discrete method can be easily observed.

Proposition 3.3. *With the usage of positivity-preserving limiter* (3.11)*, the fully discrete scheme obtained by applying RK-ERK method* (2.15) *to the semi-discrete DG method* (3.13) *with the choices*

$$\mu = gn^2 \max(|q|/h^{7/3}), \quad \Delta t \leqslant \zeta k_E = \zeta \omega_1 \Delta x/\alpha, \tag{3.14}$$

preserves the positivity of the water height h, and is also sign-preserving with respect to the momentum q. Note that the set of coefficients (2.6) *yields the optimal* $\zeta = 0.7929797388491311$.

Since the RK method is applied on the equation of water height h, the positivity-preserving feature to preserve the non-negativity of h of the semi-discrete DG method, with the aid of positivity-preserving limiter (3.11), is not affected by the RK-ERK temporal discretization. For the sign-preserving property, we follow the guideline in Theorem 2.2 to choose μ and Δt as above.

Remark 3.1. In Proposition 3.3, the choices of μ and Δt are sufficient but not necessary to preserve the positivity of water height and the sign of momentum. In practice, at each time level we can take a standard CFL condition of DG method. If one observes that the positivity preserving or the sign preserving properties are violated at the next time step, we will halve the value of Δt and restart this computation. Numerical observation show that this could lead to a saving in the computational time.

The well-balanced DG method can be simply extended to two dimensions. We divide the computational domain into cells $I_{i,j} = \left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right] \times \left[y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}\right]$. Let $\hat{f}^l_{i+\frac{1}{2}}$, $\hat{f}^r_{i-\frac{1}{2}}$, $\hat{g}^l_{j+\frac{1}{2}}$ and $\hat{g}^r_{j-\frac{1}{2}}$ be the well-balanced fluxes defined similarly as in (3.6); see also [33,34]. The spatial discretization is given by

$$\begin{aligned}\int_{I_{i,j}} \partial_t U v dx = &\int_{I_{i,j}} f(U)\partial_x v dx dy - \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \hat{f}^l_{i+\frac{1}{2}} v(x^-_{i+\frac{1}{2}}, y)dy + \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \hat{f}^r_{i-\frac{1}{2}} v(x^+_{i-\frac{1}{2}}, y)dy \\ &+ \int_{I_{i,j}} g(U)\partial_y v dx dy - \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \hat{g}^l_{j+\frac{1}{2}} v(x, y^-_{j+\frac{1}{2}})dx + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \hat{g}^r_{j-\frac{1}{2}} v(x, y^+_{j-\frac{1}{2}})dx \\ &+ \int_{I_{i,j}} s(U,b) v dx dy.\end{aligned}$$

in which $U = (h, q, p)^T$, $f(U) = \left(q,\ \frac{q^2}{h} + \frac{1}{2}gh^2,\ \frac{qp}{h}\right)^T$ and $g(U) = \left(p,\ \frac{qp}{h},\ \frac{p^2}{h} + \frac{1}{2}gh^2\right)^T$. As in one-dimensional case, we split $s(U,b)$ into non-stiff and stiff parts:

$$s_b(U,b) = \left(0,\ -ghb_x,\ -ghb_y\right)^T \quad \text{and} \quad s_{st}(U) = \left(0,\ -gn^2\frac{q\sqrt{q^2+p^2}}{h^{7/3}},\ -gn^2\frac{p\sqrt{q^2+p^2}}{h^{7/3}}\right)^T.$$

Next we apply standard RK method (2.8) to the discretization of the first equation and exponential RK method (2.7) to the second and third equations. The parameter μ is dynamically computed by $\mu = gn^2\max(\sqrt{q^2+p^2}/h^{7/3})$. The steady state solution in two-dimensional case takes the form

$$h + b = C,\ q = p = 0.$$

Following the same analysis, we conclude that the propositions above also hold in two-dimensional case, and the proposed fully discrete method satisfies the well-balanced, positivity-preserving and sign-preserving properties simultaneously.

4. Numerical examples

In this section, we will apply our new time integration method with the family of coefficients (2.6) to an ODE system and the SWEs with friction terms. Several numerical examples will be tested to illustrate the performance of our methods. DG method with $d = 2$, coupled with the third order temporal discretization, is tested. If the limiter is need, the TVD minmod slope limiter will be used. The parameter μ is dynamically chosen at each time step following the condition (3.14), unless in the accuracy test where a constant μ is set to study the accuracy of the proposed method. We take the CFL condition $\alpha\frac{\Delta t}{\Delta x} = 0.18$ unless otherwise stated, and the gravitational acceleration g is 9.80665 m/s^2.

4.1. Accuracy test on an ODE system

We first test the accuracy of the new ODE solver (2.14). Consider an ODE system

$$\begin{cases} u'(t) = \dfrac{u - cv^7}{2\sqrt{v}} - \dfrac{1}{2}, \\ v'(t) = u - \sqrt{v} - cv^7, \end{cases}$$

with $u(0) = v(0) = 1$, and its exact solutions are given by

$$u(t) = (6ct+1)^{-\frac{1}{12}}, \quad v(t) = (6ct+1)^{-\frac{1}{6}}.$$

Choose $c = 100$ and final time $T = 0.1$. In this case we take $\mu \equiv c = 100$. The l^1-error is defined as

Table 1
Accuracy test of the new 2nd-order method for the test in Section 4.1.

N	40	80	160	320	640
l^1 error	2.78e-03	7.83e-04	2.11e-04	5.45e-05	1.38e-05
Order	/	1.83	1.89	1.95	1.98

Table 2
Accuracy test of the new 3rd-order method for the test in Section 4.1.

N	40	80	160	320	640
l^1 error	2.01e-04	3.97e-05	5.20e-06	6.55e-07	8.15e-08
Order	/	2.34	2.93	2.99	3.01

Table 3
Accuracy test of the new 3rd-order method on the shallow water equations for the test in Section 4.2.

N	h		q	
	L^2 error	Order	L^2 error	Order
100	5.92e-05		1.68e-04	
200	6.27e-06	3.24	3.67e-05	2.19
400	7.42e-07	3.08	4.85e-06	2.92
800	9.32e-08	2.99	6.33e-07	2.94
1600	1.18e-08	2.99	8.10e-08	2.97

$$e_1 := |u(T) - u^N| + |v(T) - v^N|.$$

The errors and orders of the temporal discretization method (2.12) with various time step sizes are shown in Table 1, with second order convergence rate confirmed. We also tested the third order method (2.14) and reported the numerical results in Table 2, from which the third order accuracy can be observed. This validates the second and third order convergence rate of the proposed methods.

4.2. Accuracy test on the shallow water equations

In this example, we apply the temporal discretization (2.14) to the shallow water equations, coupled with the DG spatial discretization, and verify the convergence rate of the resulting algorithm. We consider the "manufactured" exact solution taking the form of $h(x,t) = q(x,t) = 2 + \sin(0.04\pi(x-t))$, which satisfies the modified shallow water equations

$$\begin{cases} h_t + q_x = 0, \\ q_t + \left(\dfrac{q^2}{h} + \dfrac{1}{2}gh^2\right)_x = -g\dfrac{|q|q}{h^{7/3}} + g[2 + \sin(0.04\pi(x-t))]^{-1/3} \\ \qquad\qquad\qquad\qquad +0.04\pi gh \cdot \cos(0.04\pi(x-t)) \end{cases} \tag{4.1}$$

with an additional source term on the right hand side. The computational domain is $[0, 100]$ and the final time is taken to be $T = 0.04$.

We apply DG spatial discretization and our new third-order time integration to this model (4.1). We simply fix μ to be 10 in the computation. The L^2 errors and numerical orders of our methods with various time and step sizes are shown in Table 3, from which we can observe the third order convergence easily. This confirms the high order accuracy of the proposed algorithm.

4.3. Test for sign-preserving property

In the simulation of the SWEs, the water height h may be very small or even zero near the wetting and drying front, in which case the friction term in (1.2) becomes stiff, therefore the use of a sign-preserving discretization would be useful. The sign-preserving property of the scheme (2.7) is also reflected in the evolution of the discharge q. In this section, we consider an example studied in [5], where the necessity of sign-preserving discretization is explored.

Following the setup in [5], we consider the system (1.2) with $b_x = -0.2$, $n = 0.09$ and the initial conditions

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

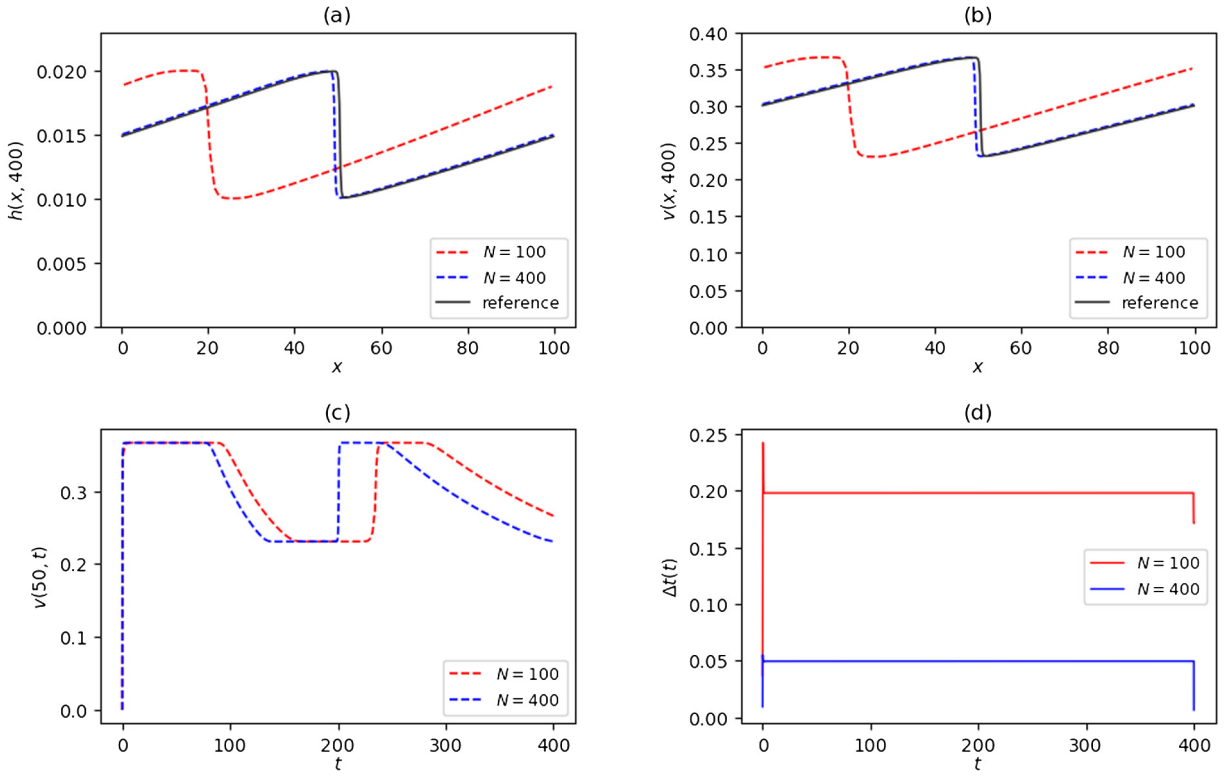

Fig. 1. Numerical solutions obtained using exponential RK methods for both equations with $N = 100$ (coarse) and $N = 400$ (fine) meshes, for the test in Section 4.3. Top row: the water height and velocity at time $T = 400$; Bottom left: the time history of velocity at $x = 50$; Bottom right: the time history of the time step size.

The computational domain is set as $[0, 100]$, which is divided into N uniform cells. The minmod slope limiter (3.8) is applied to h and q at each time step.

For this example, if the traditional explicit RK methods are used for both equations, a tiny time step size is needed as the second equation is stiff. In order to use larger time step, the implicit-explicit (IMEX) method could be utilized and has been studied in [5]. It was observed from [5, Fig. 8] that the velocity could turn into negative and the numerical result contains very large error when the coarse mesh is used. The sign-preserving property would be useful in order to produce satisfying results even on coarse mesh.

We first try to apply the sign-preserving exponential RK method (2.7) with the same μ to both equations, and the numerical results are shown in Fig. 1. We also provide the solution computed using the new method (2.14) on fine meshes ($N = 400$) as a reference solution for comparison. We can observe that the velocity does stay non-negative for all time, however the results show substantial disagreements between the coarse- and fine-grid solutions. One can notice a large phase error in h and v. Such a delay in shock propagation results from the large numerical error when we apply (2.7) to the equation of h. In other words, the parameter μ in the scheme induces a large error in the non-stiff equation. As we refine the meshes, it can be seen that the shock location converges to the correct position.

We then apply the proposed coupled RK-ERK method (2.15) (i.e., Equations (2.8) and (2.7)) to h and q respectively for time discretization and plot the numerical solutions computed at final time $T = 400$ in Fig. 2(a) and 2(b). As one can see, the results obtained on coarse ($N = 100$) and fine ($N = 400$) meshes match well. Fig. 2(c) gives the value of the velocity v at $x = 50$ as a function of time. We observe that both the velocity and the speed of the shock are captured quite accurately even with $N = 100$. Fig. 2(d) depicts the time history of the time step size Δt; Fig. 2(e) and 2(f) illustrate how the parameter μ changes with respect to the time. The rapid changes of time step sizes and the parameter μ at the very beginning of the simulation are due to the stiff friction term. As the simulation progresses, the system becomes non-stiff.

4.4. Test for well-balanced property

Several examples related to the well-balanced property will be tested in this section.

4.4.1. Steady state over a non-flat bottom containing a wet/dry interface

First, we consider the case of an initial condition being the steady state solution over a non-flat bottom containing a wet/dry interface. The Manning coefficient n is taken to be 0.09. The bottom topography is given by

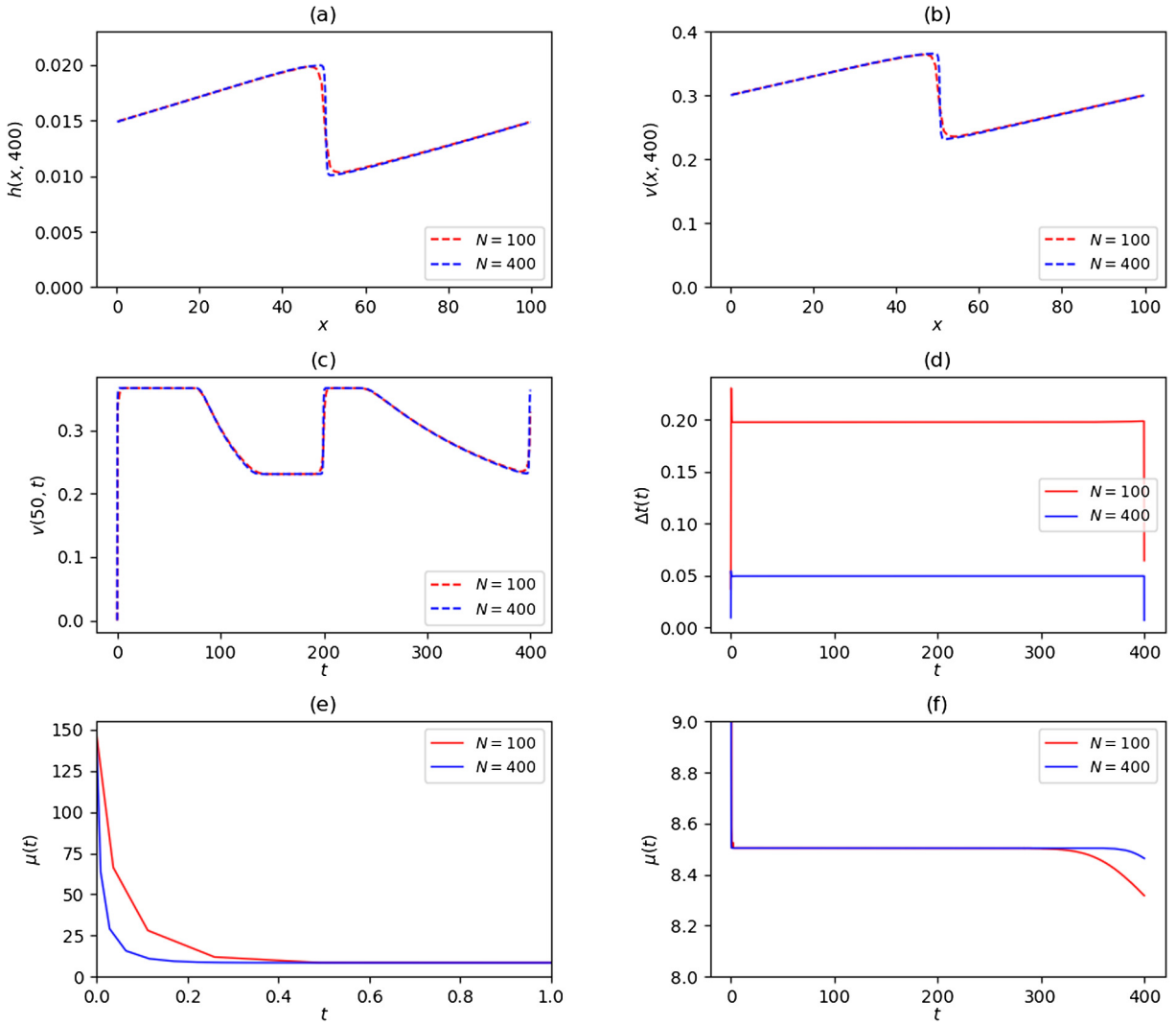

Fig. 2. Numerical solutions obtained using the proposed RK-ERK temporal discretization with $N = 100$ (coarse) and $N = 400$ (fine) meshes, for the test in Section 4.3. Top row: the water height and velocity at time $T = 400$; Middle left: the time history of velocity at $x = 50$; Middle right: the time history of the time step size; Bottom left: the time history of μ until $T = 1$; Bottom right: the time history of μ during the whole simulation.

$$b(x) = \max(0, 0.25 - 5(x - 0.5)^2), \quad 0 \leqslant x \leqslant 1. \tag{4.2}$$

The initial data are

$$h + b = \max(0.2, b), \quad q = 0,$$

which contains both wet and dry regions. We divide the computational domain $[0, 1]$ into $N = 200$ uniform cells and impose the periodic boundary conditions. The water stays still as long as initially

$$h = 0 \quad \text{or} \quad h + b = 0.2. \tag{4.3}$$

In practical implementation, we need to make sure the condition (4.3) is precisely satisfied up to round-off error when we start the computation. We denote by $h(x, t)$ and $q(x, t)$ the numerical solutions. We compute until $T = 0.5$ and compare the numerical solutions with $h(x, 0)$ and $q(x, 0) = 0$. We focus on the errors $\|h(x, T) - h_0(x)\|$ and $\|q(x, T)\|$, which are given in Table 4. The errors are at the level of round-up errors, which verifies the well-balanced property. The computed water level and discharge are shown in Fig. 3.

4.4.2. Steady state solution in part of the domain

This example explains why we need steady state preserving temporal discretization. The Manning coefficient n is still taken to be 0.09 as before. We modify (4.2) slightly and make two copies of the humps. The bottom topography is given by

Table 4
L^1 and L^∞ errors of steady state solutions for the test in Section 4.4.1.

N	L^1 error		L^∞ error	
	h	q	h	q
100	1.83e-16	5.81e-16	8.33e-16	2.26e-15
200	1.67e-18	7.15e-17	5.55e-17	7.78e-16

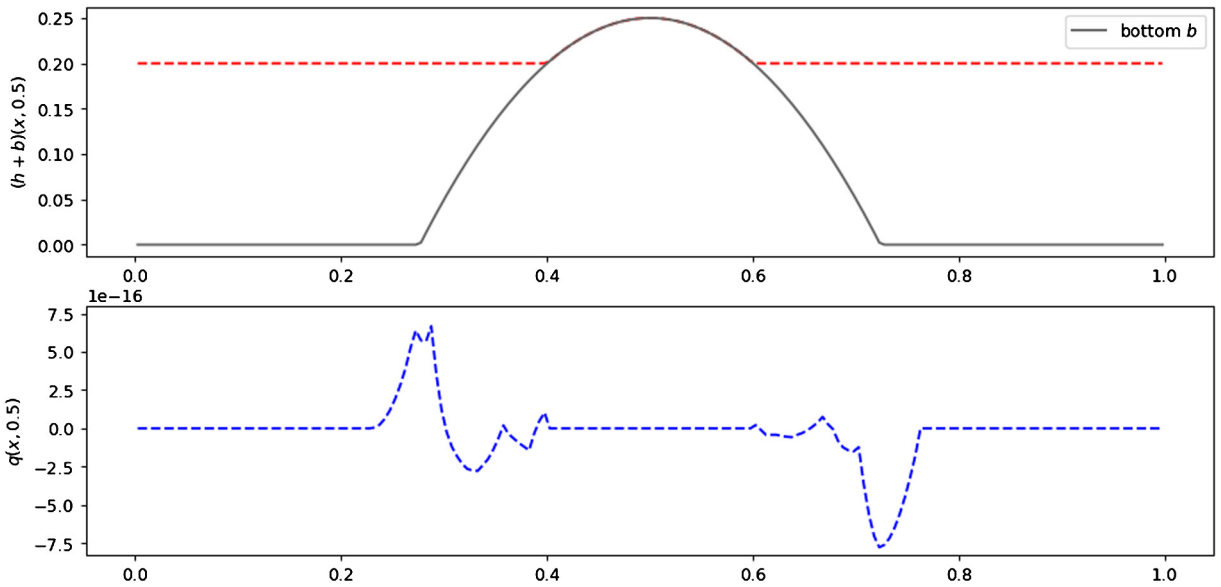

Fig. 3. Water level $h+b$ and discharge q of the steady state ($N=200$) for the test in Section 4.4.1.

$$b(x)=\begin{cases}\max(0,0.25-20(x-0.25)^2), & 0\leqslant x\leqslant 0.5,\\ \max(0,0.25-20(x-0.75)^2), & 0.5<x\leqslant 1,\end{cases}$$

and the initial data are

$$h(x,0)+b(x)=\begin{cases}1, & 0\leqslant x\leqslant 0.5\\ 0.5, & 0.5<x\leqslant 1\end{cases},\qquad q(x,0)=0.$$

We divide the computational domain into 200 uniform cells and compute until $T=0.01$. The initial wave starts to propagate from the middle. At this stopping time, the solution consists of a partial steady state, as the water surface near the boundaries remains still. We expect our method to be able to maintain the steady state near the boundaries. The numerical solutions are shown in Fig. 4. We compare the solutions at $T=0.01$ and initial data, and see that the differences are at the level of round-up errors near the left and right boundaries of computational domain (Fig. 5).

4.4.3. Small perturbation test

In this example we are studying a nearly equilibrium problem by imposing a small perturbation to the steady state problem. The system is solved over a non-flat trigonometric bottom

$$b(x)=\begin{cases}0.25\cos(10\pi(x-1.5))+1, & 1.4\leqslant x\leqslant 1.6,\\ 0, & \text{otherwise},\end{cases}$$

in the computational domain $[0,2]$. The initial conditions are given by

$$h(x,0)+b(x)=\begin{cases}1.001, & 1.1\leqslant x\leqslant 1.2\\ 1, & \text{otherwise}\end{cases},\qquad q(x,0)=0.$$

We divide the domain into 400 cells and compute until $T=0.2$. The Manning coefficient $n=1$ is considered. Although there is a friction term, this problem is not stiff since the absolute value of the discharge q stays small. Thus our new time integration method almost reduces to the traditional RK method (2.8). The numerical solutions are plotted in Fig. 6.

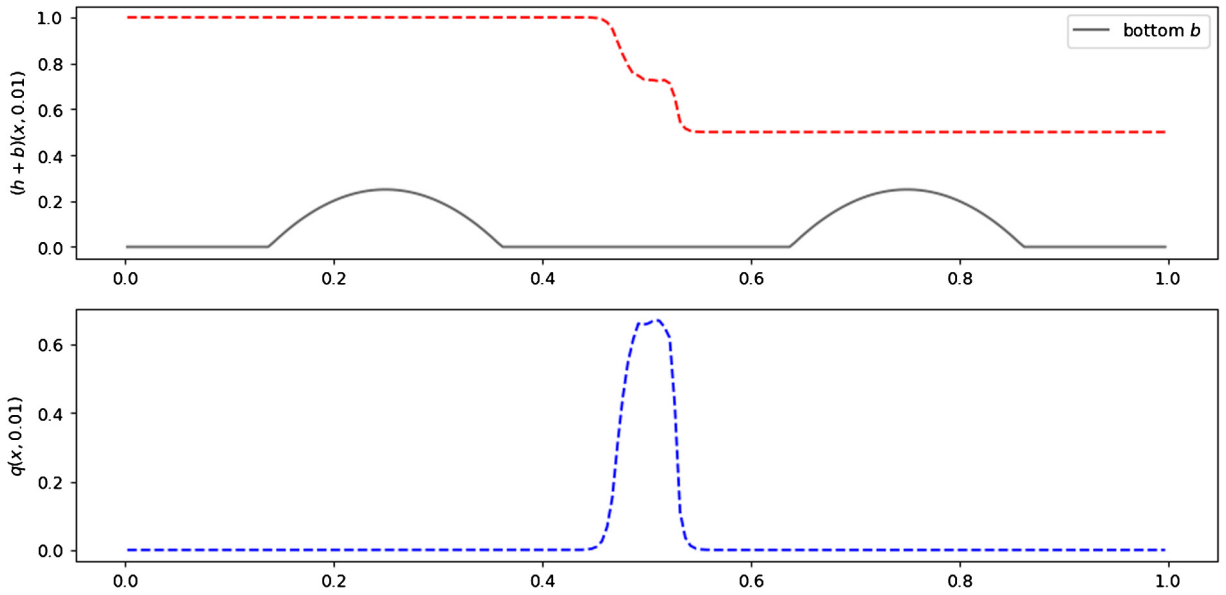

Fig. 4. Water surface and discharge of the Riemann problem in Section 4.4.2.

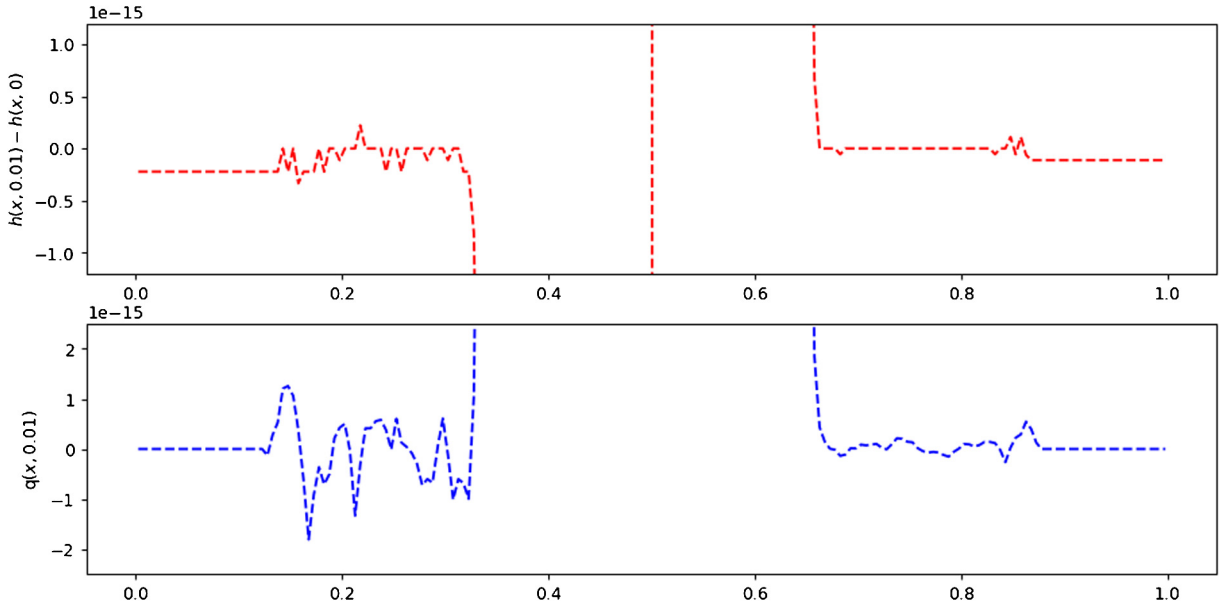

Fig. 5. Differences between the solutions at $T = 0.01$ and initial data for the test in Section 4.4.2.

4.5. Riemann problem over a flat bottom

In this subsection we consider a Riemann problem containing dry area over a flat bottom without friction terms, i.e. $b(x) = 0$ and $n = 0$. This example aims to demonstrate the positivity-preserving ability of the proposed method.

The computational domain is set to be $[0, 600]$, and the initial conditions are given by

$$h(x,0) = \begin{cases} 10, & x \leqslant 300 \\ 0, & \text{otherwise} \end{cases}, \qquad q(x,0) = 0.$$

We can see that the right half region is dry. The analytic solutions of this type of problem are given in [2]. We compute this problem using our well-balanced positivity-preserving method with simple transmissive boundary conditions. As in Section 4.4.3, our new time integration method reduces to the traditional RK method (2.8). The domain is divided into

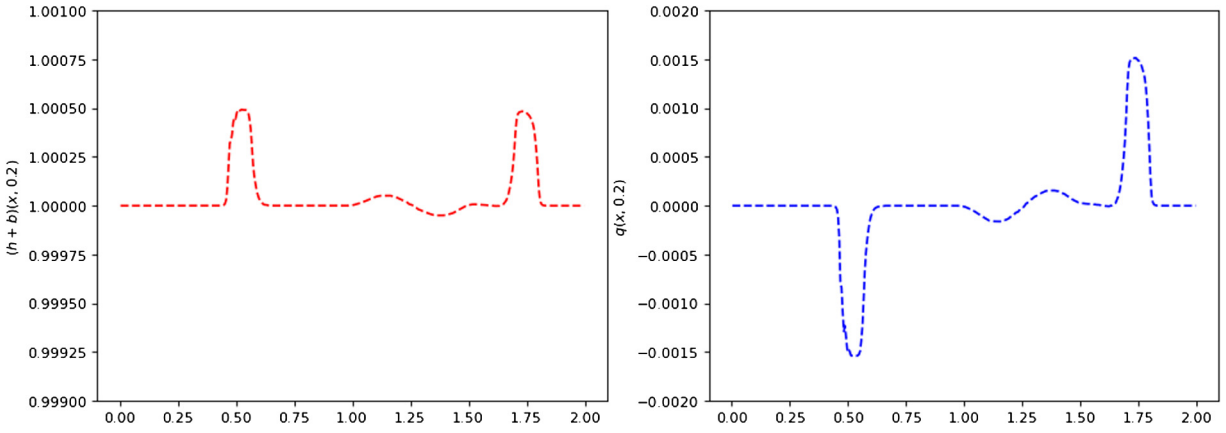

Fig. 6. Water level and discharge in the small perturbation test for the test in Section 4.4.3.

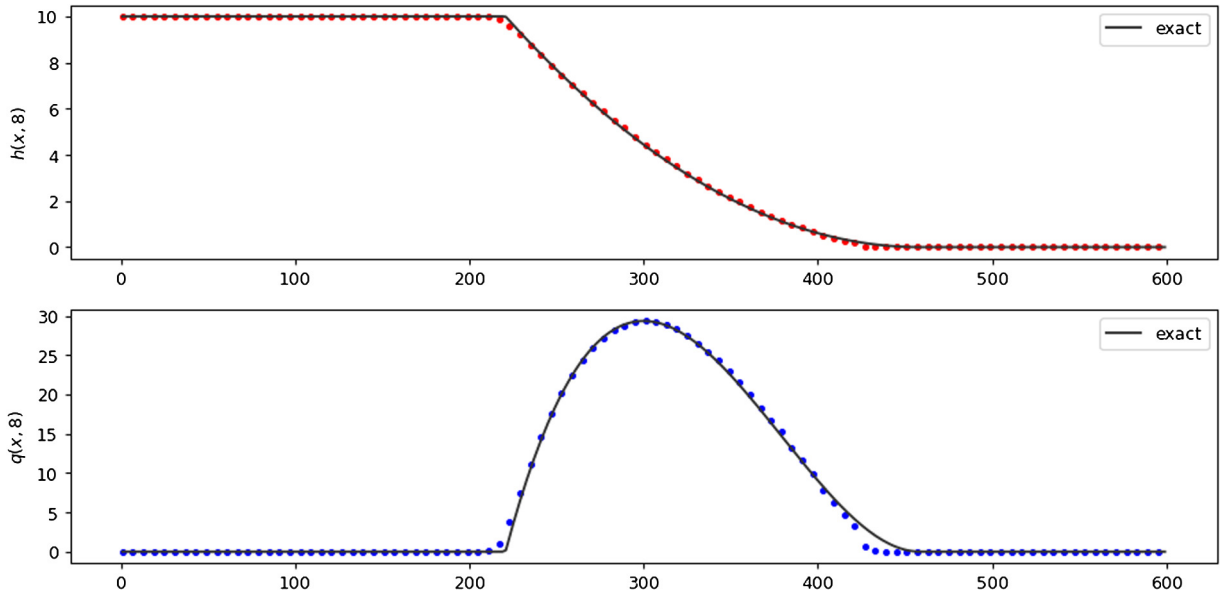

Fig. 7. Numerical and exact solutions of Riemann problem at $T = 8$ for the test in Section 4.5.

300 uniform cells and the final time is taken to be $T = 8$. We plot the numerical solutions and provide exact solutions for comparison in Fig. 7.

4.6. Two-dimensional test for sign-preserving property

We consider the two-dimensional system (1.1) with $b_x = b_y = -0.1414$, $n = 0.09$ and the initial conditions analogous to those in one-dimensional example from Section 4.3:

$$h(x, y, 0) = \begin{cases} 0.02, & 30 \leqslant x, y \leqslant 70, \\ 0.01, & \text{otherwise}, \end{cases} \quad q(x, y, 0) = \begin{cases} 0, & 30 \leqslant x, y \leqslant 70, \\ 0.02828, & \text{otherwise}, \end{cases} \quad p(x, y, 0) = 0.$$

We apply both exponential RK time integration and new RK-ERK time integration, and run the simulation until $T = 300$. Two sets of grids, with 50×50 and 100×100 meshes, are tested. Fig. 8 and Fig. 9 demonstrate the water heights generated from two temporal discretizations on 50×50 meshes, and their contour plots. A large disagreement of shock locations can be observed. In Fig. 10, we show $h(x, 50, 300)$ computed on both coarse ($N = 50$) and fine ($N = 100$) meshes. We also include the numerical solution computed by new RK-ERK time integration on 200×200 meshes as a reference solution. From these figures, one can see that the new time integration captures the shock location well even on coarse ($N = 50$) meshes while the exponential RK method does not.

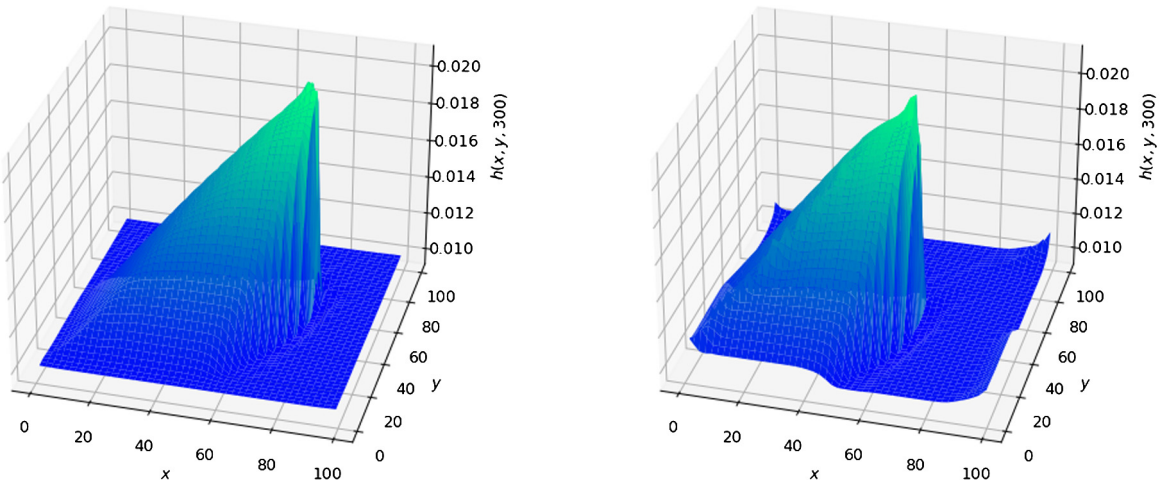

Fig. 8. Water heights at $T = 300$ on 50×50 grids, for the test in Section 4.6. Left: RK-ERK method; right: exponential RK method.

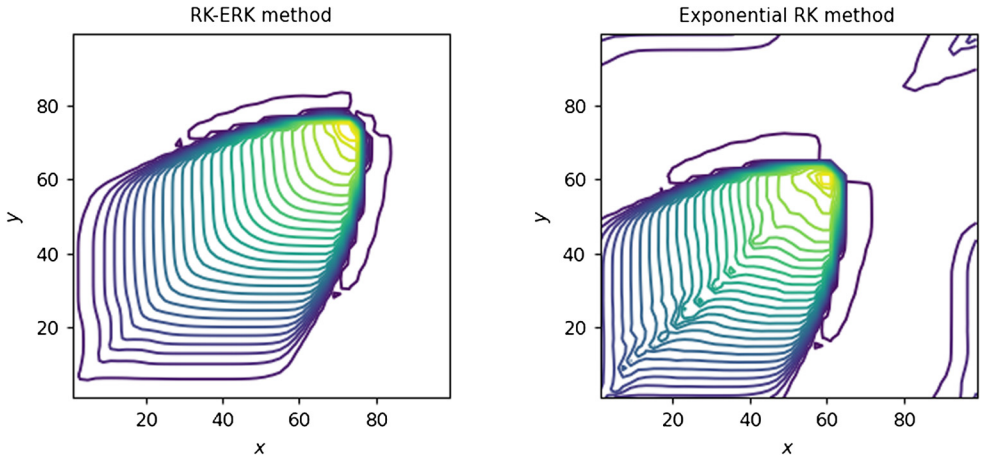

Fig. 9. The contours of water heights for the test in Section 4.6. 30 uniformly spaced contour lines.

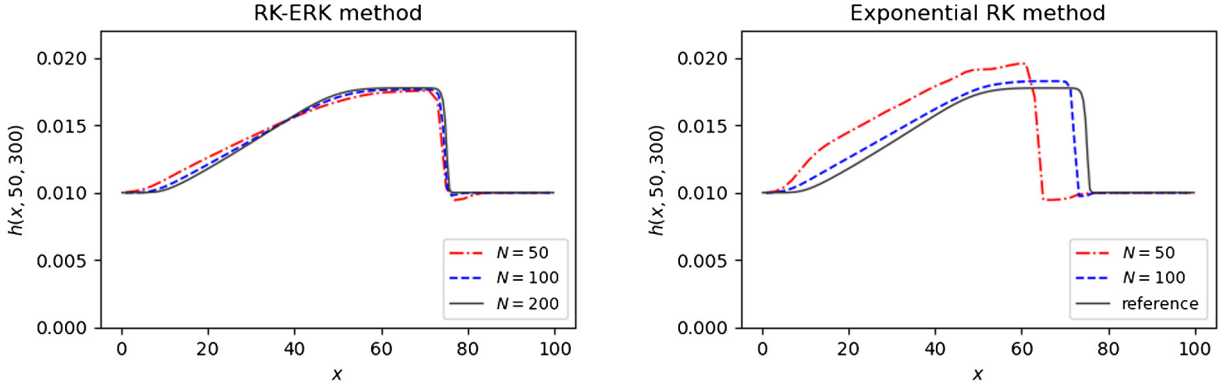

Fig. 10. Comparison of water heights at $y = 50$, $T = 300$ on different mesh sizes, for the test in Section 4.6. Left: RK-ERK method; right: exponential RK method.

4.7. Two-dimensional small perturbation test

In the last example, we extend the two dimensional small perturbation test in [29], and include friction term in the simulation. The SWEs (1.1) with $n = 0.09$ is considered, with the bottom topography

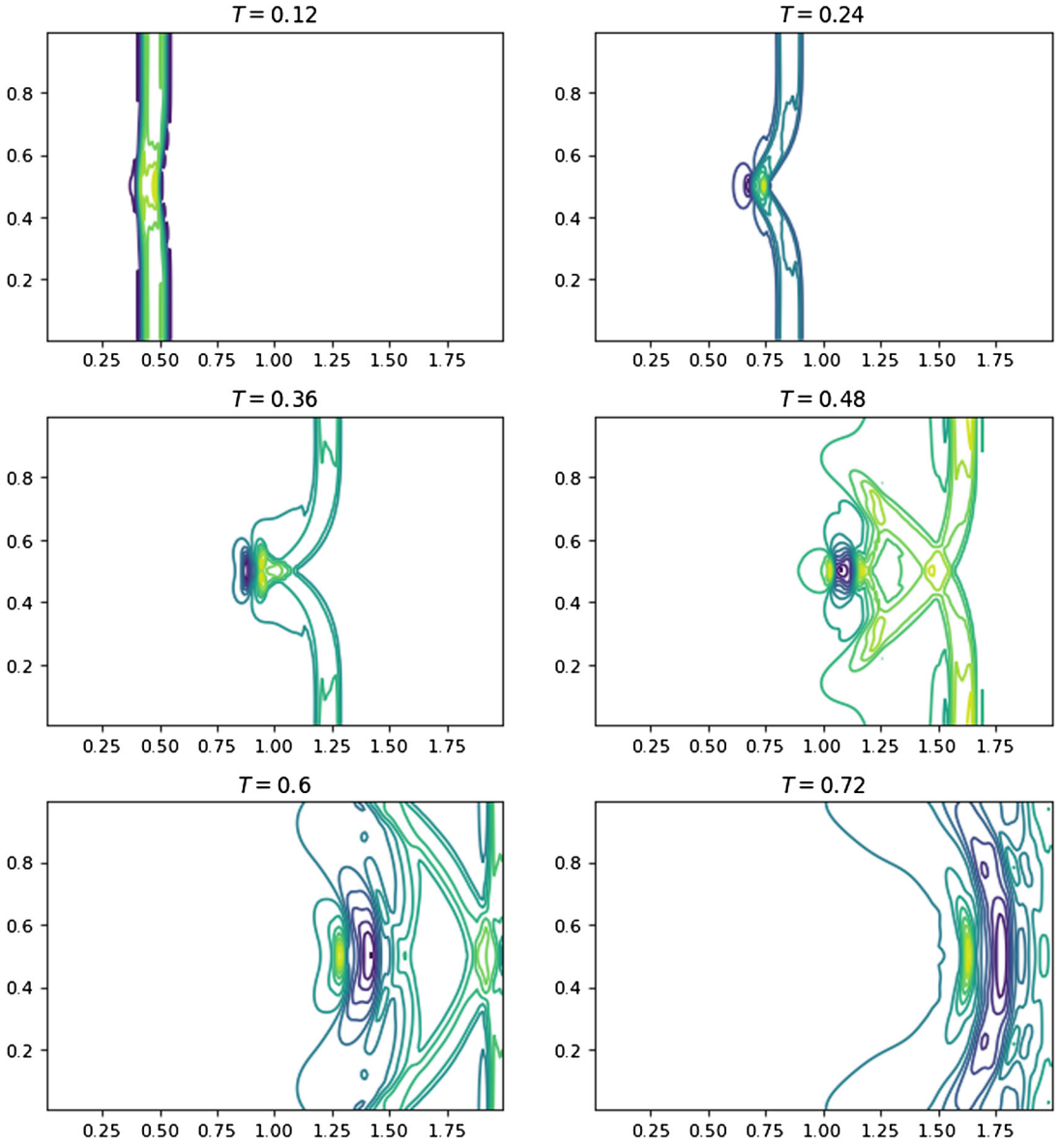

Fig. 11. The contours of the surface level $h+b$ at various times $T = 0.12, 0.24, 0.36, 0.48, 0.6, 0.72$ for the two-dimensional perturbation test in Section 4.7. 15 uniformly spaced contour lines.

$$b(x, y) = 0.8\exp(-5(x-0.9)^2 - 50(y-0.5)^2)$$

in a rectangular domain $[0, 2] \times [0, 1]$. The initial condition is given by

$$h(x, y, 0) + b(x, y) = \begin{cases} 1.01, & 0.05 \leqslant x \leqslant 0.15, \\ 1, & \text{otherwise}, \end{cases} \qquad q(x, y, 0) = p(x, y, 0) = 0.$$

We use the outlet boundary conditions. TVB constant M is taken as 10 in the test. We run the simulation on 80×160 cells, and the surface level $h + b$ at various times are presented in Fig. 11, from which we can observe the propagation of the wave to the right and its interaction with the non-flat bottom topography. Since the friction term is very little, the numerical result is almost same as [29, Fig. 19].

5. Conclusion

A family of second and third order temporal discretizations is proposed for systems of partially stiff ordinary differential equations, based on a combination of traditional RK method and exponential RK method. We provide the rigorous analysis

to show that it maintains the same order of accuracy. We considered its application in solving the SWEs with friction term, and have presented the high-order sign-preserving, positivity-preserving and well-balanced DG methods. Numerical results are given to illustrate the high-order accuracy of the new scheme and its ability to preserve signs and steady states.

CRediT authorship contribution statement

Ruize Yang: Conceptualization, Methodology, Software, Validation, Writing, Visualization. Yang Yang: Conceptualization, Methodology, Writing. Yulong Xing: Conceptualization, Methodology, Writing, Supervision.

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.

References

[1] E. Audusse, F. Bouchut, M.-O. Bristeau, R. Klein, B. Perthame, A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows, SIAM J. Sci. Comput. 25 (2004) 2050–2065.
[2] O. Bokhove, Flooding and drying in discontinuous Galerkin finite-element discretizations of shallow-water equations. Part 1: one dimension, J. Sci. Comput. 22 (2005) 47–82.
[3] A. Bollermann, G. Chen, A. Kurganov, S. Noelle, A well-balanced reconstruction of wet/dry fronts for the shallow water equations, J. Sci. Comput. 56 (2013) 267–290.
[4] A. Bermudez, M.E. Vazquez, Upwind methods for hyperbolic conservation laws with source terms, Comput. Fluids 23 (1994) 1049–1071.
[5] A. Chertock, S. Cui, A. Kurganov, T. Wu, Steady state and sign preserving semi-implicit Runge-Kutta methods for ODEs with stiff damping term, SIAM J. Numer. Anal. 53 (2015) 2008–2029.
[6] A. Chertock, S. Cui, A. Kurganov, T. Wu, Well-balanced positivity preserving central-upwind scheme for the shallow water system with friction terms, J. Numer. Methods Fluids 78 (2015) 355–383.
[7] M.J. Castro Díaz, J.A. López-García, C. Parés, High order exactly well-balanced numerical methods for shallow water systems, J. Comput. Phys. 246 (2013) 242–264.
[8] M.J. Castro, A. Pardo Milanés, C. Parés, Well-balanced numerical schemes based on a generalized hydrostatic reconstruction technique, Math. Models Methods Appl. Sci. 17 (2007) 2055–2113.
[9] L. Cea, E. Bladé, A simple and efficient unstructured finite volume scheme for solving the shallow water equations in overland flow applications, Water Resour. Res. 51 (2015) 5464–5486.
[10] B. Cockburn, C.-W. Shu, TVB Runge-Kutta local projection discontinuous Galerkin finite element methods for conservation laws II: general framework, Math. Comput. 52 (1989) 411–435.
[11] J. Du, Y. Yang, Third-order conservative sign-preserving and steady-state-preserving time integrations and applications in stiff multispecies and multi-reaction detonations, J. Comput. Phys. 395 (2019) 489–510.
[12] J. Du, C. Wang, C. Qian, Y. Yang, High-order bound-preserving discontinuous Galerkin methods for stiff multispecies detonation, SIAM J. Sci. Comput. 41 (2019) 250–273.
[13] A. Ern, S. Piperno, K. Djadel, A well-balanced Runge-Kutta discontinuous Galerkin method for the shallow-water equations with flooding and drying, Int. J. Numer. Methods Fluids 58 (2008) 1–25.
[14] J.-L. Guermond, M.Q. de Luna, B. Popov, C. Kees, M. Farthing, Well-balanced second-order finite element approximation of the shallow water equations with friction, SIAM J. Sci. Comput. 40 (2018) 3873–3901.
[15] S. Gottlieb, C.-W. Shu, E. Tadmor, Strong stability-preserving high-order time discretization methods, SIAM J. Sci. Comput. 43 (2001) 89–112.
[16] J. Huang, C.-W. Shu, Bound-preserving modified exponential Runge-Kutta discontinuous Galerkin methods for scalar hyperbolic equations with stiff source terms, J. Comput. Phys. 361 (2018) 111–135.
[17] G. Kesserwani, Q. Liang, Well-balanced RKDG2 solutions to the shallow water equations over irregular domains with wetting and drying, Comput. Fluids 39 (2010) 2040–2050.
[18] A. Kurganov, Finite-volume schemes for shallow-water equations, Acta Numer. 27 (2018) 289–351.
[19] A. Kurganov, D. Levy, Central-upwind schemes for the Saint-Venant system, Math. Model. Numer. Anal. 36 (2002) 397–425.
[20] R.J. LeVeque, Balancing source terms and flux gradients on high-resolution Godunov methods: the quasi-steady wave-propagation algorithm, J. Comput. Phys. 146 (1998) 346–365.
[21] V. Michel-Dansac, C. Berthon, S. Clain, F. Foucher, A well-balanced scheme for the shallow-water equations with topography or Manning friction, J. Comput. Phys. 335 (2017) 115–154.
[22] S. Noelle, N. Pankratz, G. Puppo, J.R. Natvig, Well-balanced finite volume schemes of arbitrary order of accuracy for shallow water flows, J. Comput. Phys. 213 (2006) 474–499.
[23] B. Perthame, C. Simeoni, A kinetic scheme for the Saint-Venant system with a source term, Calcolo 38 (2001) 201–231.
[24] C.-W. Shu, S. Osher, Efficient implementation of essentially non-oscillatory shock-capturing schemes, J. Comput. Phys. 77 (1988) 439–471.
[25] X. Wen, W.S. Don, Z. Gao, Y. Xing, Entropy stable and well-balanced discontinuous Galerkin methods for the nonlinear shallow water equations, J. Sci. Comput. 83 (2020) 66.
[26] X. Xia, Q. Liang, A new efficient implicit scheme for discretising the stiff friction terms in the shallow water equations, Adv. Water Resour. 117 (2018) 87–97.
[27] Y. Xing, Exactly well-balanced discontinuous Galerkin methods for the shallow water equations with moving water equilibrium, J. Comput. Phys. 257 (2014) 536–553.
[28] Y. Xing, C.-W. Shu, High order finite difference WENO schemes with the exact conservation property for the shallow water equations, J. Comput. Phys. 208 (2005) 206–227.
[29] Y. Xing, C.-W. Shu, High order well-balanced finite volume WENO schemes and discontinuous Galerkin methods for a class of hyperbolic systems with source terms, J. Comput. Phys. 214 (2006) 567–598.
[30] Y. Xing, C.-W. Shu, A new approach of high order well-balanced finite volume WENO schemes and discontinuous Galerkin methods for a class of hyperbolic systems with source terms, Commun. Comput. Phys. 1 (2006) 100–134.
[31] Y. Xing, C.-W. Shu, High-order finite volume WENO schemes for the shallow water equations with dry states, Adv. Water Resour. 34 (2011) 1026–1038.

[32] Y. Xing, C.-W. Shu, A survey of high order schemes for the shallow water equations, J. Math. Study 47 (2014) 221–249.
[33] Y. Xing, X. Zhang, Positivity-preserving well-balanced discontinuous Galerkin methods for the shallow water equations on unstructured triangular meshes, J. Sci. Comput. 57 (2013) 19–41.
[34] Y. Xing, X. Zhang, C.-W. Shu, Positivity-preserving high order well-balanced discontinuous Galerkin methods for the shallow water equations, Adv. Water Resour. 33 (2010) 1476–1493.
[35] X. Zhang, C.-W. Shu, On maximum-principle-satisfying high order schemes for scalar conservation laws, J. Comput. Phys. 229 (2010) 3091–3120.