HELLENIC REPUBLIC
National and Kapodistrian University of Athens
EST. 1837

Numerical methods for Shallow Water Equations

Ph.D. Thesis

Grigorios Kounadis

National and Kapodistrian University of Athens
Department of Mathematics

April, 2020

Advisory Committee

Vassilios Dougalis, Emeritus Professor, National and Kapodistrian University of Athens (Supervisor)

Theodoros Katsaounis, Professor, University of Crete

Sotirios Notaris, Professor, National and Kapodistrian University of Athens

Examination Committee

Vassilios Dougalis, Emeritus Professor, National and Kapodistrian University of Athens

Michael Dracopoulos, Assistant Professor, National and Kapodistrian University of Athens

Theodoros Katsaounis, Professor, University of Crete

Marilena Mitrouli, Professor, National and Kapodistrian University of Athens

Sotirios Notaris, Professor, National and Kapodistrian University of Athens

Ioannis Stratis, Professor, National and Kapodistrian University of Athens

Dimitrios Thilikos, Professor, National and Kapodistrian University of Athens

---

Τριμελής Συμβουλευτική Επιτροπή

Βασίλειος Δουγαλής, Ομότιμος Καθηγητής, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών (Επιβλέπων)

Θεόδωρος Κατσαούνης, Καθηγητής, Πανεπιστήμιο Κρήτης

Σωτήριος Νοτάρης, Καθηγητής, Πανεπιστήμιο Αθηνών

Επταμελής Εξεταστική Επιτροπή

Βασίλειος Δουγαλής, Ομότιμος Καθηγητής, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Μιχαήλ Δρακόπουλος, Επίκουρος Καθηγητής, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Δημήτριος Θηλυκός, Καθηγητής, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Θεόδωρος Κατσαούνης, Καθηγητής, Πανεπιστήμιο Κρήτης

Μαριλένα Μητρούλη, Καθηγήτρια, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Σωτήριος Νοτάρης, Καθηγητής, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Ιωάννης Στρατής, Καθηγητής, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Preface

This thesis has been written in the Mathematics Department of the University of Athens under the guidance of my advisor Prof. Vassilios Dougalis. I am immensely grateful for his guidance with his vast experience and knowledge.

I would also like to sincerely thank the other two members of my PhD advisory committee, Prof. Sotirios Notaris of the Mathematics Department of the University of Athens, and Prof. Theodore Katsaounis of the Mathematics Department of the University of Crete, for reading this thesis and for their help, remarks, and corrections, and for following my progress towards the PhD all these years and providing me with their invaluable support.

I also would like to thank Professor Katsaounis for making possible my visit to the Applied PDE group of Prof. T. Tzavaras at the King Abdullah University of Science and Technology (KAUST) in Saudi Arabia during November and December of 2018 and for his generous help and advice on adaptive Discontinuous Galerkin methods.

My sincere thanks are also due to my friends and former members of the Numerical PDE team at the University of Athens, Dr. Dimitrios Antonopoulos for his constant support, collaboration, and advice, especially on chapters 2 and 3 of this thesis, and Dr. Dimitrios Mitsotakis of the Victoria University of Wellington, New Zealand, for his constant support and advice, especially on chapters 1, 2, and 4 of the thesis, and for making available to me his data for the Serre-Green-Naghdi numerical simulations in [MSM17] quoted in the last two numerical experiments of chapter 2 of the thesis.

I would also like to thank sincerely the Institute of Applied and Computational Mathematics (IACM) of the Foundation of Research and Technology, Hellas (FORTH) for providing financial support in the form of scholarships and research assistantships during my doctoral studies in the period 2013-2020 through the grants KRIPIS I (PEFYKA) of ESPA, Siemens-Environment, ARCHERS (a grant of the Stavros Niarchos Foundation to FORTH), and KRIPS II (PERAN) of ESPA. My thanks go of course to the granting agencies of these grants as well.

My thanks also go to Prof. Michael Dracopoulos for all his help during my graduate studies.

Last but not least, I would like to thank Professors Marilena Mitrouli, Ioannis Stratis, Dimitrios Thilikos, and Michael Dracopoulos, all of the Department of Mathematics of the University of Athens, who , together with the advisory com-

mittee, served as members of the thesis Examination Committee.

G.Kounadis,
April 2020

Περίληψη Διατριβής

Η διατριβή αποτελείται από τέσσερα Κεφάλαια.

1. Το πρώτο κεφάλαιο, "Εισαγωγή", ξεκινά από τις εξισώσεις του Euler για την περιγραφή των κυμάτων επιφανείας ενός τέλειου ρευστού (π.χ. νερού) σε διδιάστατο κυματοδηγό πεπερασμένου βάθους με μεταβλητή τπογραφία πυθμένα. Οι εξισώσεις γράφονται σε αδιάστατη, κανονικοποιημένη μορφή με χρήση των παραμέτρων κλίμακας $\varepsilon = \alpha_0/\lambda_0$, $\mu = (D_0/\lambda_0)^2$, όπου α_0 είναι ένα τυπικό πλάτος των κυμάτων επιφανείας, λ_0 ένα τυπικό μήκος κύματος, και D_0 ένα μέσο βάθος πυθμένα.

Από τις εξισώσεις του Euler παράγεται συστηματικά μια σειρά από απλούστερα μαθηματικά μοντέλα που αποτελούν προσεγγίσεις των εξισώσεων του Euler για την περιγραφή της κίνησης μη γραμμικών, διασπειρομένων κυμάτων επιφανείας σε δύο κατευθύνσεις σε μία διάσταση, τα οποία έχουν μεγάλο μήκος κύματος σχετικά με το μέσο βάθος του πυθμένα ("κύματα ρηχών υδάτων"), δηλ. για τα οποία ισχύει $\mu \ll 1$. Το βασικό μοντέλο (ένα μη γραμμικό σύστημα Μ.Δ.Ε.) που εξάγεται είναι οι λεγόμενες εξισώσεις Serre-Green-Naghdi (SGN) με μεταβλητό πυθμένα, από τις οποίες παράγονται εν συνεχεία τρία απλούστερα μαθηματικά μοντέλα (μη γραμμικά συστήματα Μ.Δ.Ε.) σε ειδικές περιοχές των παραμέτρων κλίμακας, και τα οποία εξετάζονται λεπτομερώς στη διατριβή. Το πρώτο μοντέλο είναι το κλασσικό σύστημα Boussinesq με μεταβλητό πυθμένα γενικής τοπογραφίας (CBs), γνωστό και ως σύστημα του Peregrine, το οποίο, σε κανονικοποιημένες αδιάστατες μεταβλητές, είναι της μορφής

$$\begin{aligned} &\zeta_t + (\eta u)_x = 0, \\ &u_t + \zeta_x + \varepsilon u u_x + \mu \left(\tfrac{\beta}{2} \eta_b b'' u_t + \beta \eta_b b' u_{xt} - \tfrac{1}{3} \eta_b^2 u_{xxt} \right) = 0, \end{aligned} \tag{CBs}$$

όπου $\varepsilon\zeta = \varepsilon\zeta(x,t)$ η μεταβολή της ελεύθερης επιφάνειας του ρευστού ως προς μία ηρεμούσα κατάσταση, $\eta_b(x) = 1 - \beta b(x) > 0$ (όπου $\beta = \frac{B}{D_0}$, B ένα τυπικό μέγεθος της μεταβολής της τοπογραφίας του πυθμένα, και $b(x)$ η συνάρτηση της τοπογραφίας του πυθμένα), $u = u(x,t)$ η μέση ως προς το βάθος οριζόντια ταχύτητα του ρευστού, και $\eta = \varepsilon\zeta + \eta_b > 0$ το συνολικό βάθος της στήλης του νερού. Η περιοχή των παραμέτρων για την οποία το (CBs) είναι καλή προσέγγιση του (SGN) είναι η λεγόμενη περιοχή της "προσέγγισης Boussinesq" $\varepsilon = \mathcal{O}(\mu)$, δηλ. διασπειρόμενων κυμάτων κατάλληλα μικρού πλάτους και μεγάλου μήκους. Το β είναι της τάξης $\mathcal{O}(1)$, δηλ. η τοπογραφία μπορεί να μεταβάλλεται ισχυρά.

Αν στο (CBs) υποτεθεί επιπλέον ότι $\beta = \mathcal{O}(\varepsilon)$, δηλ. ότι ο πυθμένας παρουσιάζει σχετικά μικρή μεταβολή, προκύπτει, αν παραλείψουμε όρους ανωτέρας τάξης, ένα ακόμα πιο απλουστευμένο κλασσικό σύστημα Boussinesq μεταβλητού πυθμένα, της μορφής

$$\begin{aligned} &\zeta_t + (\eta u)_x = 0, \\ &u_t + \zeta_x + \varepsilon u u_x - \tfrac{\mu}{3} u_{xxt} = 0, \end{aligned} \tag{CBw}$$

με $\eta = \varepsilon\zeta + \eta_b$. Τέλος, αν π.χ. στο (CBw) υποτεθεί ότι $\mu = 0$, δηλ. ότι τα κύματα δεν παρουσιάζουν διασπορά, προκύπτει το (υπερβολικό) σύστημα των ρηχών υδάτων, για το οποίο το ε μπορεί να είναι της τάξης του 1. Το σύστημα αυτό εξετάζεται επίσης στην διατριβή και είναι της μορφής

$$\begin{aligned} &\eta_t + (\eta u)_x = 0, \\ &u_t + \zeta_x + \varepsilon u u_x = 0. \end{aligned} \tag{SW}$$

Η υπόλοιπη διατριβή αφορά την αριθμητική ανάλυση και αριθμητικές προσομοιώσεις των λύσεων των τριών συστημάτων (CBs), (CBw), (SW) με μεθόδους Galerkin-πεπερασμένων στοιχείων, συμπεριλαμβανομένης και της μεταξύ τους σύγκρισης, καθώς και της σύγκρισης με αντίστοιχες λύσεις του συστήματος (SGN).

2. Το δεύτερο κεφάλαιο αφορά την αριθμητική ανάλυση προβλημάτων αρχικών και συνοριακών συνθηκών για τα συστήματα (CBs), (CBw), σε πεπερασμένο διάστημα με $u = 0$ στο σύνορο. Συγκεκριμένα, αφού γίνει μία ανασκόπηση της θεωρίας ύπαρξης-μοναδικότητας των λύσεων των προβλημάτων αυτών, τα συστήματα διακριτοποιούνται ως προς την χωρική συνιστώσα με την συνήθη μέθοδο Galerkin-πεπερασμένων στοιχείων επί ημι-ομοιόμορφων διαμερισμών $\{x_i\}$ με $\max\ (x_{i+1} - x_i) = h$ με κατά τμήματα πολυωνυμικές συναρτήσεις τάξης $r \geq 3$ (δηλ. βαθμού $r - 1 \geq 2$) και εκτιμάται το σφάλμα της ημιδιακριτοποίησης αυτής στον $L^2 \times H^1$. Και για τα δύο συστήματα αποδεικνύονται με την μέθοδο της ενέργειας εκτιμήσεις των σφαλμάτων της μορφής

$$\max_{0 \leq t \leq T} (\|\zeta - \zeta_h\| + \|u - u_h\|_1) \leq C\, h^{r-1}, \tag{2.1}$$

(όπου ζ_h, u_h είναι οι ημιδιακριτές προσεγγίσεις των ζ, u, αντίστοιχα), υπό την προϋπόθεση ότι οι ζ, u είναι αρκετά ομαλές από $t = 0$ μέχρι $t =$ και έχουν προσεγγιστεί κατάλληλα οι αρχικές συνθήκες $\zeta(x, 0)$, $u(x, 0)$. Στην παραπάνω ανισότητα το C είναι μια σταθερά ανεξάρτητη του h. Στα αριθμητικά πειράματα που παρατίθενται επίσης στο δεύτερο κεφάλαιο (παράγραφος 2.3), οι αριθμητικές μέθοδοι υλοποιούνται με διαφόρου βαθμού splines ως προς x και με τη κλασσική, άμεση μέθοδο Runge-Kutta τέταρτης τάξης ακρίβειας ως προς t (η οποία είναι ευσταθής υπό την μη περιοριστική συνθήκη $\frac{k}{h} \leq c$ όπου k το βήμα της διακριτοποίησης ως προς t). Στην παράγραφο 2.3.1 επιβεβαιώνεται υπολογιστικά η εκτίμηση (2.1) για γενικό διαμερισμό και παρατηρείται ότι ισχύει και για κατά τμήματα γραμμικά πολυώνυμα (δηλ. για $r = 2$). Για ομοιόμορφη χωρική διαμέριση η τάξη ακρίβειας βελτιώνεται. Τα αριθμητικά πειράματα γίνονται με μεγάλη (τετραπλή) υπολογιστική ακρίβεια και επιτρέπουν λεπτές παρατηρήσεις όπως π.χ. την κατά τα φαινόμενα απουσία

λογαριθμικών όρων στα αντίστοιχα σφάλματα των κυβικών splines για το κλασσικό σύστημα Boussinesq με οριζόντιο πυθμένα (δηλ. $\beta = 0$) που υπάρχουν στην βιβλιογραφία.

Στην παράγραφο 2.3.2 εξετάζονται αριθμητικές μέθοδοι Galerkin-πεπερασμένων στοιχείων για τα παραπάνω συστήματα Boussinesq και τις εξισώσεις ρηχών υδάτων (SW) με απορροφητικές συνθήκες στο σύνορο, που επιτρέπουν την έξοδο κυματισμών από το υπ' όψιν πεπερασμένο υπολογιστικό διάστημα χωρίς την εμφάνιση αριθμητικών φαινομένων ανάκλασης από το σύνορο. Στην περίπτωση του (SW) υπάρχουν ακριβείς απορροφητικές (δηλ. πλήρως διαφανείς) συνθήκες, που βασίζονται στην κλασσική θεωρία των χαρακτηριστικών, ενώ στην περίπτωση των συστημάτων Boussinesq θα ήταν δυνατή ίσως η εξαγωγή απορροφητικών συνθηκών, αλλά μη τοπικών, πράγμα που θα τις καθιστούσε δύσχρηστες. Αντ' αυτού, επειδή στα συστήματα Boussinesq το μ είναι πολύ μικρό μπορεί να υποτεθεί ότι οι αναλλοίωτοι του Riemann (στις οποίες στηρίζεται η μέθοδος των χαρακτηριστικών για τις (SW)) δεν μεταβάλλονται πολύ τοπικά κοντά στο σύνορο, πράγμα που επιτρέπει την επέκταση των χαρακτηριστικών απορροφητικών συνθηκών των (SW) στα συστήματα Boussinesq (ακόμα και στις (SGN))· οι νέες συνθήκες είναι κατά προσέγγιση απορροφητικές για τα (CB) και στην παρ. 2.3.2 διερευνάται λεπτομερώς (υπολογιστικά) η ακρίβεια τους και το μέγεθος των αριθμητικών ανακλάσεων από το σύνορο στην περίπτωση εξερχομένων μοναχικών κυμάτων του κλασσικού συστήματος Boussinesq στην περίπτωση οριζόντιου πυθμένα. Επιβεβαιώνεται ότι οι χαρακτηριστικές συνθήκες απορρόφησης είναι καλές προσεγγίσεις διαφανών συνθηκών για αρκετά μικρό μ ακόμη και στην περίπτωση πυθμένων μεταβλητής τοπογραφίας.

Στην παράγραφο 2.3.3 εξετάζονται υπολογιστικά, με βάση κυρίως το μοντέλο (CBs), διάφορα φαινόμενα που αφορούν τις μεταβολές που υφίσταται ένα αρχικά μοναχικό κύμα όταν κινείται σε περιβάλλον με πυθμένα μεταβλητής τοπογραφίας. Χρησιμοποιείται το (CBs) διακριτοποιημένο με κυβικές splines και την 4ης τάξης μέθοδο RK για την αριθμητική προσομοίωση ενός μοναχικού κύματος που αναρριχάται σε κεκλιμένο επίπεδο, καθώς και σε βυθό που είναι αρχικά κεκλιμένος και μετά οριζόντιος (υφαλοκρηπίς). Αριθμητικές λύσεις για τα δύο αυτά προβλήματα είναι γνωστές από τη βιβλιογραφία. Γίνεται, μεταξύ των άλλων, λεπτομερής μελέτη της μεταβολής του σχήματος και του πλάτους του μοναχικού κύματος κατά την εξέλιξη του φαινομένου, των ανακλάσεων που δημιουργούνται λόγω της μεταβολής της τοπογραφίας του πυθμένα, καθώς και του φαινομένου της ανάλυσης του αρχικού μοναχικού κύματος (αφού αναρριχηθεί στην υφαλοκρηπίδα) σε σειρά μοναχικών κυμάτων. Εξετάζεται η επιρροή της μεταβολής της τοπογραφίας του πυθμένα και συγκρίνονται οι αριθμητικές λύσεις που προκύπτουν από τα δύο μοντέλα (CBs) και (CBw) καθώς το μοναχικό κύμα διέρχεται πάνω από μεταβλητό πυθμένα όταν το β αυξάνει σε μέγεθος από $\mathcal{O}(\varepsilon)$ σε $\mathcal{O}(1)$· επιβεβαιώνεται ότι το (CBw) δεν δίνει σωστή ποιοτική εικόνα της ροής για $\beta = \mathcal{O}(1)$. Τέλος, συγκρίνονται αριθμητικά αποτελέσματα που λαμβάνονται από το (CBs) και τις εξισώσεις SGN στην περίπτωση δύο ρεαλιστικών προβλημάτων δοκιμής της βιβλιογραφίας για τα οποία υπάρχουν αριθμητικά και πειραματικά αποτελέσματα. Όπως αναμένε-

ται το μοντέλο (SGN) είναι πιο κοντά στα πειραματικά δεδομένα αλλά διαπιστώνεται ότι και το (CBs) δίνει γενικά αρκετά καλά αποτελέσματα.

3. Στο τρίτο κεφάλαιο της διατριβής εξετάζεται το σύστημα (SW) των εξισώσεων ρηχών υδάτων με μεταβλητό πυθμένα υπό την προϋπόθεση ότι έχει ομαλές λύσεις. (Ως γνωστόν το πρόβλημα αρχικών τιμών για τις (SW) έχει, αν οι αρχικές συνθήκες είναι ομαλές, μόνο τοπικές ως προς t ομαλές λύσεις και, γενικά, εμφανίζει ασυνέχειες καθώς ο χρόνος αυξάνει.) Κατ' αρχάς αποδεικνύονται εκτιμήσεις σφαλμάτων στον χώρο $L^2 \times L^2$ για το πρόβλημα αρχικών-συνοριακών συνθηκών για τις (SW) με $u = 0$ στα άκρα πεπερασμένου διαστήματος όταν διακριτοποιείται ως προς x με την συνήθη μέθοδο Galerkin-πεπερασμένων στοιχείων. Τα σφάλματα έχουν φράγμα της μορφής $C\, h^{r-1}$ αν $r \geq 3$. Στη συνέχεια αποδεικνύονται παρόμοιες εκτιμήσεις σφαλμάτων για το πρόβλημα αρχικών-συνοριακών τιμών για τις (SW) με χαρακτηριστικές συνοριακές συνθήκες απορρόφησης όπου οι ροές είναι υπερκρίσιμες ή υποκρίσιμες. Ακολουθεί η υπολογιστική μελέτη των συστημάτων, τα οποία διακριτοποιούνται ως προς t πάλι με την κλασσική, άμεση $(4, 4)$ μέθοδο RK. Τα αριθμητικά πειράματα δείχνουν ότι οι εκτιμήσεις των σφαλμάτων της ημιδιακριτής προσέγγισης για όλες τις σ.σ. υπό μελέτη ισχύουν και για $r = 2$ (δηλ. και για διακριτοποίησεις με συνεχείς, κατά τμήματα γραμμικές συναρτήσεις). Μάλιστα για $r = 2$ και στην περίπτωση ομοιόμορφου διαμερισμού φαίνεται ότι τα σφάλματα είναι βέλτιστης τάξης ακρίβειας, δηλ. $\mathcal{O}(h^2)$.

Το σύστημα (SW) έχει λύσεις *σταθερής μορφής*, ανεξάρτητες του χρόνου, στις οποίες π.χ. τείνουν ομαλές ολικές λύσεις καθώς ο χρόνος αυξάνει. Εξετάζεται υπολογιστικά η ικανότητα της συνήθους μεθόδου Galerkin να προσεγγίζει τις χρονικά ανεξάρτητες αυτές καταστάσεις· τα αποτελέσματα είναι πολύ ικανοποιητικά και τα σφάλματα εξετάζονται ποσοτικά στην περίπτωση αρκετών παραδειγμάτων. Τέλος, εξετάζεται αν η διακριτοποίηση με την συνήθη μέθοδο Galerkin του συστήματος (SW), γραμμένου σε μορφή νόμου ισορροπίας, διατηρεί τις λύσεις του συστήματος της μορφής "ηρεμουσών ροών", δηλ. με $u = 0$ και οριζόντια ελεύθερη επιφάνεια. (Αυτό δεν είναι προφανές για τυχαία αριθμητική μέθοδο στην περίπτωση μεταβλητού πυθμένα). Αν όμως κάτι τέτοιο συμβαίνει, οι μέθοδοι λέγονται "καλώς εξισορροπημένες"). Αποδεικνύεται ότι η συνήθης μέθοδος Galerkin είναι καλώς εξισορροπημένη όταν ο όρος πηγής υπολογιστεί με μεγαλύτερης ακρίβειας κανόνα αριθμητικής ολοκλήρωσης, σε σύγκριση με την ακρίβεια του κανόνα αριθμητικής ολοκλήρωσης που διατηρεί απλώς το σφάλμα διακριτοποίησης της μεθόδου.

4. Στο τέταρτο κεφάλαιο της διατριβής εξετάζεται η ασυνεχής μέθοδος Galerkin-πεπερασμένων στοιχείων (DG) για το σύστημα εξισώσεων ρηχών υδάτων, γραμμένο σε μορφή νόμου ισορροπίας. Οι μέθοδοι αυτές χρησιμοποιούνται ευρέως σήμερα, μεταξύ των άλλων για την διακριτοποίηση μη γραμμικών υπερβολικών προβλημάτων στα οποία δημιουργούνται ασυνέχειες (ωστικά κύματα, υδραυλικά άλματα, κ.α.). Στην παράγραφο 4.1.1 γίνεται μία ανασκόπηση των μεθόδων RKDG, (δηλ. ασυνεχών μεθόδων Galerkin για την ημιδιακριτοποίηση ως προς την χωρική συνιστώσα και αμέσων μεθόδων RK για την διακριτοποίηση ως προς t· επιλέγεται η μέθοδος Shu-Osher η οποία είναι άμεση RK τρίτης τάξης ακρίβειας), όταν εφαρμόζονται σε υπερβολικά συστήματα νόμων διατήρησης σε μία χωρική διάσταση.

Εξετάζονται οι επιλογές για τεχνικές περιορισμού κλίσης (slope limiter), απαραίτητου εξαρτήματος της μεθόδου DG για την επιτυχή προσομοίωση ασυνεχειών και γίνεται επιλογή του λεγόμενου minmod limiter.

Στην παράγραφο 4.2 εξετάζεται η εφαρμογή των μεθόδων RKDG στην περίπτωση των εξισώσεων ρηχών υδάτων με μεταβλητό πυθμένα, οι οποίες γράφονται σε μορφή νόμου ισορροπίας με την συνάρτηση τοπογραφίας πυθμένα σε όρο "πηγής" στο δεύτερο μέλος. Παρ' ότι οι βασικές αρχές της μεθόδου DG για τις SW είναι γνωστές από τη βιβλιογραφία, στη διατριβή γίνεται ενδελεχής εξέταση της μεθόδου και λεπτομερής κατασκευή του αλγορίθμου της με συμπλήρωση και έλεγχο με αριθμητικά πειράματα όλων των βημάτων του ώστε να μπορεί να εφαρμόζεται σε πολύπλοκα μονοδιάστατα προβλήματα. Εξετάζονται θέματα και αλγόριθμοι καλής εξισορρόπησης, διατήρησης του μη-αρνητικού βάθους της στήλης νερού στην περίπτωση που η τοπογραφία του πυθμένα πλησιάζει ή υπερβαίνει την ελεύθερη επιφάνεια, περιορισμού της κλίσης σε περίπτωση ασυνεχειών κ.α.. Τέλος παρατίθεται μια σειρά αριθμητικών παραδειγμάτων της βιβλιογραφίας τα οποία ο αλγόριθμος της διατριβής προσεγγίζει με μεγάλη ακρίβεια, συμπεριλαμβανομένων προβλημάτων Riemann και θραυομένου φράγματος, προβλημάτων όπου το νερό αποσύρεται από κεκλιμένο πυθμένα, σχεδόν περιοδικών προβλημάτων ταλαντώσεων σε παραβολικό δοχείο, κ.α. Παρουσιάζεται και ένα νέο πρόβλημα δοκιμής με πολύπλοκη τοπογραφία πυθμένα, του οποίοι η επίλυση απαιτεί τον συνδυασμό όλων των τεχνικών που προαναφέρθηκαν.

Contents

Chapter 1

Introduction

In this dissertation we will study numerical methods for some nonlinear systems of pde's that model two-way propagation of long (i.e. shallow-water) surface water waves in a channel with variable-bottom topography. The systems that will be considered are two Boussinesq-type models with dispersive terms and also the Shallow Water equations that are a quasilinear hyperbolic system of pde's of balance-law form (i.e. with a source term).

The systems are derived in appropriate scaling regimes from the Euler equations of water-wave theory. They all follow from a more general "fully nonlinear" dispersive model of two-way long-wave propagation, the Serre-Green-Naghdi system of equations. This system was first derived by Serre [Ser53a],[Ser53b], in the case of horizontal bottom in one space dimension, and subsequently rederived by Su and Gardner, [SG69], and Green, Laws, and Naghdi, [GLN74], and Green and Naghdi, [GN76]; In the latter two references the system was extended to two-space dimensions. We will call it therefore *Serre-Green-Naghdi* (SGN) system.

In what follows we will formally derive the (SGN) in one space dimension from the 2D Euler equations with variable bottom and then formally derive from (SGN) the two 'classical' Boussinesq type systems and the Shallow Water equations that will be considered in the rest of the thesis.

1.1 Derivation of the SGN system

(i) The Euler equations

The 2D Euler equations of water wave theory, [Whi74], for an ideal (incompressible, irrotational) fluid, say water, in a finite-depth channel, are given for $x \in \mathbb{R}$, $t \geq 0$, in the case of dimensional, unscaled variables, by the following equations

Conservation of horizontal momentum:

$$u_t + (u^2)_x + (wu)_z = -\frac{1}{\rho} p_x, \quad -D \leq z \leq \zeta, \tag{1.1}$$

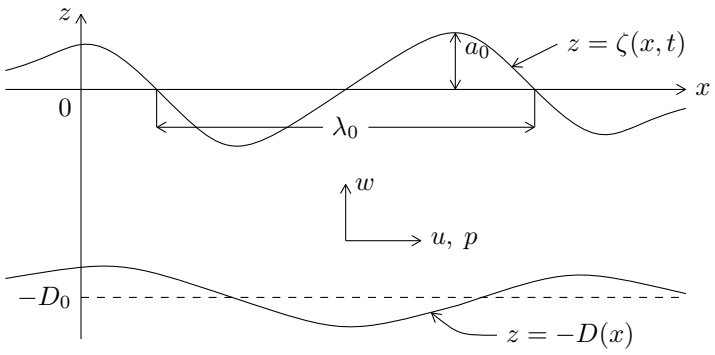


Figure 1.1: Notation for 2D Euler equations in dimensional, unscaled variables

Conservation of vertical momentum:

$$w_t + uw_x + ww_z = -g - \frac{1}{\rho}p_z, \quad -D \le z \le \zeta, \tag{1.2}$$

Incompressibility condition (continuity equation):

$$u_x + w_z = 0, \quad -D \le z \le \zeta, \tag{1.3}$$

Irrotationality condition:

$$u_z - w_x = 0, \quad -D \le z \le \zeta, \tag{1.4}$$

Kinematic surface boundary condition:

$$\zeta_t + u\zeta_x - w = 0, \quad \text{at } z = \zeta(x,t), \tag{1.5}$$

Dynamic surface boundary condition:

$$p = 0, \quad \text{at } z = \zeta(x,t), \tag{1.6}$$

Bottom boundary condition:

$$uD_x + w = 0, \quad \text{at } z = -D(x). \tag{1.7}$$

Here x is distance along the channel and $t \ge 0$ the time. The depth variable z (positive upwards) ranges, in the domain of interest, for given (x,t), between the given bottom topography function $z = -D(x)$ and the free surface which is given by the unknown function $z = \zeta(x,t)$. The undistributed water surface will be at $z = 0$. We will assume that D is sufficiently smooth for our purposes and that always $D > 0$ and $\zeta > -D$. In (1.1)–(1.7) $u = u(x,z,t)$ and $w = w(x,z,t)$ denote the horizontal and vertical component, respectively, of the velocity of the fluid at (x,z,t), $p = p(x,z,t)$ is the pressure and ρ he density of the fluid, assumed constant. We note that the conservation of horizontal momentum equation is usually

written as $u_t + uu_x + wu_z = -\frac{1}{\rho}p_x$. Using the continuity equation we see that $(u^2)_x+(wu)_z = 2uu_x+w_tu+wu_x = u(u_x+w_z)+uu_x+wu_z = uu_x+wu_z$. The form of these terms in (1.1) is more useful in the derivation of the (SGN) equations to follow.

(ii) Nondimensional variables and scaling parameters

In order to study surface wave propagation phenomena in detail, one usually derives model equations from the full 2D Euler system; these simplified equations are valid approximation of the Euler equations in specific scaling regimes of interest. To derive these models, we first write (1.1)–(1.7) in nondimensional and scaled variables following [Per72], [Dou14], [ADM].

We first nondimensionalize the Euler equations. Let D_0 denote a characteristic length of the problem, chosen here naturally as the mean depth of the channel. Then $c_0 = \sqrt{gD_0}$ defines a characteristic velocity, $\sqrt{D_0/g}$ a characteristic time and $\rho g D_0$ a characteristic pressure. We introduce *nondimensional (unscaled) variables*, denoted by $'$, by the equations

$$x' = \frac{x}{D_0}, \quad t' = \frac{t}{\sqrt{D_0/g}} = \frac{c_0 t}{D_0}, \quad z' = \frac{z}{D_0},$$
$$u' = \frac{u}{c_0}, \quad w' = \frac{w}{c_0}, \quad \zeta' = \frac{\zeta}{D_0}, \quad D' = \frac{D}{D_0}, \quad p' = \frac{p}{\rho g D_0} = \frac{p}{\rho c_0^2}.$$

Since long wavelength will be typical in our models (and in some, small amplitude,) we define the *scaling parameters*

$$\sigma = \frac{D_0}{\lambda_0}, \quad \varepsilon = \frac{a_0}{D_0},$$

where λ_0 is a typical wavelength of the problem and a_0 a typical amplitude of the surface waves. For the time being we make no assumption about their magnitude. The nondimensional, scaled variables will be denoted by *, and they are derived in terms of the nondimensional (unscaled) variables denoted by $'$, and the usual dimensional variables bearing no superscript, by the formulas

$$x^* = \sigma x' = \frac{\sigma}{D_0}x = \frac{x}{\lambda_0}, \quad z^* = z' = \frac{z}{D_0}, \quad \zeta^* = \frac{\zeta'}{\varepsilon} = \frac{\zeta}{D_0\varepsilon} = \frac{\zeta}{a_0},$$
$$D^* = D' = \frac{D}{D_0}, \quad t^* = \sigma t' = \frac{c_0}{\lambda_0}t, \quad u^* = \frac{1}{\varepsilon}u' = \frac{u}{\varepsilon c_0} = \frac{D_0 u}{a_0 c_0},$$
$$w^* = \frac{1}{\varepsilon\sigma}w' = \frac{\lambda_0}{a_0 c_0}w, \quad p^* = p' = \frac{p}{\rho g D_0}.$$

(iii) The Euler equations in nondimensional, scaled variables

Using the chain rule and some simple algebra one may easily transform the

2D-Euler equations (1.1)–(1.7) into the following equations, respectively:

$$(1.1) \Leftrightarrow \varepsilon u^*_{t^*} + \varepsilon^2 (u^{*2})_{x^*} + \varepsilon^2 (w^* u^*)_{z^*} = -p^*_{x^*}, \quad -D^* \le z^* \le \varepsilon\zeta^*,$$
$$(1.2) \Leftrightarrow \varepsilon\sigma^2 w^*_{t^*} + \varepsilon^2\sigma^2 u^* w^*_{x^*} + \varepsilon^2\sigma^2 w^* w^*_{z^*} = -1 - p^*_{z^*}, \quad -D^* \le z^* \le \varepsilon\zeta^*,$$
$$(1.3) \Leftrightarrow u^*_{x^*} + w^*_{z^*} = 0, \quad -D^* \le z^* \le \varepsilon\zeta^*,$$
$$(1.4) \Leftrightarrow u^*_{z^*} - \sigma^2 w^*_{x^*} = 0, \quad -D^* \le z^* \le \varepsilon\zeta^*,$$
$$(1.5) \Leftrightarrow \zeta^*_{t^*} + \varepsilon u^* \zeta^*_{x^*} - w^* = 0, \quad \text{at } z^* = \varepsilon\zeta^*,$$
$$(1.6) \Leftrightarrow p^* = 0, \quad \text{at } z^* = \varepsilon\zeta^*,$$
$$(1.7) \Leftrightarrow u^* D^*_{x^*} + w^* = 0, \quad \text{at } z^* = -D^*.$$

Here u^*, w^*, p^*, are functions of (x^*, z^*, t^*), $\zeta^* = \zeta^*(x^*, t^*)$, $D^* = D^*(x^*)$. The main advantage of the equations written in their nondimensional, scaled form is that when specific assumptions of the magnitudes of the scaling parameters ε, σ are made, then the order of magnitude of each specific term of the equation is determined by the order of magnitude of the monomial $\varepsilon^\alpha\sigma^\beta$ that multiplies it as coefficient. The starred variables, independent or dependent, along with their spatial or temporal derivatives will all be of $\mathcal{O}(1)$.

Finally, since using the starred variables throughout the rest of the chapter will be typographically cumbersome, we simplify the notation and revert to denoting by unstarred variables the nondimensional, scaled quantities. Hence, finally, the 2D Euler equations in nondimensional, scaled form are given by

$$\text{Horizontal momentum:} \quad \varepsilon u_t + \varepsilon^2 (u^2)_x + \varepsilon^2 (wu)_z = p_x, \quad -D \le z \le \varepsilon\zeta, \tag{1.1$'$}$$

$$\text{Vertical momentum:} \quad \varepsilon\sigma^2 w_t + \varepsilon^2\sigma^2 u w_x + \varepsilon^2\sigma^2 w w_z = -1 - p_z, \quad -D \le z \le \varepsilon\zeta, \tag{1.2$'$}$$

$$\text{Continuity:} \quad u_x + w_z = 0, \quad -D \le z \le \varepsilon\zeta, \tag{1.3$'$}$$

$$\text{Irrotationality:} \quad u_z - \sigma^2 w_x = 0, \quad -D \le z \le \varepsilon\zeta, \tag{1.4$'$}$$

$$\text{Kinematic b.c. surface:} \quad \zeta_t + \varepsilon u \zeta_x - w = 0, \quad \text{at } z = \varepsilon\zeta, \tag{1.5$'$}$$

$$\text{Dynamic b.c. surface:} \quad p = 0, \quad \text{at } z = \varepsilon\zeta, \tag{1.6$'$}$$

$$\text{Bottom b.c.:} \quad D_x u + w = 0, \quad \text{at } x = -D, \tag{1.7$'$}$$

(iv) Depth-averaged quantities; mass conservation in depth-averaged form

For a continuous function $v(x, z, t)$ defined for $-\infty < x < \infty$, $t \ge 0$, $-D \le z \le \varepsilon\zeta$ we define its depth-averaged $\overline{v} = \overline{v}(x, t)$ as the mean

$$\overline{v}(x,t) = \frac{1}{\eta(x,t)} \int_{-D(x)}^{\varepsilon\zeta(x,t)} v(x,z,t)\,\mathrm{d}z, \tag{1.8}$$

where the *water depth* $\eta = \eta(x, t)$ is defined by $\eta = \varepsilon\zeta + D$ and will always be positive. For fixed x, t, integrating (1.3′) with respect to z in the interval $[-D, \varepsilon\zeta]$ we obtain

$$0 = \int_{-D}^{\varepsilon\zeta} u_x\,\mathrm{d}z + \int_{-D}^{\varepsilon\zeta} w_z\,\mathrm{d}z = \int_{-D}^{\varepsilon\zeta} u_z\,\mathrm{d}z + w\big|_{z=\varepsilon\zeta} - w\big|_{z=D}.$$

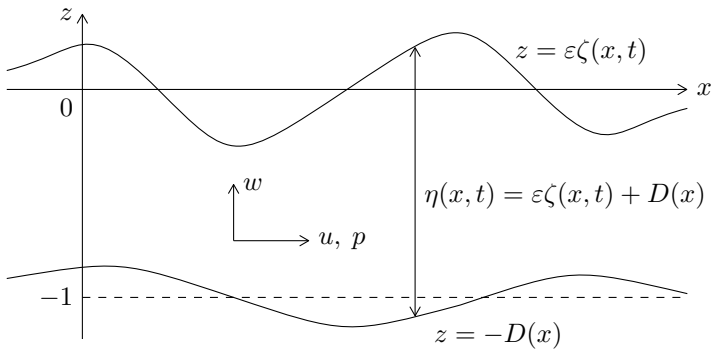


Figure 1.1′: Notation for 2D Euler equations in nondimensional, scaled variables

Let $\overline{u}$ be the depth-averaged u. Then $\eta\overline{u} = \int_{-D}^{\varepsilon\zeta} u\,\mathrm{d}z$. Therefore, by Leibniz's rule for differentiation of integrals,

$$(\eta\overline{u})_x = \int_{-D}^{\varepsilon\zeta} u_x\,\mathrm{d}z + u\big|_{z=\varepsilon\zeta}\varepsilon\zeta_x - u\big|_{z=-D}(-D_x).$$

Therefore, by the two above relations we have

$$\begin{aligned}\eta_t + \varepsilon(\eta\overline{u})_x &= \varepsilon\zeta_t + \varepsilon\int_{-D}^{\varepsilon\zeta} u_x\,\mathrm{d}z + \varepsilon^2\zeta_x u\big|_{z=\varepsilon\zeta} + \varepsilon D_x u\big|_{z=-D}\\ &= \varepsilon\zeta_t - \varepsilon w\big|_{z=\varepsilon\zeta} + \varepsilon w\big|_{z=-D} + \varepsilon^2\zeta_x u\big|_{z=\varepsilon\zeta} + \varepsilon D_x u\big|_{z=-D}\\ &= \varepsilon\left(\zeta_t + \varepsilon\zeta_x u\big|_{z=\varepsilon\zeta} - w\big|_{z=\varepsilon\zeta}\right) + \varepsilon\left(D_x u\big|_{z=-D} + w\big|_{z=-D}\right) = 0,\end{aligned}$$

where in the last equality we used the surface kinematic b.c. (1.5′) and the bottom b.c. (1.7′). We conclude that the continuity equation (1.3′) and the b.c.'s (1.5′) and (1.7′) imply the continuity equation in depth-averaged form:

$$\boxed{\eta_t + \varepsilon(\eta\overline{u})_x = 0} \tag{1.9}$$

which is an *exact* equation, whose terms in the l.h.s. depend on x, and since $D = D(x)$ it is sometimes written as $\zeta_t + \varepsilon(\zeta\overline{u})_x + (D\overline{u})_x = 0$.

(v) Horizontal momentum equation in integrodifferential form with a pressure term

Integrating the horizontal equation (1.1′) w.r.t. z in the interval $[-D, \varepsilon\zeta]$ we obtain

$$\varepsilon\int_{-D}^{\varepsilon\zeta} u_t\,\mathrm{d}z + \varepsilon^2\int_{-D}^{\varepsilon\zeta}(u^2)_x\,\mathrm{d}z + \varepsilon^2\int_{-D}^{\varepsilon\zeta}(wu)_z\,\mathrm{d}z = -\int_{-D}^{\varepsilon\zeta} p_x\,\mathrm{d}z. \tag{1.10}$$

For the third term in the l.h.s. of (1.10) we have

$$\varepsilon^2\int_{-D}^{\varepsilon\zeta}(wu)_z\,\mathrm{d}z = \varepsilon^2(wu)\big|_{z=\varepsilon\zeta} - \varepsilon^2(wu)\big|_{z=-D}. \tag{1.11}$$

For the first term in the l.h.s. of (1.10) we obtain, using Leibniz's rule:

$$\varepsilon\int_{-D}^{\varepsilon\zeta} u_t\,\mathrm{d}z = \varepsilon\frac{d}{dt}\int_{-D}^{\varepsilon\zeta} u\,\mathrm{d}z - \varepsilon^2\zeta_t u\big|_{z=\varepsilon\zeta} = \varepsilon(\eta\overline{u})_t - \varepsilon^2\zeta_t u\big|_{z=\varepsilon\zeta}. \tag{1.12}$$

For the second term in the l.h.s. of (1.10) we get, using again Leibniz's rule

$$\varepsilon^2\int_{-D}^{\varepsilon\zeta}(u^2)_x\,\mathrm{d}z = \varepsilon^2\left(\int_{-D}^{\varepsilon\zeta} u^2\,\mathrm{d}z\right)_x - \varepsilon^3\zeta_x u^2\big|_{z=\varepsilon\zeta} - \varepsilon^2 D_x u^2\big|_{z=-D}. \tag{1.13}$$

Inserting the expressions (1.11)–(1.13) in (1.10) gives

$$\varepsilon(\eta\overline{u})_t + \varepsilon^2\left(\int_{-D}^{\varepsilon\zeta} u^2\,\mathrm{d}z\right)_x + \varepsilon^2(wu)\big|_{z=\varepsilon\zeta} - \varepsilon^2(wu)\big|_{z=-D} \\ - \varepsilon^2\zeta_t u\big|_{z=\varepsilon\zeta} - \varepsilon^3\zeta_x u^2\big|_{z=\varepsilon\zeta} - \varepsilon^2 D_x u^2\big|_{z=-D} = \int_{-D}^{\varepsilon\zeta} p_x\,\mathrm{d}z.$$

Note that the boundary terms in the above vanish (due to the b.c. (1.5′), (1.7′)), since they may be written as

$$\varepsilon^2 u\big|_{z=\varepsilon\zeta}\left(-\zeta_t - \varepsilon\zeta_x u\big|_{z=\varepsilon\zeta} + w\big|_{z=\varepsilon\zeta}\right) - \varepsilon^2 u\big|_{z=-d}\left(w\big|_{z=-D} + D_x u\big|_{z=-D}\right).$$

Therefore the depth-integrated horizontal momentum equation (1.10) gives, in view of the b.c. (1.5′), (1.7′), that

$$\varepsilon(\eta\overline{u})_t + \varepsilon^2\left(\int_{-D}^{\varepsilon\zeta} u^2\,\mathrm{d}z\right)_x = -\int_{-D}^{\varepsilon\zeta} p_x\,\mathrm{d}z. \tag{1.14}$$

Using the depth-averaged form (1.9) of the continuity equation gives now

$$\varepsilon(\eta\overline{u})_t = \varepsilon\eta_t\overline{u} + \varepsilon\eta\overline{u}_t = -\varepsilon^2(\eta\overline{u})_x\overline{u} + \varepsilon\eta\overline{u}_t = \\ = -\varepsilon^2(\eta\overline{u}^2)_x + \varepsilon^2\eta\bar{u}\bar{u}_x + \varepsilon\eta\overline{u}_t = -\varepsilon^2\left(\int_{-D}^{\varepsilon\zeta}\overline{u}^2\,\mathrm{d}z\right)_x + \varepsilon^2\eta\bar{u}\bar{u}_x + \varepsilon\eta\overline{u}_t,$$

where of course $\overline{u}_x = (\overline{u})_x$, $\overline{u}_t = (\overline{u})_t$. Therefore (1.14) is finally written as

$$\varepsilon\eta\overline{u}_t + \varepsilon^2\eta\bar{u}\bar{u}_x + \varepsilon^2\left(\int_{-D}^{\varepsilon\zeta}(u^2 - \overline{u}^2)\,\mathrm{d}z\right)_x = -\int_{-D}^{\varepsilon\zeta} p_x\,\mathrm{d}z. \tag{1.15}$$

which is an exact integrodifferential equation with a pressure integral term, obtained from (1.10) using (1.9) and the b.c.'s (1.5′), (1.7′).

(v) Elimination of the pressure term in (1.15)

We consider now the vertical momentum equation in (1.2′) that we write as

$$\varepsilon\sigma^2\Gamma = -1 - p_z, \quad -D \le z \le \varepsilon\zeta, \tag{1.16}$$

where

$$\Gamma = \Gamma(x,z,t) = w_t + \varepsilon u w_x + \varepsilon w w_z. \tag{1.17}$$

Integrating (1.16) with respect to the depth variable and using the b.c. (1.6′) for the pressure at the free surface, we have

$$\varepsilon\sigma^2 \int_z^{\varepsilon\zeta} \Gamma(x,z',t)\,\mathrm{d}z' = -\int_z^{\varepsilon\zeta} (1+p_{z'})\,\mathrm{d}z' = -(\varepsilon\zeta - z) + p(x,z,t),$$

i.e. that

$$p(x,z,t) = \varepsilon\zeta - z + \varepsilon\sigma^2 \int_z^{\varepsilon\zeta} \Gamma(x,z',t)\,\mathrm{d}z', \quad -D \le z \le \varepsilon\zeta. \tag{1.18}$$

Therefore the pressure in the waveguide is given by the hydrostatic part $\varepsilon\zeta - z$ plus a term involving the indefinite integral of Γ with respect to z. Note that (1.18) and a little algebra (recalling that $\eta = \varepsilon\zeta + D$) gives for the depth-averaged pressure that

$$\eta\overline{p} = \int_{-D}^{\varepsilon\zeta} p(x,z,t)\,\mathrm{d}z = \frac{\eta^2}{2} + \varepsilon\sigma^2 \int_{-D}^{\varepsilon\zeta} \mathrm{d}z \int_z^{\varepsilon\zeta} \Gamma(x,z',t)\,\mathrm{d}z'. \tag{1.19}$$

(Note that both sides of (1.19) are functions of x and t.) By Leibniz's formula we have for the term in the r.h.s. of (1.15), using (1.6′), and the formula (1.18) for the pressure, that

$$\begin{aligned}
\int_{-D}^{\varepsilon\zeta} p_x\,\mathrm{d}z &= \left(\int_{-D}^{\varepsilon\zeta} p\,\mathrm{d}z\right)_x - D_x p\big|_{z=-D} = (\eta\overline{p})_x - D_x p\big|_{z=-D} \\
&= (\eta\overline{p})_x - D_x\left(\varepsilon\zeta + D + \varepsilon\sigma^2 \int_{-D}^{\varepsilon\zeta} \Gamma(x,z',t)\,\mathrm{d}z'\right) \\
&= (\eta\overline{p})_x - D_x\eta - \varepsilon\sigma^2 D_x \int_{-D}^{\varepsilon\zeta} \Gamma(x,z',t)\,\mathrm{d}z'.
\end{aligned}$$

Hence, from (1.19)

$$\int_{-D}^{\varepsilon\zeta} p_x\,\mathrm{d}z = \eta\eta_x + \varepsilon\sigma^2\partial_x \int_{-D}^{\varepsilon\zeta}\mathrm{d}z \int_z^{\varepsilon\zeta} \Gamma(x,z',t)\,\mathrm{d}z' - D_x\eta - \varepsilon\sigma^2 D_x \int_{-D}^{\varepsilon\zeta} \Gamma(x,z',t)\,\mathrm{d}z'.$$

Substituting this expression in (1.15) we see that

$$\boxed{\begin{aligned}
&\varepsilon\,\eta\,\overline{u}_t + \varepsilon^2\eta\,\overline{u}\,\overline{u}_x + \varepsilon^2\left(\int_{-D}^{\varepsilon\zeta} (u^2 - \overline{u}^2)\,\mathrm{d}z\right)_x + \varepsilon\,\eta\,\zeta_x = \\
&\quad = -\varepsilon\sigma^2\partial_x \int_{-D}^{\varepsilon\zeta}\mathrm{d}z \int_z^{\varepsilon\zeta} \Gamma(x,z',t)\,\mathrm{d}z' + \varepsilon\sigma^2 D_x \int_{-D}^{\varepsilon\zeta} \Gamma(x,z',t)\,\mathrm{d}z'
\end{aligned}} \tag{1.20}$$

The equation (1.20) is an integrodifferential equation, both of the sides of which are functions of x and t. Like (1.9) it is an *exact* equation, i.e. it holds for the solution

of the Euler equations exactly. It was obtained by the depth-integrated horizontal momentum equation (1.15), and has no pressure term at the expense of involving a single and a double integral of Γ and the integral of $u^2 - \overline{u}^2$. Thus, in addition of η (or ζ) and $\overline{u}$ it involves two more unknowns, the components of the pointwise velocity u and w in the fluid domain. In order to obtain a system of two equations of two unknown functions of x and t (we will choose as such the functions η and $\overline{u}$), i.e. a simpler model than the Euler equations that involve four unknown functions and a third independent variable, we need to make hypotheses on the magnitude of ε and/or σ, i.e. select a *scaling regime* to which ε and/or σ will belong.

(vi) Asymptotic expansions of u, w in powers of σ^2 in the scaling regime $\sigma \ll 1$.

The scaling regime that we will consider is $\sigma \ll 1$, i.e. that of long (or "shallow-water") waves. At this stage we make no assumptions about ε, which could be $\mathcal{O}(1)$. Thusfar we have not used the irrotationality condition (1.4′). We note that (1.4′) and (1.3′) form an elliptic system of equations for u and w that could be formally solved if we had one more suitable datum on u or w, say at $z = -D$. So we introduce the (unknown) function $u_b = u_b(x,t) = u\big|_{z=-D}$ and try to express u and w in terms of u_b in expansions of powers of σ.

For this purpose we integrate with respect to z both sides of (1.4′) to get

$$u = u_b + \sigma^2 \int_{-D}^{z} w_x \,\mathrm{d}z', \quad -D \le z \le \varepsilon\zeta. \tag{1.21}$$

Integrating now with respect to z the continuity equation (1.3′) we obtain

$$w = w\big|_{z=-D} - \int_{-D}^{z} u_x \,\mathrm{d}z',$$

and using the bottom b.c. (1.7′) we conclude

$$w = -u_b D_x - \int_{-D}^{z} u_x \,\mathrm{d}z'. \tag{1.22}$$

We now proceed iteratively. From (1.21) we get by Leibniz's rule

$$u_x = u_{b,x} + \sigma^2 \partial_x \int_{-D}^{z} w_x \,\mathrm{d}z' = u_{b,x} + \sigma^2 \int_{-D}^{z} w_{xx} \,\mathrm{d}z' + \sigma^2 D_x w_x\big|_{z=-D}.$$

Then, inserting this in (1.22) we see that

$$w = -u_b D_x - u_{b,x}(z+D) + \mathcal{O}(\sigma^2), \tag{1.23}$$

which, in view of (1.21), yields

$$u = u_b + \sigma^2 \int_{-D}^{z} \left[-(u_b D_x)_x - u_{b,xx}(z'+D) - u_{b,x} D_x\right] \mathrm{d}z' + \mathcal{O}(\sigma^4),$$

i.e.

$$u = u_b - \sigma^2 \left[\big(u_{b,x} D_x + (u_b D_x)_x\big)(z+D) + u_{b,xx} \frac{(z+D)^2}{2} \right] + \mathcal{O}(\sigma^4). \tag{1.24}$$

The expansions (1.23), (1.24) suffice for our purposes. Now, since we want to work with η (or ζ) and $\overline{u}$ as dependent variables, we should express u_b in terms of $\overline{u}$ and then use (1.23) and (1.24) to express u and w in terms of $\overline{u}$.

Since, from (1.24), $u_b = u + \mathcal{O}(\sigma^2)$ and $u_b = u_b(x,t)$, it follows that

$$u_b = \overline{u} + \mathcal{O}(\sigma^2). \tag{1.25}$$

This suffices to give us, in view of (1.23), the required expression for w in terms of $\overline{u}$ with $\mathcal{O}(\sigma^2)$ remainder:

$$w = -\overline{u}D_x - \overline{u}_x(z+D) + \mathcal{O}(\sigma^2). \tag{1.26}$$

Now, from (1.24)

$$u_b = u + \sigma^2\left[\big(u_{b,x}D_x + (u_bD_x)_x\big)(z+D) + u_{b,xx}\frac{(z+D)^2}{2}\right] + \mathcal{O}(\sigma^4),$$

and in view of (1.25)

$$u_b = u + \sigma^2\left[\big(\overline{u}_xD_x + (\overline{u}D_x)\big)(z+D) + \overline{u}_{xx}\frac{(z+D)^2}{2}\right] + \mathcal{O}(\sigma^4).$$

Therefore, since $u_b = u_b(x,t)$, $\overline{u} = \overline{u}(x,t)$, depth-averaging in the above gives

$$u_b = \overline{u} + \sigma^2\left[\big(\overline{u}_xD_x + (\overline{u}D_x)_x\big)\frac{\eta}{2} + \overline{u}_{xx}\frac{\eta^2}{6}\right] + \mathcal{O}(\sigma^4),$$

which is the required expansion of u_b in terms of $\overline{u}$. Inserting this in (1.24) and simplifying the resulting expression, we get

$$u = \overline{u} + \sigma^2\big(\overline{u}_xD_x + (\overline{u}D_x)_x\big)\left(\frac{\eta}{2} - (z+D)\right) + \sigma^2\overline{u}_{xx}\left(\frac{\eta^2}{6} - \frac{(z+D)^2}{2}\right) + \mathcal{O}(\sigma^4), \tag{1.27}$$

which is the desired expansion of u with $\mathcal{O}(\sigma^4)$ remainder with $\overline{u}$ in the coefficients.

(vii) Asymptotic expansions in powers of σ^2 of the three integrals in (1.20)

We now use the expansions (1.23) and (1.27) in the three integral terms in (1.20) with the aim of getting a pde involving only η (or ζ) and $\overline{u}$ upon neglecting $\mathcal{O}(\sigma^4)$ terms.

First, squaring both sides of (1.27) gives

$$u^2 = \overline{u}^2 + \sigma^2\overline{u}\big(\overline{u}_xD_x + (\overline{u}D_x)_x\big)\big(\eta - 2(z+D)\big) + \sigma^2\overline{u}\,\overline{u}_{xx}\left(\frac{\eta^2}{3} - (z+D)^2\right) + \mathcal{O}(\sigma^4).$$

Integrating then the difference $u^2 - \overline{u}^2$ with respect to z on the interval $[-D, \varepsilon\zeta]$ we see that the $\mathcal{O}(\sigma^2)$ terms of the integral vanish. Therefore

$$\int_{-D}^{\varepsilon\zeta}(u^2 - \overline{u}^2)\,\mathrm{d}z = \mathcal{O}(\sigma^4). \tag{1.28}$$

The two integrals involving Γ in (1.20) have $\mathcal{O}(\sigma^2)$ coefficients, so we need their expansions up to $\mathcal{O}(\sigma^2)$ terms. Using the definition of Γ, i.e. equation (1.17), (1.26), and the fact that $u = \overline{u} + \mathcal{O}(\sigma^2)$ (which follows form (1.27)), we obtain, for $-D \le z \le \varepsilon\zeta$

$$\Gamma(x,z,t) = -D_x(\overline{u}_t + \varepsilon\,\overline{u}\,\overline{u}_x) - \overline{u}_{xt}(z+D) \\ - \varepsilon(\overline{u})^2 D_{xx} - \varepsilon(z+D)(\overline{u}\,\overline{u}_{xx} - (\overline{u}_x)^2) + \mathcal{O}(\sigma^2).$$

Therefore integration gives

$$\int_{-D}^{\varepsilon\zeta} \Gamma(x,z',t)\,\mathrm{d}z' = -D_x\eta(\overline{u}_t + \varepsilon\,\overline{u}\,\overline{u}_x) - \varepsilon\eta(\overline{u})^2 D_{xx} \\ - (\overline{u}_{xt} + \varepsilon\,\overline{u}\,\overline{u}_{xx} - \varepsilon(\overline{u}_x)^2)\frac{\eta^2}{2} + \mathcal{O}(\sigma^2). \quad (1.29)$$

Finally, for the double integral we get, after straightforward calculations:

$$\int_{-D}^{\varepsilon\zeta} \mathrm{d}z \int_{z}^{\varepsilon\zeta} \Gamma(x,z',t)\,\mathrm{d}z' = -\left[D_x(\overline{u}_t + \varepsilon\,\overline{u}\,\overline{u}_x) + \varepsilon(\overline{u})^2 D_{xx}\right]\frac{\eta^2}{2} \\ - \left[\overline{u}_{xt} + \varepsilon\,\overline{u}\,\overline{u}_{xx} - \varepsilon(\overline{u}_x)^2\right]\frac{\eta^3}{3} + \mathcal{O}(\sigma^2). \quad (1.30)$$

(viii) Derivation of an $\mathcal{O}(\sigma^4)$ approximation of (1.20)

Inserting now the expressions (1.28)–(1.30) into (1.20) we obtain

$$\varepsilon\eta\overline{u}_t + \varepsilon^2\eta\,\overline{u}\,\overline{u}_x + \varepsilon\eta\zeta_x = \varepsilon\sigma^2\left(A\frac{\eta^2}{2} + B\frac{\eta^3}{3}\right)_x - \varepsilon\sigma^2 D_x\left(A\eta + B\frac{\eta^2}{2}\right) + \mathcal{O}(\sigma^4), \quad (1.31)$$

where

$$A = D_x(\overline{u}_t + \varepsilon\,\overline{u}\,\overline{u}_x) + \varepsilon(\overline{u})^2 D_{xx}, \quad (1.32)$$

$$B = \overline{u}_{xt} + \varepsilon\,\overline{u}\,\overline{u}_{xx} - \varepsilon(\overline{u}_x)^2. \quad (1.33)$$

If we disregard the $\mathcal{O}(\sigma^4)$ term in (1.31) we obtain a pde that involves η (or ζ) and $\overline{u}$, and which, together with the (exact) pde (1.9) gives us the SGN system over variable bottom, which is formally an $\mathcal{O}(\sigma^4)$ approximation to the 2D Euler equations (1.1′)–(1.7′), formally valid if $\sigma \ll 1$.

(ix) The final form of the SGN system.

We simplify now somewhat the SGN system. We will henceforth put $\mu = \sigma^2$ and drop the bar from the depth-averaged horizontal velocity which will be denoted simply by $u = u(x,t)$.

We replace the $\mathcal{O}(\sigma^4)$ terms of (1.31) by zero and (since η will always be positive) we divide both sides of the resulting equation by $\varepsilon\eta$. Straightforward but

long calculations in the r.h.s. of this equation yield the first form of the system in nondimensional, scaled variables

$$\begin{aligned}
&\eta_t + \varepsilon(\eta u)_x = 0,\\
&\left(I + \frac{\mu}{\eta}\widetilde{\mathcal{T}}[\eta, D]\right) u_t + \zeta_x + \varepsilon u u_x\\
&\qquad + \mu\varepsilon\left\{-\frac{1}{3\eta}\left(\eta^3(uu_{xx} - u_x^2)\right)_x + \widetilde{\mathcal{Q}}[\eta, D]u\right\} = 0,
\end{aligned} \tag{SGN$'$}$$

where $\eta = \varepsilon\zeta + D$, and the differential operators $\widetilde{\mathcal{T}}[\eta, D]$ and $\widetilde{\mathcal{Q}}[\eta, D]$ are defined by

$$\begin{aligned}
\widetilde{\mathcal{T}}[\eta, D]w &= -\frac{1}{3}(\eta^3 w_x)_x - \frac{1}{2}(D_x\eta^2 w)_x + \frac{1}{2}D_x\eta^2 w_x + D_x^2\eta w,\\
\widetilde{\mathcal{Q}}[\eta, D]w &= -\frac{1}{2\eta}\left\{\left[(D_x w w_x + D_{xx}w^2)\eta^2\right]_x - D_x(ww_{xx} - w_x^2)\eta^2\right\}\\
&\qquad + D_x^2 w w_x + D_x D_{xx} w^2.
\end{aligned}$$

Alternative derivations of the Serre-Green-Naghdi system have been given in the literature, in one or two space dimensions. For example in [Lan13] and [LB09] the starting point is not the Euler equations written in the 'primitive' variables u, w, p, ζ, i.e. equalities (1.1)–(1.7), but an equivalent system of nonlocal equations (the Zakharov formulation, cf. [Lan13]). In the sequel we will use the SGN system in the final form obtained in [LB09]. To this end, we introduce another scaling parameter β, defined as $\beta = \frac{B}{D_0}$, where B is a characteristic measure of the variation of the bottom, and put $D(x) = 1 - \beta b(x)$, still in scaled, nondimensional variables. A little algebra shows that (SGN$'$) may be written using the new bottom topography function in the form used in [LB09], i.e. as

$$\begin{aligned}
&\eta_t + \varepsilon(\eta u)_x = 0,\\
&\left(I + \frac{\mu}{\eta}T[\eta, \beta b]\right) u_t + \zeta_x + \varepsilon u u_x\\
&\qquad + \mu\varepsilon\left\{-\frac{1}{3\eta}\left(\eta^3(uu_{xx} - u_x^2)\right)_x + Q[\eta, \beta b]u\right\} = 0,
\end{aligned} \tag{SGN}$$

where the variables u, ζ and the constants ε, μ are as before, the water depth is now defined as $\eta = \varepsilon\zeta + 1 - \beta b$, and the operators T and Q, obtained by $\widetilde{\mathcal{T}}$ and $\widetilde{\mathcal{Q}}$ by putting $D = 1 - \beta b(x)$, are given by

$$\begin{aligned}
T[\eta, \beta b]w &= -\frac{1}{3}(\eta^3 w_x)_x + \frac{\beta}{2}\left[(b'\eta^2 w)_x - b'\eta^2 w_x\right] + \beta^2(b')^2\eta w,\\
Q[\eta, \beta b]w &= \frac{\beta}{2\eta}\left\{(b'\eta^2 w w_x)_x + (b''\eta^2 w^2)_x - b'\eta^2(ww_{xx} - w_x^2)\right\}\\
&\qquad + \beta^2(b')^2 w w_x + \beta^2 b' b'' w^2.
\end{aligned}$$

Note that in the derivation of (SGN) or (SGN′) it was tacitly assumed that b (or D) is three times differentiable. Note also that $\widetilde{\mathcal{Q}}$ or Q is a purely bathymetric term, i.e. is zero for horizontal bottoms, while T contains also the dispersive term $-\frac{1}{3}(\eta^3 w_x)_x$.

When the bottom is horizontal, i.e. when $\beta = 0$, $T[\eta, 0]w = -\frac{1}{3}(\eta^3 w_x)_x$, $Q[h, 0] = 0$, and (SGN) becomes (with $\eta = 1 + \varepsilon\zeta$):

$$\eta_t + (\eta u)_x = 0, \tag{1.34}$$

$$\left[1 - \frac{\mu}{3\eta}\partial_x(\eta^3\partial_x)\right] u_t + \zeta_x + \varepsilon u u_x - \frac{\varepsilon\mu}{3\eta}\left[\eta^3(uu_{xx} - (u_x)^2)\right]_x = 0. \tag{1.35}$$

The last equation may be written in the form

$$u_t + \zeta_x + \varepsilon u u_x - \frac{\mu}{3\eta}\left[\eta^3(u_{xt} + \varepsilon u u_{xx} - \varepsilon u_x^2)\right]_x = 0, \tag{1.36}$$

and we see that the system given by (1.34), (1.36) is indeed the classical system of the Serre equations for a horizontal bottom in their scaled form, cf. [Ser53a], [Ser53b].

1.2 Derivation of two ‘classical’ Boussinesq type systems with variable bottom

We now simplify the SGN system when the scaling parameters belong to the *Boussinesq regime*, i.e. when $\varepsilon = \mathcal{O}(\mu)$, $\mu \ll 1$, following [LB09]. We will recover the two ‘classical’ Boussinesq (CB) models with bottom topography variation, called (a) and (b) in [LB09, section III C].

(a) CB system with strongly varying bottom topography, $\beta = \mathcal{O}(1)$ (CBs)

Using $\varepsilon = \mathcal{O}(\mu)$, $\mu \ll 1$, $\beta = \mathcal{O}(1)$ in (SGN) we see that the continuity equation does not change since $\eta = \varepsilon\zeta + 1 - \beta b$. In addition note that

$$\eta = \varepsilon\zeta + 1 - \beta b = (1 - \beta b)\left(1 + \frac{\varepsilon}{1 - \beta b}\right) = \eta_b\left(1 + \frac{\varepsilon}{\eta_b}\right),$$

where $\eta_b = 1 - \beta b$, following now the notation of [LB09], i.e. using η_b instead of $D(x)$. Then

$$\frac{\mu}{\eta} = \frac{\mu}{\eta_b}\left(1 + \frac{\varepsilon}{\eta_b}\right)^{-1} = \frac{\mu}{\eta_b}\left(1 - \frac{\varepsilon}{\eta_b} + \mathcal{O}(\varepsilon^2)\right) = \frac{\mu}{\eta_b} + \mathcal{O}(\mu^2). \tag{1.37}$$

Now, since $\eta = \eta_b + \mathcal{O}(\mu)$, we have by (1.37) and the definition of $T[\eta, \beta b]$ that

$$\begin{aligned}
&\frac{\mu}{\eta} T[\eta, \beta b]\, w = \frac{\mu}{\eta_b} T[\eta, \beta b]\, w + \mathcal{O}(\mu^2) = \\
&= \frac{\mu}{\eta_b} \Bigg\{ -\frac{1}{3} \left[(\eta_b^3 + \mathcal{O}(\mu)) w_x \right]_x + \frac{\beta}{2} \left\{ \left[(\eta_b^2 + \mathcal{O}(\mu)) b' w \right]_x - (\eta_b^2 + \mathcal{O}(\mu)) b' w_x \right\} \\
&\qquad\qquad + \beta^2 (\eta_b + \mathcal{O}(\mu)) (b')^2 w + \mathcal{O}(\mu^2) \Bigg\} = \\
&= \frac{\mu}{\eta_b} \left\{ -\frac{1}{3} (\eta_b^3 w_x)_x + \frac{\beta}{2} \left[(\eta_b^2 b' w)_x - \eta_b^2 b' w_x \right] + \beta^2 \eta_b (b')^2 w + \mathcal{O}(\mu^2) \right\} \\
&= \frac{\mu}{\eta_b} T[\eta_b, \beta b] + \mathcal{O}(\mu^2).
\end{aligned}$$

Therefore, since $\mu\varepsilon = \mathcal{O}(\mu^2)$, if we ignore $\mathcal{O}(\mu^2)$ terms, the second pde in (SGN) becomes

$$\left(1 + \frac{\mu}{\eta_b} T[\eta_b, \beta b] \right) u_t + \zeta_x + \varepsilon u u_x = 0. \tag{1.38}$$

The system consisting of the continuity equation and (1.38) is precisely system (a) of [LB09, section III C] and will be called (CBs) in the sequel.
Notice that by the definition of T we have

$$T[\eta_b, \beta b]\, w = -\frac{1}{3} (\eta_b^3 w_x)_x + \frac{\beta}{2} \left[(\eta_b^2 b' w)_x - \eta_b^2 b' w_x \right] + \beta^2 \eta_b (b')^2 w. \tag{1.39}$$

Therefore (1.38) may be written as

$$u_t - \frac{\mu}{3\eta_b} (\eta_b^3 u_{xt})_x + \frac{\beta\mu}{2\eta_b} \left[(\eta_b^2 b' u_t)_x - \eta_b^2 \beta' u_{xt} \right] + \beta^2 \mu (b')^2 u_t + \zeta_x + \varepsilon u u_x = 0.$$

Simplifying the above a bit further, we finally see that the system (CBs) is:

$$\begin{cases}
\zeta_t + (\eta u)_x = 0, & (1.40) \\
u_t + \zeta_x + \varepsilon u u_x + \mu \underbrace{\left(\frac{\beta}{2} \eta_b b'' u_t + \beta \eta_b b' u_{xt} - \frac{1}{3} \eta_b^2 u_{xxt} \right)}_{\substack{\text{linear, dispersive terms} \\ \text{depending on bottom topography}}} = 0, & (1.41)
\end{cases}$$

where

$$\eta = \eta_b + \varepsilon\zeta > 0, \quad \eta_b = 1 - \beta b > 0.$$

This system reduces to the scaled nondimensional 'classical' Boussinesq system (CB) in the case of horizontal bottom ($\beta = 0, \eta_b = 1$).

It is easy to see that (1.40)–(1.41) is the usual 'Peregrine' system, [Per67], in its nondimensional, scaled form. The first equation of the Peregrine system is just (1.40). The second equation of the Peregrine system is usually written as

$$u_t + \zeta_x + \varepsilon u u_x + \mu \left(-\frac{\gamma}{2} (\gamma u_t)_{xx} + \frac{\gamma^2}{6} u_{xxt} \right) = 0, \tag{1.42}$$

where γ turns out, cf. [Per67], to be just η_b. Therefore, the dispersive term in (1.42) is μ times

$$\begin{aligned} -\frac{\gamma}{2}(\gamma u_t)_{xx} + \frac{\gamma^2}{6}u_{xxt} &= -\frac{\gamma}{2}\left(\gamma_{xx}u_t + 2\gamma_x u_{xt} + \gamma u_{xxt}\right) + \frac{\gamma^2}{6}u_{xxt} \\ &= -\frac{1}{2}\gamma\gamma_{xx}u_t - \gamma\gamma_x u_{xt} - \frac{1}{3}\gamma^2 u_{xxt}. \end{aligned}$$

Since $\gamma = \eta_b$, $\gamma_x = -\beta b'$, $\gamma_{xx} = -\beta b''$, we see that (1.42) is precisely (1.41).

(b) CB system with weakly varying bottom topography, $\beta = \mathcal{O}(\mu)$ (CBw)

The second system (b) in [LB09] is derived under the additional assumption that $\beta = \mathcal{O}(\mu)$. (Recall that $\mu \ll 1$ and that $\varepsilon = \mathcal{O}(\mu)$.) This means that the new system has small topography variation. Under these assumptions (1.40)–(1.41) become

$$\zeta_t + (\eta u)_x = 0, \tag{1.43}$$

$$u_t + \zeta_x + \varepsilon u u_x - \tfrac{\mu}{3}u_{xxt} = 0, \tag{1.44}$$

since now $\eta_b = 1 + \mathcal{O}(\mu)$ and the right-hand side of (1.44) consists of $\mathcal{O}(\mu^2)$ terms that are ignored. In the new system, the bottom topography appears only in (1.43), since $\eta = 1 + \varepsilon\zeta - \beta b$. Hence, the system may be written as

$$\begin{cases} \zeta_t + u_x + \varepsilon(\zeta u)_x - \beta(bu)_x = 0, \\ u_t + \zeta_x + \varepsilon u u_x - \dfrac{\mu}{3}u_{xxt} = 0, \end{cases} \tag{CBw}$$

which is just the CB system with a $\mathcal{O}(\mu)$ perturbation due to the bottom topography in the first pde. We will consider numerical methods for (CBs) and (CBw) in Chapter 2.

Needless to say, when we omit the dispersive terms in (CBw), i.e. put $\mu = 0$ in the second equation and let $\varepsilon, \beta = \mathcal{O}(1)$, we obtain the Shallow Water equations with variable bottom. Their numerical solution will be the object of Chapters 3 and 4 of this thesis. We will discuss issues of validity and rigorous existence-uniqueness theory of the ivp's for those systems at the appropriate places in Chapters 2 and 3.

Chapter 2

Standard Galerkin Finite Element methods for the numerical solution of two classical-Boussinesq type systems over variable bottom

2.1 Introduction

The 'Classical' Boussinesq system, [Whi74], in one spatial dimension is the nonlinear, dispersive system of pde's

$$\begin{aligned} &\zeta_t + u_x + \varepsilon(\zeta u)_x = 0, \\ &u_t + \zeta_x + \varepsilon u u_x - \frac{\mu}{3} u_{xxt} = 0. \end{aligned} \tag{CB}$$

and corresponds to $\beta = 0$, i.e. horizontal bottom in (CBw) of Chapter 1. As mentioned in Ch. 1, the variables in (CB) are nondimensional and scaled. We recall that the scaling parameters ε, μ are related under the Boussinesq hypothesis that $\varepsilon \ll 1$, $\mu \ll 1$, and that $\varepsilon = \mathcal{O}(\mu)$. We recall that the first pde in (CB) is exact while the second is an $\mathcal{O}(\varepsilon^2)$ approximation to a relation obtained from the Euler equations. We recall that in the variables of (CB), the horizontal bottom lies at $z = -1$ so that the water depth is given by $\eta = 1 + \varepsilon\zeta(x,t)$.

The initial-value problem for (CB) with initial data $\zeta(x,0) = \zeta_0(x)$, $u(x,0) = u_0(x)$ on the real line has been studied by Schonbek [Sch81] and Amick [Ami84], who established global existence and uniqueness of smooth solutions under the assumption that $1 + \varepsilon \inf_x \zeta_0(x) > 0$. One conclusion of this theory is that for all $t \geq 0$, $1 + \varepsilon \inf_x \zeta(x,t) > 0$, i.e. that there is always water in the channel. Existence-uniqueness of solutions globally in time in Sobolev spaces were established in [BCS04]. The initial-boundary-value problem (ibvp) for (CB) posed on a

finite interval, say $[0,1]$, with zero boundary conditions for u at $x=0$ and $x=1$, and no boundary conditions for ζ, was proved in [Ada11] to possess global weak (distributional) solutions.

The system (CB) has been used and solved numerically extensively in the engineering literature. We will refer here just to [AD13] and [AD12] for error estimates of Galerkin-finite element methods for the ibvp for (CB) mentioned above and a computational study of the properties of the solitary-wave solutions of the system. For the numerical analysis of the periodic ivp we refer to [ADM10].

In this chapter we will be interested in the numerical solution of extensions of (CB) valid in channels of variable-bottom topography. Several such extensions have been derived in the literature. As mentioned in Chapter 1, we will follow [LB09] and consider two specific such variable-bottom models that may be derived from the Serre-Green-Naghdi (SGN) system of equations. Their formal derivation was done in Chapter 1. For their rigorous theory of validity we refer to [LB09] and [Lan13] and their references.

Recalling from Chapter 1 the Serre-Green-Naghdi equations (SGN), we mention that the ivp of the system has been analyzed in some generality in [Lan13]. For initial positive depth, the depth remains positive while solution of (SGN) exist. Also, in the case of the 1d (SGN) a local in time theory of existence and uniqueness of solutions of the ivp with energy methods has been given by Israwi, [Isr11].

The model (SGN) has been used in many computational studies of long-surface wave propagation over uneven bottoms. We refer, for example, to [Bar04], [CBB07] and [Bon+11] and their references for computations with finite differences and finite volume methods, and to [MSM17] for a finite element scheme. An error analysis of the Galerkin-finite element method in the case of a horizontal bottom (i.e. when $\beta=0$), appears in [ADM17] in the case of periodic ivp.

As was mentioned previously, our aim in this chapter is to consider two simplifications of (SGN) that are variable-bottom extensions of (CB), namely the systems (CBs) and (CBw). For the convenience of the reader we rewrite the system here: the *'Classical' Boussinesq system with strongly varying bottom topography*, abbreviated as (CBs), is given by the system

$$\begin{aligned} &\zeta_t + (\eta u)_x = 0, \\ &\left(1 + \frac{\mu}{\eta_b}\mathcal{T}[\eta_b, \beta b]\right) u_t + \zeta_x + \varepsilon u u_x = 0, \end{aligned} \tag{CBs}$$

where $\eta = \eta_b + \varepsilon\zeta > 0$, $\eta_b = 1 - \beta b > 0$, $\varepsilon = \mathcal{O}(\mu) \ll 1$, and $\mathcal{T}[\eta_b, \beta b]w$ is given by its expression in (SGN) when we replace η by η_b. It was mentioned in Ch. 1 that (CBs) coincides with the system that was first derived from the Euler equations by Peregrine in [Per67]; it is usually called the 'Peregrine system' in the literature and has been used widely in coastal dynamics computations. We will refer to several computational studies with (CBs) in Section 2.3 in the sequel. We recall that if we assume in (CBs) following [LB09] that $\beta = \mathcal{O}(\varepsilon)$, i.e. that the variation of bottom is small and specifically of the order ε of the nonlinear and dispersion

terms in (CBs), we obtain a system that we call the *'Classical' Boussinesq system with weakly varying bottom topography*, (CBw) which is given by

$$\begin{aligned} &\zeta_t + (\eta u)_x = 0, \\ &u_t + \zeta_x + \varepsilon u u_x - \frac{\mu}{3} u_{xxt} = 0, \end{aligned} \tag{CBw}$$

where of course we still assume that $\varepsilon = \mathcal{O}(\mu)$, $\mu \ll 1$. The dependence on the bottom topography occurs now explicitly (but weakly) through the first equation, since $\eta = \eta_b + \varepsilon\zeta = 1 - \beta b + \varepsilon\zeta$ with $\beta = \mathcal{O}(\varepsilon)$. This system has also been used widely in computations in the engineering literature.

The theory of existence and uniqueness of solutions, at least locally in time, for the ivp for (CBs) may be easily inferred from the analogous theory of (SGN), cf. e.g. [Isr11], while that of (CBw) is practically the same as the one for (CB) plus a 'source'-type linear term of the form $-\beta(bu)_x$ in the left-hand side of the first equation.

In this chapter we will discretize in space ibvp's for the systems (CBs) and (CBw), with zero b.c. for u at the endpoints of $[0,1]$ and no b.c. for ζ, by the standard Galerkin-finite element method on a quasiuniform mesh and prove L^2-error estimates in Section 2.2 for the resulting semidiscretizations. Under certain standard assumptions on the finite element spaces we will prove error estimates of the form

$$\|\zeta - \zeta_h\| + \|u - u_h\|_1 \leq Ch^{r-1},$$

where ζ_h, u_h are the semidiscrete approximations of ζ and u, respectively, $h = \max_i h_i$, and $r - 1 \geq 2$ is the degree of the piecewise polynomials in the finite element space. ($\|\cdot\|$ and $\|\cdot\|_1$ denote, respectively the L^2 and H^1 norms of functions on $[0,1]$.) This type of error estimate is of the same type as the one proved in [AD13] for the analogous ibvp for (CB) in the case of a quasiuniform mesh.

In Section 2.3 we show the results of several numerical experiments that we performed with both systems using a fully discrete scheme with the above spatial discretization and using as a time marching scheme the classical, 4$^{\text{th}}$ order, 4-stage Rugne-Kutta method. The resulting schemes are stable under a mild Courant number restriction and highly accurate. In Section 2.3.1 we check that the schemes also work for piecewise linear continuous functions (i.e. for $r = 2$ and are of optimal order in L^2 for both u and ζ in the case of uniform mesh. In Section 2.3.2 we discuss the application of simple, approximate, absorbing boundary conditions for the systems as an alternative to the reflection b.c. $u = 0$ at the endpoints. In Section 2.3.3 we perform a series of numerical experiments aimed at describing in detail the changes that solitary waves undergo when evolving under (CBs) or (CBw) in a variety of variable-bottom environments. We assess the efficacy of these systems in approximating these flows by comparing them with each other and with the (SGN) system and available experimental data.

In the sequel, we denote, for integer $k \geq 0$, $C^k = C^k[0,1]$ the spaces of k-times continuously differentiable functions on $[0,1]$ and by $H^k = H^k(0,1)$ the

usual L^2-based Sobolev spaces of functions on $(0,1)$. $\overset{0}{H}{}^1$ will denote the elements of H^1 which vanish at $x=0$ and $x=1$. The inner product in $L^2=L^2(0,1)$ will be denoted by $(\cdot,\cdot)$, its norm by $\|\cdot\|$, and the norm on H^k by $\|\cdot\|_k$. The norms on W^k_∞ and L^∞ on $(0,1)$ are denoted by $\|\cdot\|_{k,\infty}$ and $\|\cdot\|_\infty$, respectively. $\mathbb{P}_r$ are the polynomials of degree at most r.

2.2 Error analysis of the Galekin semidiscretization

2.2.1 The finite element spaces

Let $0\le x_1<x_2<\ldots<x_{N+1}=1$ be a quasiuniform partition of $[0,1]$ with $h:=\max_i(x_{i+1}-x_i)$. For integers $r\ge 2$ and $0\le k\le r-2$ we consider the finite element space $S_h=\{\phi\in C^k:\phi\big|_{[x_i,x_{i+1}]}\in\mathbb{P}_{r-1}\}$ and $S_{h,0}=\{\phi\in S_h:\phi(0)=\phi(1)=0\}$. It is well known, see [Cia78], that if $w\in H^r$ there exists $\chi\in S_h$ such that

$$\|w-\chi\|+h\|w'-\chi'\|\le Ch^r\|w\|_r \tag{2.1}$$

for some constant C independent of h and w, and that a similar property holds in $S_{h,0}$ provided $w\in H^r\cap H^1_0$. In addition, if P is the L^2 prejection operator onto S_h, then it holds, cf. [DDW75], that

$$\|\mathrm{P}\,v\|_\infty\le C\|v\|_\infty,\quad \forall v\in L^\infty, \tag{2.2a}$$
$$\|\mathrm{P}\,v-v\|_\infty\le Ch^r\|v\|_{r,\infty},\quad \forall v\in C^r. \tag{2.2b}$$

Due to the quasiuniformity of the mesh, cf. [Cia78], the inverse inequalities

$$\|\chi\|_1\le Ch^{-1}\|\chi\|,\quad \|\chi\|_\infty\le Ch^{-1/2}\|\chi\| \tag{2.3}$$

are valid for $\chi\in S_h$ (or $\chi\in S_{h,0}$).

2.2.2 Semidiscretization in the case of a strongly varying bottom

Using the notation of Chapter 1 we consider the following initial-boundary-value problem (ibvp) for (CBs). For $T>0$ we seek $\zeta=\zeta(x,t)$, $u=u(x,t)$, for $(x,t)\in[0,1]\times[0,T]$, such that

$$\begin{aligned}
&\zeta_t+(\eta u)_x=0,\\
&\left(1+\frac{\mu}{\eta_b}\mathcal{T}[\eta_b,\beta b]\right)u_t+\zeta_x+\varepsilon uu_x=0, && 0\le x\le 1,\ \ 0\le t\le T,\\
&\zeta(x,0)=\zeta_0(x),\ \ u(x,0)=u_0(x), && 0\le x\le 1,\\
&u(0,t)=u(1,t)=0, && 0\le t\le T,
\end{aligned} \tag{2.4}$$

where

$$\eta=\varepsilon\zeta+\eta_b>0,\quad \eta_b(x)=1-\beta b(x)>0,$$

ε, μ, β, are positive constants with $\varepsilon = \mathcal{O}(\mu)$, $\mu \ll 1$, $\beta = \mathcal{O}(1)$, and $b \in C^2[0,1]$. The operator $\mathcal{T}[\eta_b, \beta b]$ is defined as in Chapter 1 by

$$\mathcal{T}[\eta_b, \beta b]w = -\frac{1}{3}(\eta_b^3 w_x)_x + \frac{\beta}{2}[(\eta_b^2 b' w)_x - \eta_b^2 b' w_x] + \beta^2 \eta_b (b')^2 w.$$

All the variables above are nondimensional and scaled. We will assume that the ibvp (2.4) has a unique solution that is smooth enough for the purposes of the error estimates to follow. Taking into account that

$$\mathcal{T}[\eta_b, \beta b]w = -\frac{1}{3}(\eta_b^3 w_x)_x + \frac{\beta}{2}(\eta_b^2 b')' w + \beta^2 \eta_b (b')^2 w,$$

and that $\eta_b' = -\beta b'$, we have

$$\mathcal{T}[\eta_b, \beta b]w = -\frac{1}{3}(\eta_b^3 w_x)_x - \frac{1}{2}\eta_b^2 \eta_b'' w. \tag{2.5}$$

Using in first equation of (2.4) the definition of η, multiplying the second equation by η_b, and taking into account (2.5), we rewrite the ibvp (2.4) in the form

$$\begin{aligned}
&\zeta_t + \varepsilon(\zeta u)_x + (\eta_b u)_x = 0, \\
&\left(\eta_b - \frac{\mu}{2}\eta_b^2 \eta_b''\right) u_t - \frac{\mu}{3}(\eta_b^3 u_{tx})_x + \eta_b \zeta_x + \varepsilon \eta_b u u_x = 0, \qquad (x,t) \in [0,1] \times [0,T], \\
&\zeta(x,0) = \zeta_0(x), \;\; u(x,0) = u_0(x), \quad x \in [0,1], \\
&u(0,t) = u(1,t) = 0, \quad t \in [0,T].
\end{aligned} \tag{2.6}$$

We assume that there are positive constants c_1 and c_2 such that

$$\eta_b(x) \geq c_1, \tag{2.7a}$$

$$\eta_b(x) - \frac{\mu}{2}\eta_b^2(x)\eta_b''(x) \geq c_2, \tag{2.7b}$$

for all $x \in [0,1]$. Since η_b and its derivatives are $\mathcal{O}(1)$, (2.7b) holds for μ sufficiently small. We also consider the bilinear form $A : H_0^1 \times H_0^1 \to \mathbb{R}$ defined by

$$A(v,w) = \left((\eta_b - \frac{\mu}{2}\eta_b^2 \eta_b'')v, w\right) + \frac{\mu}{3}(\eta_b^3 v', w'), \tag{2.8}$$

which is symmetric, bounded on $H^1 \times H^1$, and, because of (2.7), coercive, with

$$A(v,v) \geq c_2 \|v\|^2 + \frac{\mu c_1^3}{3}\|v'\|^2 \geq c_\mu \|v\|_1^2, \quad \forall v \in H^1, \tag{2.9}$$

where $c_\mu := \min(c_2, \mu c_1^3/3)$. Consider now a weighted H^1 ('elliptic') projection associated with the bilinear form (2.9) as the map $\mathrm{R}_h : \overset{0}{H}{}^1 \to S_{h,0}$ defined by

$$A(\mathrm{R}_h v, \chi) = A(v, \chi), \quad \forall \chi \in S_{h,0}, \tag{2.10}$$

for which, cf. e.g. [DDW75], it holds that

$$\|\,\mathrm{R}_h\,v - v\| + h\|\,\mathrm{R}_h\,v - v\|_1 \le Ch^r\|v\|_r, \quad \text{if } v \in H^r \cap H_0^1, \tag{2.11}$$

$$\|\,\mathrm{R}_h\,v - v\|_\infty \le Ch^r\|v\|_{r,\infty}, \quad \text{if } v \in W^{r,\infty} \cap H_0^1. \tag{2.12}$$

We now define the standard Galerkin finite element semidiscretization of the ibvp (2.6). We seek $\zeta_h : [0,T] \to S_h$, $u_h : [0,T] \to S_{h,0}$ such that

$$(\zeta_{ht},\phi) + \varepsilon\big((\zeta_h u_h,\phi)_x,\phi\big) + \big((\eta_b u_h)_x,\phi\big) = 0, \quad \forall \phi \in S_h, \quad 0 \le t \le T, \tag{2.13}$$

$$A(u_{ht},\chi) + (\eta_b \zeta_{hx},\chi) + \varepsilon(\eta_b u_h u_{hx},\chi) = 0, \quad \forall \chi \in S_{h,0}, \tag{2.14}$$

with initial conditions

$$\zeta_h(0) = \mathrm{P}\,\zeta_0, \quad u_h(0) = \mathrm{R}_h\,u_0. \tag{2.15}$$

The ode ivp given by (2.13)–(2.15) has a unique solution locally in time. As part of Theorem 2.1 below we will prove that for sufficiently small h, its solution may be extended up to $t = T$.

Theorem 2.1. *Suppose that the solution* (ζ, u) *of* (2.6) *is sufficiently smooth and that the conditions* (2.7) *hold. Then, if* h *sufficiently small, there exists a constant* C *independent of* h *such that the semidiscrete problem* (2.13)–(2.15) *has a unique solution* (ζ_h, u_h) *for* $0 \le t \le T$*, that satisfies*

$$\max_{0\le t\le T} (\|\zeta(t) - \zeta_h(t)\| + \|u(t) - u_h(t)\|_1) \le Ch^{r-1}. \tag{2.16}$$

Proof. Let $\rho = \zeta - \mathrm{P}\,\zeta$, $\theta = \mathrm{P}\,\zeta - \zeta_h$, $\sigma = u - \mathrm{R}_h\,u$, $\xi = \mathrm{R}_h\,u - u_h$. From (2.6) and (2.13)–(2.15) we get

$$(\theta_t,\phi) + \varepsilon\big((\zeta u - \zeta_h u_h)_x,\phi\big) + \big((\eta_b\sigma + \eta_b\xi)_x,\phi\big) = 0, \quad \forall \phi \in S_h, \tag{2.17}$$

$$A(\xi_t,\chi) + \big(\eta_b(\rho_x + \theta_x),\chi\big) + \varepsilon\big(\eta_b(uu_x - u_h u_{hx}),\chi\big) = 0, \quad \forall \chi \in S_{h,0}, \tag{2.18}$$

that are valid while the semidiscrete problem has a unique solution. For the nonlinear terms we have

$$\begin{aligned} \zeta u - \zeta_h u_h &= \zeta(\sigma+\xi) + u(\rho+\theta) - (\rho+\theta)(\sigma+\xi), \\ uu_x - u_h u_{hx} &= (u\sigma)_x + (u\xi)_x - (\sigma\xi)_x - \sigma\sigma_x - \xi\xi_x. \end{aligned}$$

Let now $t_h \in (0,T]$ be the maximal temporal instance for which the solution of (2.6) exists and it holds that $\|\theta(t)\|_\infty + \|\xi(t)\|_\infty \le 1$, for $t \le t_h$. Putting $\phi = \theta$ in (2.17), using (2.1), (2.2b), (2.11), (2.12), (2.3), and integrating by parts we have

for $t \le t_h$

$$\begin{aligned}
\tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}\|\theta\|^2 =& -\varepsilon\big((\zeta\sigma)_x,\theta\big)-\varepsilon\big((\zeta\xi)_x,\theta\big)-\varepsilon\big((u\rho)_x,\theta\big)-\varepsilon\big((u\theta)_x,\theta\big)+\varepsilon\big((\rho\sigma)_x,\theta\big)\\
&+\varepsilon\big((\rho\xi)_x,\theta\big)+\varepsilon\big((\sigma\theta)_x,\theta\big)+\varepsilon\big((\theta\xi)_x,\theta\big)-\big((\eta_b\sigma)_x,\theta\big)-\big((\eta_b\xi)_x,\theta\big)\\
\le&\,\varepsilon(\|\zeta_x\|_\infty\|\sigma\|+\|\zeta\|_\infty\|\sigma_x\|)\|\theta\|+\varepsilon(\|\zeta_x\|_\infty\|\xi\|+\|\zeta\|_\infty\|\xi_x\|)\|\theta\|\\
&+\varepsilon(\|u_x\|_\infty\|\rho\|+\|u\|_\infty\|\rho_x\|)\|\theta\|+\tfrac{\varepsilon}{2}\|u_x\|_\infty\|\theta\|^2\\
&+\varepsilon(\|\sigma\|_\infty\|\rho_x\|+\|\rho\|_\infty\|\sigma_x\|)\|\theta\|+\varepsilon(\|\xi\|_\infty\|\rho_x\|+\|\rho\|_\infty\|\xi_x\|)\|\theta\|\\
&+\tfrac{\varepsilon}{2}\|\theta\|_\infty\|\sigma_x\|\|\theta\|+\tfrac{\varepsilon}{2}\|\theta\|_\infty\|\xi_x\|\|\theta\|+(\|\eta_\beta'\|_\infty\|\sigma\|+\|\eta_b\|_\infty\|\sigma_x\|)\|\theta\|\\
&+(\|\eta_\beta'\|_\infty\|\xi\|+\|\eta_b\|_\infty\|\xi_x\|)\|\theta\|\\
\le&\,C(h^{r-1}+\|\xi\|_1+\|\theta\|)\|\theta\|,
\end{aligned}\tag{2.19}$$

for some constant C independent of h.

In addition, with $\chi=\xi$ in (2.18) we obtain for $t\le t_h$

$$\begin{aligned}
\tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}A(\xi,\xi) =& -(\eta_b\rho_x+\eta_b\theta_x,\xi)-\varepsilon(\eta_b(u\sigma)_x,\xi)-\varepsilon(\eta_b(u\xi)_x,\xi)+\varepsilon(\eta_b(u\xi)_x,\xi)\\
&+\varepsilon(\eta_b\sigma\sigma_x,\xi)+\varepsilon(\eta_b\xi\xi_x,\xi)\\
=&(\rho,\eta_b'\xi+\eta_b\xi_x)+(\theta,\eta_b'\xi+\eta_b\xi_x)+\varepsilon(u\sigma,\eta_b'\xi+\eta_b\xi_x)-\varepsilon(\eta_b(u\xi)_x,\xi)\\
&-\varepsilon(\sigma\xi,\eta_b'\xi+\eta_b\xi_x)-\tfrac{\varepsilon}{2}(\sigma^2,\eta_b'\xi+\eta_b\xi_x)-\tfrac{\varepsilon}{3}\eta_b'\xi^2,\xi).
\end{aligned}$$

With estimates analogous to those used in (2.19) we get

$$\begin{aligned}
\tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}A(\xi,\xi) \le&\,(\|\eta_b'\|_\infty\|\xi\|+\|\eta_b\|_\infty\|\xi_x\|)(\|\rho\|+\|\theta\|)+\varepsilon\|u\eta_b'\|_\infty\|\sigma\|\|\xi\|\\
&+\varepsilon\|u\eta_b\|_\infty\|\sigma\|\|\xi_x\|+\varepsilon\|\eta_b u_x\|_\infty\|\xi\|^2+\varepsilon\|\eta_b u\|_\infty\|\xi_x\|\|\xi\|\\
&+\varepsilon\|\sigma\eta_b'\|_\infty\|\xi\|^2+\varepsilon\|\sigma\eta_b\|_\infty\|\xi\|\|\xi_x\|+\tfrac{\varepsilon}{2}\|\sigma\eta_b'\|_\infty\|\sigma\|\|\xi\|\\
&+\tfrac{\varepsilon}{2}\|\sigma\eta_b\|_\infty\|\sigma\|\|\xi_x\|+\tfrac{\varepsilon}{3}\|\eta_b'\|_\infty\|\xi\|_\infty\|\xi\|^2\\
\le&\,C(h^r+\|\xi\|_1+\|\theta\|)\|\xi\|_1,
\end{aligned}\tag{2.20}$$

where C is independent of h. From (2.19) and (2.20) we see that

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\big(\|\theta\|^2+A(\xi,\xi)\big)\le C_1h^{2r-2}+C_2\left(\|\theta\|^2+\|\xi\|_1^2\right),$$

where C_1, C_2 are independent of h. From this inequality and (2.9) it follows that

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\big(\|\theta\|^2+A(\xi,\xi)\big)\le C_1h^{2r-2}+C_\mu\big(\|\theta\|^2+A(\xi,\xi)\big),$$

for $t\le t_h$, where $C_\mu=C_2\max(1,1/c_\mu)$. Using Gronwall's lemma in the above we obtain for $t\le t_h$,

$$\|\theta(t)\|^2+A(\xi(t),\xi(t))\le \mathrm{e}^{C_\mu T}\big(\|\theta(0)\|^2+A(\xi(0),\xi(0))\big)+\tfrac{C_1}{C_\mu}\mathrm{e}^{C_\mu T}h^{2r-2},$$

from which, in view of (2.9) and since $\theta(0)=\xi(0)=0$, we see that

$$\|\theta(t)\|+\|\xi(t)\|_1\le\left(\frac{2C_1}{C_\mu\widetilde{C}_\mu}\mathrm{e}^{C_\mu T}\right)^{1/2}h^{r-1},\tag{2.21}$$

for $t \le t_h$, where $\widetilde{C}_\mu = \min(1, c_\mu)$. Now, since (2.3) gives $\|\theta\|_\infty \le Ch^{-1/2}\|\theta\|$ and $\|\xi\|_\infty \le \|\xi\|_1$, if h is taken sufficiently small, we have that $\|\theta\|_\infty + \|\xi\|_\infty < 1$ for $0 \le t \le t_h$, and therefore we may take $t_h = T$. The result follows from (2.20) and the approximation properties of the finite element spaces. □

As suggested by numerical experiments for the (CB) on a *horizontal bottom*, shown in [AD13], the convergence rates in the error estimate (2.16) are sharp in the case of a horizontal bottom; they are sharp in the case of variable-bottom models as well. The H^1 convergence rate of the error of u_h is optimal, while the L^2 rate for η_h suboptimal, as expected, since the first pde in (2.4) is of hyperbolic type and we are using the standard Galerkin method on a nonuniform mesh. (For $r = 2$ the numerical experiments in [AD13] also suggest the improved estimate $\|u - u_h\| = \mathcal{O}(h^2)$.) In the case of *uniform mesh*, better results were proved in [AD13] in the case of horizontal bottom. The numerical experiments in Section 2.3 in the sequel suggest that such improved rates of convergence for uniform mesh persist in the presence of a variable bottom as well.

2.2.3 Semidiscretization in the case of a weakly varying bottom

In the case of a weakly varying bottom, following the remarks in Chapter 1, we consider the following ibvp for the system (CBw). For $T > 0$ we seek $\zeta = \zeta(x,t)$, $u = u(x,t)$, for $(x,t) \in [0,1] \times [0,T]$, such that

$$\begin{aligned} &\zeta_t + (\eta u)_x = 0, \\ &u_t + \zeta_x + \varepsilon u u_x - \frac{\mu}{3} u_{xxt} = 0, \qquad 0 \le x \le 1, \ \ 0 \le t \le T, \\ &\zeta(x,0) = \zeta_0(x), \ \ u(x,0) = u_0(x), \quad 0 \le x \le 1, \\ &u(0,t) = u(1,t), \quad 0 \le t \le T, \end{aligned} \tag{2.22}$$

where

$$\eta = \varepsilon\zeta + \eta_b > 0, \quad \eta_b(x) = 1 - \beta b(x) > 0,$$

and ε, μ, β, are positive constants with $\varepsilon = \mathcal{O}(\mu)$, $\beta = \mathcal{O}(\mu)$, $\mu \ll 1$, and $b = C^2[0,1]$. All the variables above are nondimensional and scaled. We assume that (2.22) has a unique solution, smooth enough for the purposes of the error estimate below.

Let $a : H_0^1 \times H_0^1 \to \mathbb{R}$ denote the weighted H^1-inner product defined by $a(v,w) = (v,w) + \frac{\mu}{3}(v', w')$ and consider the weighted H^1 ('elliptic') projection associated with $a(\cdot,\cdot)$, defined as the map $\widetilde{\mathrm{R}}_h : \overset{0}{H}{}^1 \to S_{h,0}$ such that

$$a(\widetilde{\mathrm{R}}_h v, \chi) = a(v, \chi), \quad \forall \chi \in S_h. \tag{2.23}$$

Obviously, $\widetilde{\mathrm{R}}_h$ satisfies the properties (2.11) and (2.12).

The standard Galerkin finite element semidiscretization of the ibvp (2.22) is the following. We seek $\zeta_h : [0,T] \to S_h$, $u_h : [0,T] \to S_{h,0}$, such that

$$(\zeta_{ht}, \phi) + \varepsilon\big((\zeta_h u_h)_x, \phi\big) + \big((\eta_b u_h)_x, \phi\big) = 0, \quad \forall \phi \in S_h, \tag{2.24}$$

$$a(u_{ht}, \chi) + (\zeta_{hx}, \chi) + \varepsilon(u_h u_{hx}, \chi) = 0, \quad \forall \chi \in S_{h,0}, \tag{2.25}$$

with initial conditions

$$\zeta_h(0) = \mathrm{P}\,\zeta_0, \quad u_h(0) = \widetilde{\mathrm{R}}_h u_0. \tag{2.26}$$

In analogy with Theorem 2.1, the following error estimate holds for the semidiscrete scheme (2.24)–(2.26).

Theorem 2.2. *Suppose that the solution (ζ, u) of* (2.22) *is sufficiently smooth. Then, if h is sufficiently small, there exists a constant C independent of h such that the semidiscrete problem* (2.24)–(2.26) *has a unique solution (ζ_h, u_h) for $0 \le t \le T$, that satisfies*

$$\max_{0\le t\le T} \left(\|\zeta(t) - \zeta_h(t)\| + \|u(t) - u_h(t)\|_1\right) \le C h^{r-1}. \tag{2.27}$$

Proof. Let $\rho = \zeta - \mathrm{P}\,\zeta$, $\theta = \mathrm{P}\,\zeta - \zeta_h$, $\sigma = u - \widetilde{\mathrm{R}}_h u$, $\xi = \widetilde{\mathrm{R}}_h u - u_h$. Using (CB) and (2.24), (2.25), we have

$$(\theta_t, \phi) + \varepsilon\big((\zeta u - \zeta_h u_h)_x, \phi\big) + \big((h_b\sigma + h_b\xi)_x, \phi\big) = 0, \quad \forall \phi \in S_h, \tag{2.28}$$

$$a(\xi_t, \chi) + (\rho_x + \theta_x, \chi) + \varepsilon(u u_x - u_h u_{hx}, \chi) = 0, \quad \forall \chi \in S_{h,0}, \tag{2.29}$$

until the time for which the semidiscrete problem has unique solution. Is holds that

$$\begin{aligned}
\zeta u - \zeta_u u_h &= \zeta(\sigma + \xi) + u(\rho + \theta) - (\rho + \theta)(\sigma + \xi),\\
u u_x + u_h u_{hx} &= (u\sigma)_x + (u_\xi)_x - (\sigma\xi)_x - \sigma\sigma_x - \xi\xi_x.
\end{aligned}$$

Let $t_h \le T$ be the largest time for which $\|\theta(t)\|_\infty \le 1$, for $t \le t_h$. Setting $\phi = \theta$ in (2.28), using (2.1), (2.2b), (2.11), (2.12), (2.3), and integrating by parts, we get for $t \le t_h$

$$\begin{aligned}
\tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}\|\theta\|^2 =& -\varepsilon((\zeta\sigma)_x, \theta) - \varepsilon((\zeta\xi)_x\theta) - \varepsilon((u\rho)_x, \theta) - \varepsilon((u\theta)_x, \theta) + \varepsilon((\rho\sigma)_x, \theta)\\
&+ \varepsilon((\rho\xi)_x, \theta) + \varepsilon((\sigma\theta)_x, \theta) + \varepsilon((\theta\xi)_x, \theta) - ((h_b\sigma)_x, \theta) - ((h_b\xi)_x, \theta)\\
\le& \varepsilon(\|\zeta_x\|_\infty\|\sigma\| + \|\zeta\|_\infty\|\sigma_x\|)\|\theta\| + \varepsilon(\|\zeta_x\|_\infty\|\xi\| + \|\zeta\|_\infty\|\xi_x\|)\|\theta\|\\
&+ \varepsilon(\|u_x\|_\infty\|\rho\| + \|u\|_\infty\|\rho_x\|)\|\theta\| + \tfrac{\varepsilon}{2}\|u_x\|_\infty\|\theta\|^2\\
&+ \varepsilon(\|\sigma\|_\infty\|\rho_x\| + \|\rho\|_\infty\|\sigma_x\|)\|\theta\| + \varepsilon(\|\xi\|_\infty\|\rho_x\| + \|\rho\|_\infty\|\xi_x\|)\|\theta\|\\
&+ \tfrac{\varepsilon}{2}\|\theta\|_\infty\|\sigma_x\|\|\theta\| + \tfrac{\varepsilon}{2}\|\theta\|_\infty\|\xi_x\|\|\theta\| + (\|h_b'\|_\infty\|\sigma\| + \|h_b\|_\infty\|\sigma_x\|)\|\theta\|\\
&+ (\|h_b'\|_\infty\|\xi\| + \|h_b\|_\infty\|\xi_x\|)\|\theta\|\\
\le& C(h^{r-1} + \|\xi\|_1 + \|\theta\|)\|\theta\|,
\end{aligned} \tag{2.30}$$

for some constant C which depends on ε, μ, β. Furthermore, setting $\chi = \xi$ in (2.29) we have, for $0 \le t \le t_h$

$$\begin{aligned}\tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}a(\xi,\xi) =& -(\rho_x+\theta_x,\xi) - \varepsilon((u\sigma)_x,\xi) - \varepsilon((u\xi)_x,\xi) + \varepsilon((\sigma\xi)_x,\xi)\\ &+\varepsilon(\sigma\sigma_x,\xi)+\varepsilon(\xi\xi_x,\xi)\\ =&(\rho+\theta,\xi_x)+\varepsilon(u\sigma,\xi_x)-\varepsilon((u\xi)_x,\xi)-\varepsilon(\sigma\xi,\xi_x)-\tfrac{\varepsilon}{2}(\sigma^2,\xi_x)\end{aligned}$$

therefore

$$\begin{aligned}\tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}a(\xi,\xi) \le&(\|\rho\|+\|\theta\|)\|\xi_x\|+\varepsilon\|u\|_\infty\|\sigma\|\|\xi_x\|+\varepsilon\|u_x\|_\infty\|\xi\|^2+\varepsilon\|u\|_\infty\|\xi_x\|\|\xi\|\\ &+\varepsilon\|\sigma\|_\infty\|\xi\|\|\xi_x\|+\tfrac{\varepsilon}{2}\|\sigma\|_\infty\|\sigma\|\|\xi_x\|\\ &\le C(h^r+\|\xi\|+\|\theta\|)\|\xi\|_1,\end{aligned} \tag{2.31}$$

where the constant C depends on ε, μ, β. From (2.30) and (2.31) we can see that

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\big(\|\theta\|^2+a(\xi,\xi)\big) \le C_1h^{2r-2}+C_2(\|\theta\|^2+\|\xi\|_1^2)$$

where the constants C_1, C_2 depend on ε, μ, β. From this relationship, since $a(v,v) = \|v\|^2+\frac{\mu}{3}\|u'\|^2 \ge \frac{\mu}{3}\|v\|_1^2$, we get

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\big(\|\theta\|^2+a(\xi,\xi)\big) \le C_1h^{2r-2}+C_\mu\big(\|\theta\|^2+a(\xi,\xi)\big),$$

for $t \le t_h$, where $C_\mu = C_2\cdot 3/\mu$. From this inequality and Gronwall's lemma we have, for $t \le t_h$,

$$\|\theta(t)\|^2+a(\xi(t),\xi(t)) \le \mathrm{e}^{C_\mu T}\big(\|\theta(0)\|^2+A(\xi(0),\xi(0))\big)+\tfrac{C_1}{C_\mu}\mathrm{e}^{C_\mu T}h^{2r-2},$$

which results, taking into consideration that $a(v,v) \ge \frac{\mu}{3}\|v\|_1^2$ and that $\theta(0) = \xi(0) = 0$,

$$\|\theta(t)\|+\|\xi(t)\|_1 \le \left(\frac{2C_1}{C_\mu\widetilde{C}_\mu}\mathrm{e}^{C_\mu T}\right)^{1/2}h^{r-1} = \left(\frac{2C_1}{C_2}\mathrm{e}^{C_\mu T}\right)^{1/2}h^{r-1}, \tag{2.32}$$

for $t \le t_h$ and $\widetilde{C}_\mu = \mu/3$. Since, from (2.3) $\|\theta\|_\infty \le Ch^{-1/2}\|\theta\|$, if h is small enough, we have $\|\theta\|_\infty < 1$ and consequently we can take $t_h = T$, therefore from the relation (2.32) the proof of the Theorem is complete. □

2.3 Numerical experiments

In this section we present results of numerical experiments that we performed using the two models (CBs) and (CBw) of the classical Bousinesq system with variable bottom. We discretized the two systems in space using the Galerkin finite element method analyzed in the previous section. For the temporal discretization we used the 'classical', explicit, 4-stage, 4$^{\text{th}}$ order Runge-Kutta scheme (RK4). The convergence of this fully discrete scheme was analyzed, in the case of the ibvp for the

(CB) with horizontal bottom and $u = 0$ at the endpoints in [AD13], where it was shown that under a Courant number stability restriction of the form $\frac{k}{h} \leq \alpha$ the scheme is stable, is fourth-order accurate in time, and preserves the spatial order of convergence of the semidiscrete problem; here k denotes the (uniform) time step.

2.3.1 Convergence rates

The spatial convergence rates proved in Theorems 2.1 and 2.2 in the case of a general quasiuniform mesh are sharp as is suggested by numerical experiments (not shown here). In the case of a uniform spatial mesh better convergence rates may be achieved. This was proved in [AD13] for the (CB) (horizontal bottom and $u = 0$ at the endpoints of the spatial interval) in the case of piecewise linear continuous functions ($r = 2$) and cubic splines ($r = 4$). The numerical results to be presented in the sequel suggest that the improved rates persist in the case of a variable bottom as well for both CB models. (We do not show the optimal-order results for the piecewise linear case ($r = 2$) but concentrate instead in the case of cubic splines ($r = 4$).)

The exact solution of the test problem used for the error rate computations is $\zeta(x,t) = \mathrm{e}^{2t}(\cos(\pi x)+x+2)$, $u(x,t) = \mathrm{e}^{xt}(\sin(\pi x)+x^3-x^2)$ for $(x,t) \in [0,1]\times[0,1/4]$; the bottom topography was given by the function $\eta_b(x) = 1 - \beta \sin \pi x$. The scaling parameters (not important for the convergence rate computations) were taken as $\varepsilon = 1$, $\mu = 1/10$, $\beta = 1/10$. Appropriate right-hand sides and initial conditions were found from the above data. We solved numerically the ibvp's (2.6) and (2.22) with the above exact solution and bottom profile using the spatial discretizations (2.13)–(2.15) and (2.24)–(2.26), respectively, with cubic splines with uniform mesh of meshlength $h = 1/N$. The temporal discretization was realized by the RK4 scheme with stability restriction $\frac{k}{h} \leq \frac{1}{4}$; the resulting time steps were small enough so that the temporal errors were much smaller than the spatial ones. We used 3-point Gauss quadrature to evaluate the finite element integrals in every mesh interval. (Since we wished to obtain detailed information about the spatial convergence rates, we computed throughout with quadruple precision and evaluated he L^2-errors using 5-point Gauss quadrature and the L^∞ errors by taking the maximum value of the error on all these quadrature points.)

In Table 2.1 we show the L^2, L^∞, and H^1 (seminorm) spatial errors and convergence rates in the case of the (CBw) model. The numerical results suggest strongly that the L^2 rates for ζ and u are equal to 3.5 and 4, respectively, the L^∞ rates equal to 3 and 4, while the H^1 ones 2.5 and 3, respectively. The same rates are observed (cf. Table 2.2) in the numerical integration by the same method of the analogous ibvp for the (CBs) model.

As a remark of theoretical interest we point out that in the case of the analogous ibvp for (CB) on a horizontal bottom L^2 error estimates on a uniform mesh were proved in [AD13]. The error estimates were $\|\zeta - \zeta_h\| \leq ch^{3.5}\sqrt{|\ln h|}$, $\|u - u_h\| \leq ch^4\sqrt{|\ln h|}$. The increased accuracy of our present code affords investigating computationally if the logarithmic factors are actually present in these

N	L_2 error	rate	L_∞ error	rate	H_1 semi-nrm	rate
8	1.1154e-04	-	3.4102e-04	-	4.9273e-03	-
16	9.5884e-06	3.5401	3.5426e-05	3.2670	7.5831e-04	2.6999
32	8.1075e-07	3.5640	3.9539e-06	3.1635	1.2300e-04	2.6241
64	6.9606e-08	3.5420	4.6509e-07	3.0877	2.0801e-05	2.5640
128	6.0526e-09	3.5236	5.6333e-08	3.0455	3.5964e-06	2.5320
256	5.3006e-10	3.5133	6.9294e-09	3.0232	6.2857e-07	2.5164
512	4.6605e-11	3.5076	8.5918e-10	3.0117	1.1046e-07	2.5086
1024	4.1074e-12	3.5042	1.0696e-10	3.0059	1.9466e-08	2.5045
2048	3.6250e-13	3.5022	1.3343e-11	3.0029	3.4357e-09	2.5023
4096	3.2016e-14	3.5011	1.6662e-12	3.0015	6.0686e-10	2.5012
8192	2.8287e-15	3.5006	2.0816e-13	3.0007	1.0724e-10	2.5006

(a) ζ

N	L_2 error	rate	L_∞ error	rate	H_1 semi-nrm	rate
8	2.1716e-05	-	5.1473e-05	-	1.0232e-03	-
16	1.2560e-06	4.1119	2.7324e-06	4.2356	1.2429e-04	3.0413
32	7.6917e-08	4.0294	1.5843e-07	4.1083	1.5438e-05	3.0092
64	4.7794e-09	4.0084	9.6737e-09	4.0336	1.9270e-06	3.0020
128	2.9812e-10	4.0029	6.0043e-10	4.0100	2.4080e-07	3.0004
256	1.8618e-11	4.0011	3.7468e-11	4.0023	3.0098e-08	3.0001
512	1.1632e-12	4.0005	2.3407e-12	4.0006	3.7623e-09	3.0000
1024	7.2689e-14	4.0002	1.4628e-13	4.0002	4.7029e-10	3.0000
2048	4.5427e-15	4.0001	9.1419e-15	4.0001	5.8786e-11	3.0000
4096	2.8391e-16	4.0001	5.7136e-16	4.0000	7.3483e-12	3.0000
8192	1.7744e-17	4.0000	3.5710e-17	4.0000	9.1853e-13	3.0000

(b) u

Table 2.1: Spatial errors and rates of convergence, $t = 1/4$, (CBw), cubic splines on uniform mesh, $h = 1/N$, (a): ζ, (b): u.

N	L_2 error	rate	L_∞ error	rate	H_1 semi-nrm	rate
8	1.0080e-04	-	2.5352e-04	-	4.1903e-03	-
16	9.1376e-06	3.4635	3.0130e-05	3.0728	6.9419e-04	2.5936
32	7.9145e-07	3.5292	3.6312e-06	3.0527	1.1747e-04	2.5631
64	6.8773e-08	3.5246	4.4523e-07	3.0278	2.0318e-05	2.5314
128	6.0165e-09	3.5149	5.5101e-08	3.0144	3.5538e-06	2.5153
256	5.2848e-10	3.5090	6.8528e-09	3.0073	6.2481e-07	2.5079
512	4.6535e-11	3.5054	8.5440e-10	3.0037	1.1012e-07	2.5043
1024	4.1044e-12	3.5031	1.0666e-10	3.0019	1.9436e-08	2.5023
2048	3.6236e-13	3.5017	1.3324e-11	3.0009	3.4331e-09	2.5012
4096	3.2010e-14	3.5009	1.6650e-12	3.0005	6.0663e-10	2.5006
8192	2.8284e-15	3.5004	2.0809e-13	3.0002	1.0721e-10	2.5003

(a) ζ

N	L_2 error	rate	L_∞ error	rate	H_1 semi-nrm	rate
8	2.1831e-05	-	5.2866e-05	-	1.0249e-03	-
16	1.2603e-06	4.1146	2.7930e-06	4.2425	1.2441e-04	3.0423
32	7.7005e-08	4.0326	1.5971e-07	4.1283	1.5442e-05	3.0102
64	4.7813e-09	4.0095	9.6939e-09	4.0422	1.9271e-06	3.0023
128	2.9818e-10	4.0032	6.0086e-10	4.0120	2.4081e-07	3.0005
256	1.8621e-11	4.0012	3.7476e-11	4.0030	3.0099e-08	3.0001
512	1.1634e-12	4.0005	2.3411e-12	4.0007	3.7623e-09	3.0000
1024	7.2699e-14	4.0002	1.4630e-13	4.0002	4.7029e-10	3.0000
2048	4.5433e-15	4.0001	9.1432e-15	4.0001	5.8786e-11	3.0000
4096	2.8395e-16	4.0001	5.7144e-16	4.0000	7.3483e-12	3.0000
8192	1.7746e-17	4.0000	3.5714e-17	4.0000	9.1853e-13	3.0000

(b) u

Table 2.2: Spatial errors and rates of convergence, $t = 1/4$, (CBs), cubic splines on uniform mesh, $h = 1/N$, (a): ζ, (b): u.

estimates. To this end we considered the ibvp for (CB) with the exact solution given previously, but now in the case of the horizontal bottom $\eta_b = 1$, and found that the rates $\frac{\|\zeta-\zeta_h\|}{h^{3.5}}$ stabilized to the value 0.124 (the values of h used went down to $1/4096$ and the errors $\|\zeta - \zeta_h\|$ were very small,) while the ratio $\frac{\|\zeta-\zeta_h\|}{h^{3.5}\sqrt{|\ln h|}}$ did not stabilize for the same range of h's (see Table 2.3 and Figure 2.1). Similar ob-

N	$\|\zeta - \zeta_h\|$	$\frac{\|\zeta - \zeta_h\|}{h^{3.5}}$	$\frac{\|\zeta - \zeta_h\|}{h^{3.5}\sqrt{\|\ln(h)\|}}$
8	1.017830463352483e-04	0.1473975957	0.1022155669
16	8.652711027620317e-06	0.1417660175	0.0851391702
32	7.239701695841898e-07	0.1341978618	0.0720854914
64	6.180357192069458e-08	0.1296114845	0.0635557911
128	5.361327720539382e-09	0.1272058982	0.0577491370
256	4.690111486140556e-10	0.1258992215	0.0534644767
512	4.121623437941200e-11	0.1251737244	0.0501163218
1024	3.631637661581006e-12	0.1247821199	0.0473957750
2048	3.204686496602256e-13	0.1245777215	0.0451160750
4096	2.830196188452161e-14	0.1244733447	0.0431591622

Table 2.3

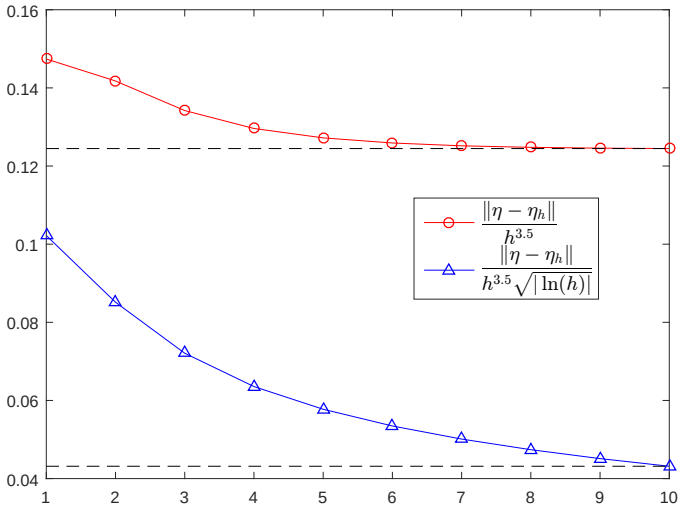


Figure 2.1: Graphs of $\|\zeta - \zeta_h\|/h^{3.5}$, $\|\zeta - \zeta_h\|/(h^{3.5}\sqrt{|\ln h|})$ as h diminishes. (Horizontal axis is $\log N$, for $N = 1/h$.)

servations were made for the u component of the error. Therefore these increased accuracy experiments suggest that the logarithmic factors are not actually present in these error estimates of [AD13].

2.3.2 Approximate absorbing boundary conditions

In the case of the shallow water (SW) equations on a horizontal bottom, obtained if we set $\mu = 0$ in the (CB) system, i.e. for the equations

$$\begin{aligned} &\zeta_t + u_x + \varepsilon(\zeta u)_x = 0, \\ &u_t + \zeta_x + \varepsilon u u_x = 0, \end{aligned} \tag{SW}$$

(written here in nondimensional, scaled variables, and where it is assumed that $1+\varepsilon\zeta > 0$), it is well known that using Riemann invariants and the theory of characteristics. [Whi74], one may derive transparent, *characteristic boundary conditions* at the endpoints of a finite spatial interval, say $[0, 1]$. These boundary conditions allow an initial pulse that is generated in the interior of $(0, 1)$ and travels in both directions to exit the interval cleanly. In the case of a subcritical flow, which will be of interest here, i.e. when the solution of (SW) satisfies $u^2 < (1 + \varepsilon\zeta)/\varepsilon^2$, the characteristic boundary conditions are of the form

$$\begin{aligned} &\varepsilon u(0,t) + 2\sqrt{1+\varepsilon\zeta(0,t)} = \varepsilon u_0 + 2\sqrt{1+\varepsilon\zeta_0}, \\ &\varepsilon u(1,t) - 2\sqrt{1+\varepsilon\zeta(1,t)} = \varepsilon u_0 - 2\sqrt{1+\varepsilon\zeta_0}. \end{aligned} \tag{2.33}$$

Here it is assumed that outside the interval $[0, 1]$ the flow is uniform and satisfies $\zeta(x, t) = \zeta_0$, $u(x, t) = u_0$, where η_0, u_0 are constants such that $u_0^2 < (1+\varepsilon\zeta_0)/\varepsilon^2$. In addition, the initial conditions $\zeta(x, 0)$, $u(x, 0)$, of (SW) should satisfy the subcriticality conditions and be compatible at $x = 0$ and $x = 1$ with the uniform flow outside $[0, 1]$. In [AD17] the authors analyzed the space discretization of (SW) with characteristic boundary conditions (both in the subcritical and supercritical case) using Galerkin finite element methods. Analytical and computational evidence in [AD17] suggests that the discretized characteristic boundary conditions, although not exactly transparent, are nevertheless highly absorbent. We note that the same type of characterisic absorbing conditions may be used for the (SW) over a variable bottom, at least in the case where the bottom is locally horizontal at the endpoints cf. e.g. Chapter 3 and its references.

Finding (exact) transparent boundary conditions for the (CB) is not easy, as a nonlocal problem should be solved for this nonlinear system. In practice, for small μ, it is reasonable to assume that the Riemann invariants do not change much over short distances along the characteristics, and, consequently, to pose the b.c. (2.33) as approximate, absorbing b.c.'s for (CB) as well. This has been widely done in practice, for example in numerical simulations of the Serre equations cf. e.g. [CBB07], [Bon+11]; in [DM10] the related problem of deriving one-way approximations of the Serre equations is discussed. Our aim in this subsection is to assess, by numerical experiment, the accuracy of (2.33) as approximate absorbing boundary conditions for the (CB), paying special attentions to their efficacy in simulating outgoing *solitary-wave* solutions of the (CB).

To derive (classical) solitary-wave solutions of (CB) on the real line, we let $\zeta = \zeta_s(x - c_s t)$, $u = u_s(x - c_s t)$, where c_s is the speed of the solitary wave and

$\zeta_s(\xi)$, $u_s(\xi)$ are smooth functions that tend to zero, along with their derivatives, as $|\xi| \to \infty$. Inserting these expressions in (CB) and integrating we see that the equations for η_s and u_s decouple and give

$$\zeta_s = \frac{u_s}{c_s - \varepsilon u_s}, \quad \frac{c_s\mu}{3}u_s'' + \frac{\varepsilon}{2}u_s^2 - c_s u_s + \frac{u_s}{c_s - \varepsilon u_s} = 0, \tag{2.34}$$

A further integration yields that u_s satisfies the ode

$$\frac{c_s\mu}{6}(u_s')^2 + \frac{\varepsilon}{6}u_s^3 - \frac{c_s}{2}u_s^2 - \frac{1}{\varepsilon}u_s - \frac{c_s}{\varepsilon^2}\ln\frac{c_s - \varepsilon u_s}{c_s} = 0. \tag{2.35}$$

It is straightforward to see that ζ_s and u_s have a single positive maximum at some point ξ_0 (we assume that $\xi_0 = 0$). Denoting $A = \max \zeta_s$, $B = \max u_s$, we get

$$A = \frac{B}{c_s - \varepsilon B}, \quad \varepsilon B^3 - \frac{c_s}{2}B^2 - \frac{1}{\varepsilon}B - \frac{c_s}{\varepsilon^2}\ln\left(\frac{c_s - \varepsilon B}{c_s}\right) = 0, \tag{2.36}$$

from which one may compute the speed-amplitude relation

$$c_s = \frac{\sqrt{6}(1+\varepsilon A)}{\sqrt{3+2\varepsilon A}}\frac{\sqrt{(1+\varepsilon A)\ln(1+\varepsilon A) - \varepsilon A}}{\varepsilon A}. \tag{2.37}$$

For fixed ε, c_s is monotonically increasing with A but stays below the straight line $c_s = 1 + \frac{\varepsilon A}{2}$, which is the speed-amplitude relation of the solitary waves of the Serre equations. (The formulas (2.34)–(2.37) were derived in [AD12] in the case of the unscaled (CB). Note that there are some typographical errors in [AD12]: In equation (1.58) of [AD12] the last term in the left-hand side of the equation should have the sign $+$, while in the equation preceding (2.34) in [AD12] the third term in the left-hand side should have the sign $+$ and the last term the sign $-$. However formulae (3.1) and (3.2) of [AD12], which are the analogous of (2.37) and (2.36) above, are correct.)

When εA is not large, i.e. when (CB) is a valid model for surface waves, it may be seen by (2.37) and also by numerical simulations that the solitary-wave solutions of (CB) satisfy the subcriticality condition. (Since there is no closed-form formula for the solitary waves we generate them numerically by solving for given c_s the nonlinear o.d.e. (2.34) that u_s satisfies, taking zero boundary conditions for u_s and u_s' at the endpoints of a large enough spatial interval using the routine `bvp4c` of [MAT18].)

In the numerical experiments to be described in the sequel we solved the (SW) and the (CB), unless otherwise specified, by the standard Galerkin-finite element method on the spatial interval $[0, 50]$ using cubic splines on a uniform mesh with $h = 0.025$, coupled with RK4 time stepping with time step satisfying $\frac{k}{h} = \frac{1}{2}$, up to $T = 50$.

We set the stage by solving numerically the (SW) with $\varepsilon = 1$ with the b.c. (2.33), posed now at the endpoints of $x = 0$ and $x = 50$. As initial condition we take the solitary wave of (CB) with $\mu = \varepsilon = 1$ of speed $c_s = 1.18112$, centered

at $x = 25$, which we multiply by a factor 0.1 (thus it is no longer a solitary wave), so that no discontinuities develop in its evolution under (SW) for the duration of the experiment. As expected, the initial single-hump wave is split in two pulses: a larger one of amplitude of about 0.04 traveling to the right with a speed of about 1.057 and starts exiting the computational interval at $x = 50$ at about $t = 22.5$, (the exit is completed by about $t = 30$), and a smaller one of amplitude of about 0.0035 that travels to the left with speed 1.005 and exits the interval at $x = 0$ at about $t = 24.5$ (see Figure 2.2).

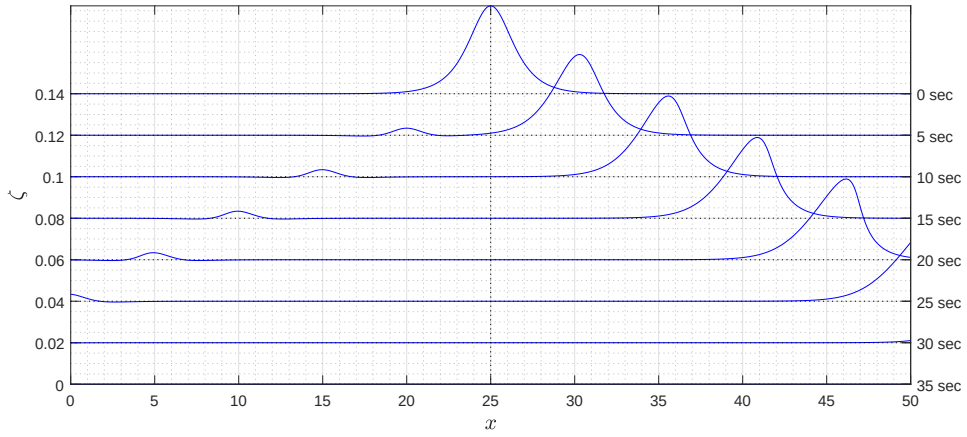


Figure 2.2: Two-way propagation and exit of pulses, (SW) with b.c. (2.33).

In Figure 2.3 we present some graphs that are relevant for assessing the accuracy of the absorbing b.c.'s for this example. (All graphs refer to ζ.) In Fig. 2.3(a) we observe the temporal variation of the wavefield at $x = 40$. The pulse that travels to the right passes this gauge and exits the interval. What remains after $t \simeq 30$ is a small residual consisting of small-amplitude oscillations reflected from the boundary due to the inexactness of the discretized b.c.'s and shown in the magnification of 2.3(a) to be of $\mathcal{O}(10^{-9})$. In 2.3(b) we show the maximum amplitude of ζ with respect to x over the whole interval as a function of t, while 2.3(c) shows the small oscillations still present in the computational interval at the end of the experiment ($t = 50$). The are all of magnitude at most 10^{-9} and consist of a main wavepacket of high frequency and amplitude of about 4×10^{-10} centered at about $x = 40$ and moving to the right, and three larger amplitude 'thin' wavetrains of small support centered at about $x = 5$ (moving to the right), $x = 20$ (moving to the left) and $x = 37.5$ (moving to the left), respectively. The main oscillatory wavepacket is produced when the right-traveling pulse exits the boundary at $x = 50$. This wavepacket moves to the left with speed equal to about 7 and has undergone three reflections at the boundary by $T = 50$. The thinner wavetrains (of speed about 1) are generated by the interaction of this wavepacket with the boundaries (The left-traveling pulse produced by the splitting of the initial condition produces, when it hits the boundary at $x = 0$, artificial reflections with amplitude well below 10^{-10}.)

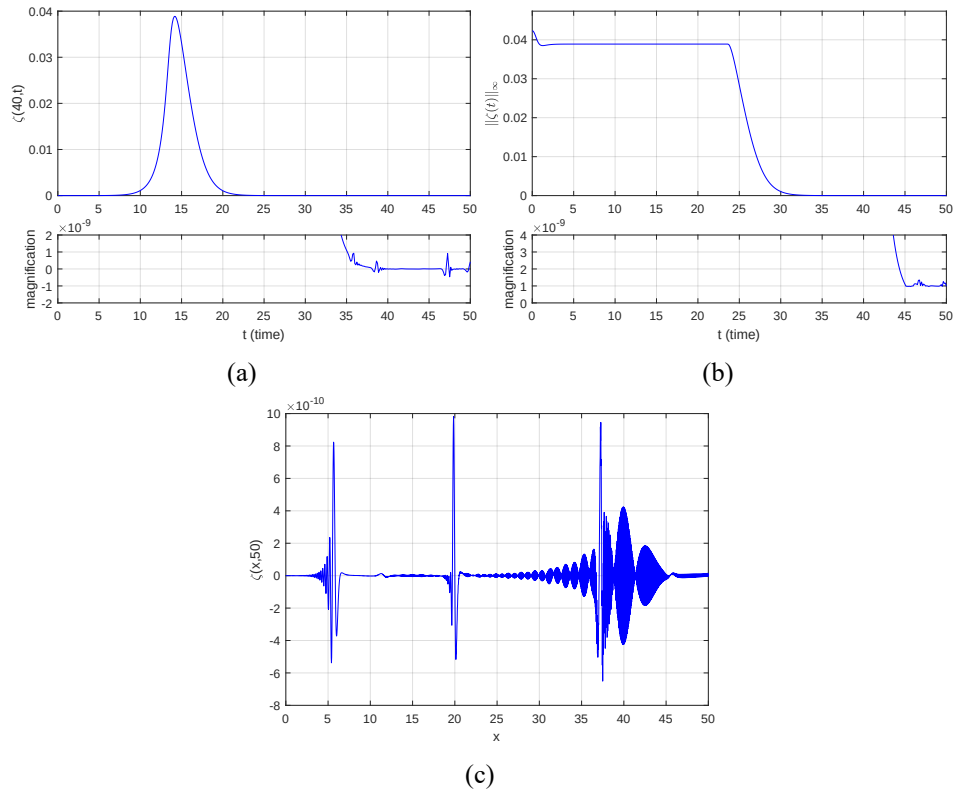


Figure 2.3: Accuracy of the numerical characteristic b.c.'s for the (SW), $\varepsilon = 1$, (a): $\zeta(40,t)$ with magnification underneath, (b): $\max_x \zeta(x,t)$ with magnification underneath, (c): Magnification of $\zeta(x,50)$

In Figure 2.4, resp. 2.5, we show analogous graphs in the case of the (CB) system in the cases $\varepsilon = \mu = 0.1$, resp. $\varepsilon = \mu = 0.01$. As initial condition we took now the exact solitary-wave profile of (CB) for these values of ε, μ, and of speed $c_s = 1.18112$. As a consequence, the wave moves to the right without changing its shape. The fact that the characteristic b.c.'s are no longer exactly transparent for the continuous system is manifested by the larger magnitudes of the residual artificial oscillations, which are now of $\mathcal{O}(10^{-3})$, resp. $\mathcal{O}(10^{-4})$. (Note their dispersive character in the larger μ case, Fig. 2.3(c).) The main pulse in graph (c)

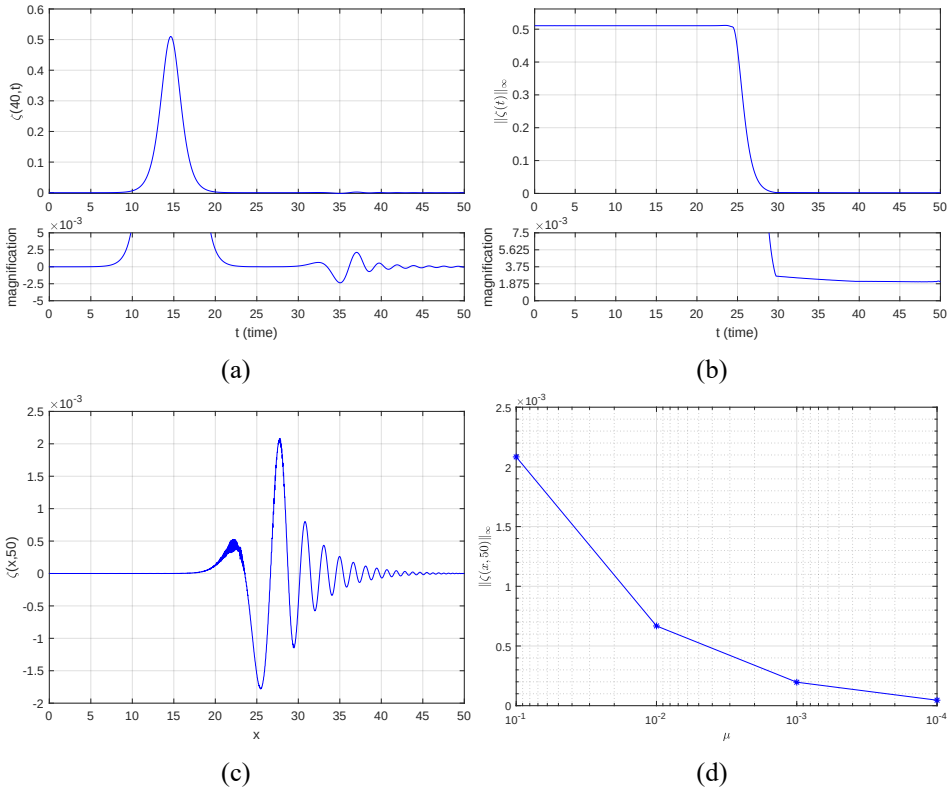


Figure 2.4: Accuracy of the numerical characteristic b.c.'s for the (CB), $\varepsilon = \mu = 0.1$, (a): $\zeta(40,t)$ with magnification underneath, (b): $\max_x \zeta(x,t)$ with magnification underneath, (c): Magnification of $\zeta(x,50)$, (d): $\|\zeta(\cdot,50)\|_\infty$ for $\varepsilon = 0.1$ versus μ.

of Figures 2.4 and 2.5 is due to the modelling, i.e. the approximate character of the characteristic b.c.'s, while the superimposed noise in Fig. 2.5(c) disappears as h is decreased. The amplitude of the residual was equal to about 2.1×10^{-3} for $\varepsilon = \mu = 0.1$ and fell to 3.2×10^{-4} for $\varepsilon = \mu = 0.01$, and to 3.3×10^{-5} for $\varepsilon = \mu = 0.001$ (figure not shown). We thus observe that it decreases linearly with μ when $\varepsilon = \mu$. As expected, for fixed ε we observed that this amplitude decreased with μ. For example, for $\varepsilon = 0.01$ and $\mu = 10^{-3}$ it was equal to about 3.6×10^{-5},

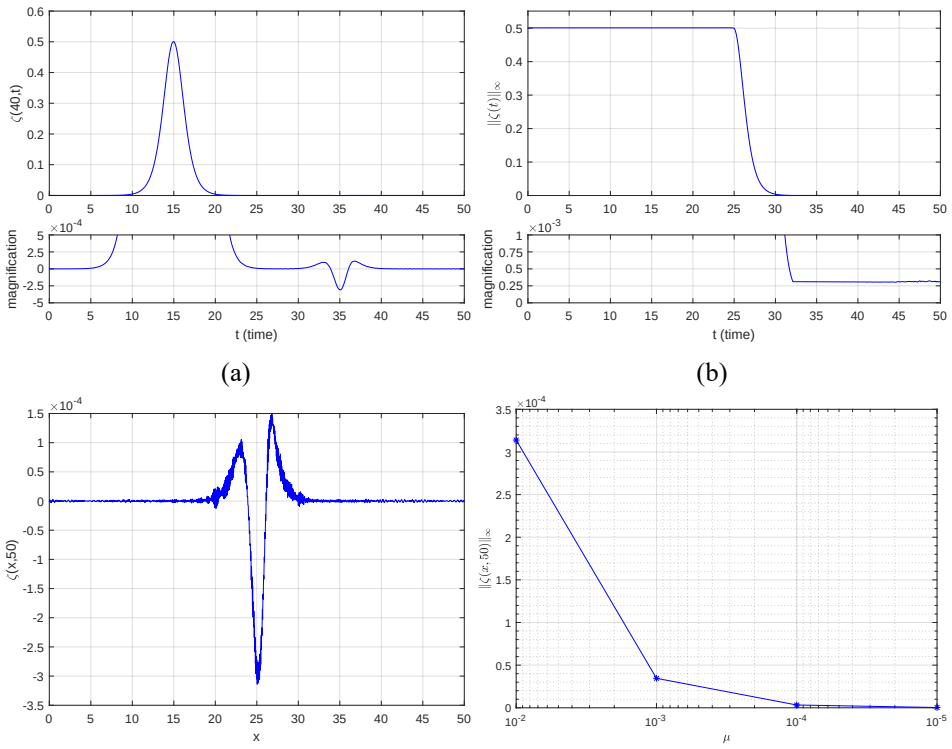


Figure 2.5: Accuracy of the numerical characteristic b.c.'s for the (CB), $\varepsilon = \mu = 0.01$, (a): $\zeta(40,t)$ with magnification underneath, (b): $\max_x \zeta(x,t)$ with magnification underneath, (c): Magnification of $\zeta(x,50)$, (d): $\|\zeta(\cdot,50)\|_\infty$ for $\varepsilon = 0.01$ versus μ.

for $\mu = 10^{-4}$ it was of $\mathcal{O}(10^{-6})$, cf. Figures 2.4, 2.5, (d).

Our conclusion is that for small $\varepsilon = \mu$, i.e. when the (CB) is a valid model, the (approximate) characteristic b.c.'s for the (CB) are satisfactorily absorbing. We extended these b.c.'s in the case of the variable bottom models (CBw) and (CBs) and used them in numerical experiments that will be reported in the next subsection.

2.3.3 Propagation of solitary waves over a variable bottom

In this subsection we present the results of several numerical experiments we performed with the variable-bottom models (CBw) and (CBs) in order to validate the numerical methods used for their solution, compare the two models, and compare the results of (CBs) with those obtained by the Serre-Green-Naghdi system and with experimental measurements. We mainly use test problems already considered in the literature, whose main theme is the study of the changes that solitary-wave pulses undergo when propagating over an uneven bottom.

2.3.3.1 Solitary waves on a sloping beach

We first consider the problem of a solitary wave climbing a sloping beach of mild slope that was studied by Peregrine in his pioneering study [Per67], in which he derived the (CBs) system and solved it numerically by a finite differene scheme. In our experiments we used the (CBs) in unscaled, nondimensional variables (i.e. setting $\varepsilon = \mu = 1$) and solved it with our fully discrete scheme using cubic splines on a uniform mesh with $N = 2000$ spatial intervals and $M = 2N$ temporal steps. Following [Per67] we consider, using out notation, a bottom of uniform slope $\alpha > 0$ given by $\eta_b(x) = \alpha x$ on a spatial interval of the form $[0, L_\alpha]$. As initial condition we take as in [Per67] a solitary wave of the form

$$\zeta_0(x) = a_0 \operatorname{sech}^2 \left[\tfrac{1}{2}\sqrt{3a_0}(x - x_0) \right], \tag{2.38}$$

where $x_0 = 1/\alpha$. This is a solitary wave of the KdV type equation $\zeta_t + \zeta_x + \frac{3}{2}\zeta\zeta_x + \frac{1}{6}\zeta_{xxx} = 0$ with speed $c_s = 1 + a_0/2$. The KdV equation in this form is obtained as a one-way approximation of the (CB) with $\varepsilon = \mu = 1$ in the standard manner, cf. [Whi74]. The particular solitary wave (2.38) is centered at $x_0 = 1/\alpha$, where the (undisturbed) water depth is equal to one. The initial velocity of the pulse, found by inserting (2.38) in the continuity equation, is given by

$$u_0(x) = \frac{-\left(1 + \frac{1}{2}a_0\right)\zeta_0(x)}{\alpha x + \zeta_0(x)}. \tag{2.39}$$

Thus the initial condition (2.38)–(2.39) is not an exact solitary-wave solution of (CB) but a close approximation thereof. We took an interval of length $L_\alpha = 1/\alpha + 20$ to ensure that the support of the initial pulse was well within the spatial interval of integration. At $x = 0$ we used the b.c. $u = 0$ (which produced no reflections as the wave did not reach the left boundary within the temporal range of

the experiment), posed absorbing (characteristic) boundary conditions at $x = L_\alpha$, and ran the experiment up to $t = 25$.

During this temporal interval the wave moves to the left, steepens (wave 'shoaling') and grows in amplitude, cf. Figure 2.6; its evolution resembles that of Fig. 1 of [Per67], which corresponds to $\alpha = 1/30$, $a_0 = 0.1$. We compared our numer-

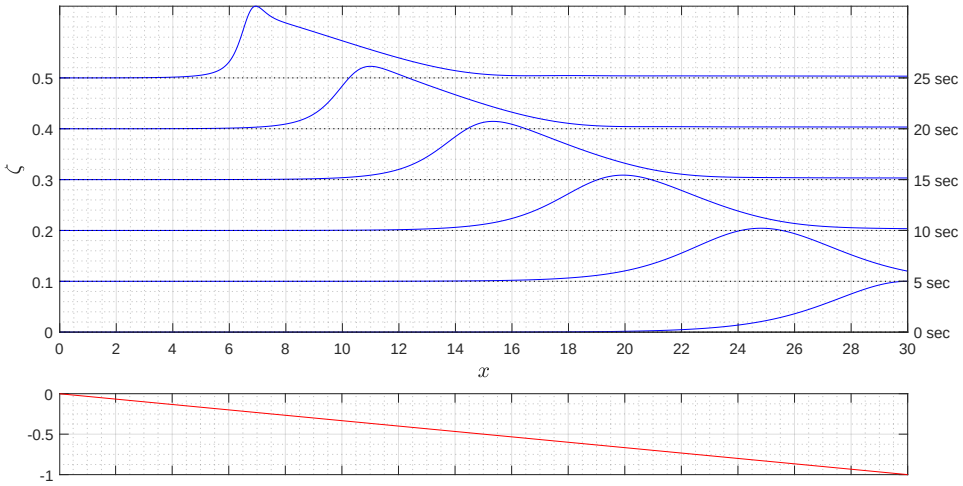


Figure 2.6: (CBs) system. Solitary wave (2.38)–(2.39) climbing a sloping beach, $\alpha = 1/30$, $a_0 = 0.1$.

ical results with those of the finite-difference scheme of Peregrine (given in the Appendix of [Per67]) that we implemented. (Note that there is a misprint in the last equation of this scheme in [Per67]: In the discretization of the term $\eta_b u_x$, the denominator should be $4\Delta x$.) We observed that the maximum discrepancy in the amplitude of ζ approximated by the two methods occurred at $t = 25$ where the values were 0.14100 for our scheme and 0.13634 for the scheme of [Per67] (implemented with $\Delta x = \Delta t = 0.1$), which corresponds to a difference of about 3.4% (Fig. 1 of [Per67] shows a ζ-amplitude of about 0.15 at $t = 25$ which does not correspond to the actual numerical results that the scheme of [Per67] gives and is probably due to some inaccuracy in the graphics.)

We also repeated with our scheme the numerical experiments leading to Fig. 2 of [Per67] that depicts the change of amplitude of the solitary wave with depth for various values of a_0 in the case of a beach of slope $\alpha = 1/20$. We show our results in Figure 2.7. There was good agreement for low values of a_0; however the values given in [Per67] for $a_0 = 0.2$ seem too high as the depth approaches 0.4. (All the amplitudes computed by our scheme stay below the curve of Green's law for depths larger than 0.5.)

As the solitary wave climbs the sloping beach a small-amplitude flat wave of elevation is reflected backwards due to the presence of the sloping bed. The results of our computations, cf. Figure 2.8, agree with those of Fig. 3 in [Per67]. Peregrine, *op. cit.*, derives an approximate expression for the amplitude of the reflected wave

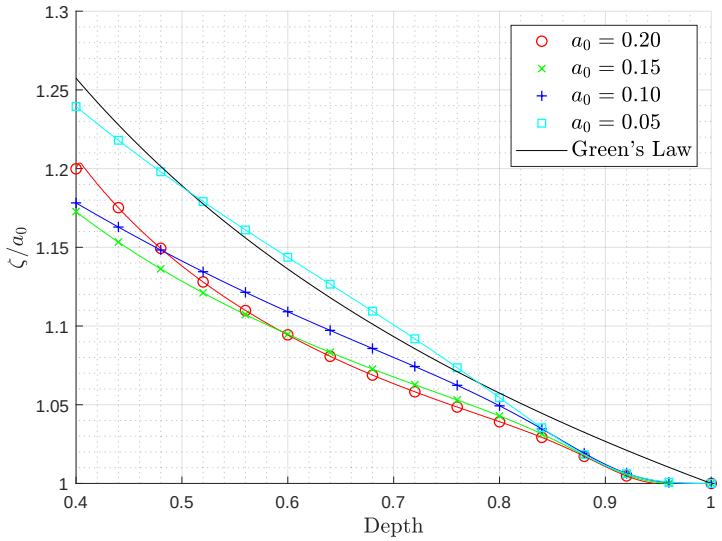


Figure 2.7: Change of amplitude with depth. (CBs), solitary wave, $\alpha = 1/20$, various initial amplitudes.

of the form

$$\zeta_{\text{max,refl}} \simeq \frac{1}{2}\alpha\left(\frac{1}{3}a_0\right)^{\frac{1}{2}}, \tag{2.40}$$

using characteristic variables for the linearized shallow water equations. We found quite a good agreement between our numerical results and the values computed by (2.40). For example, for $\alpha = 1/40$, $a_0 = 0.1$, our computations gave $\zeta_{\text{max,refl}} = 0.0023$, while (2.40) gives 0.0025. We will return to the reflections due to the uneven bottom in subsection 2.3.3.3 in the sequel.

As was previously mentioned, we used the approximate characteristic boundary conditions discussed in subsection 2.3.2 at the right-hand boundary $x = L_\alpha$. We found that this b.c. also works for a sloping bottom provided the length of the

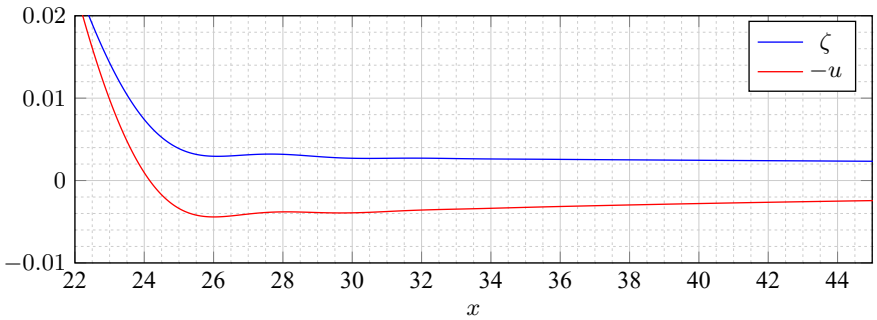


Figure 2.8: (CBs): Reflexion, due to the sloping bottom, from a solitary wave. $\alpha = 1/40, a_0 = 0.1, t = 25$.

domain is taken sufficiently large so that the artificial oscillations created at the boundary do not interfere as they travel to the left with the reflected wave due to the slope. As an example we consider a beach of slope $\alpha = 1/40$ on the spatial interval $[0, 70]$. As initial condition we took $\zeta(x, 0) = \zeta_0(x)$ given by (2.38) with $x_0 = 40$, $a_0 = 0.1$, and $u(x, 0) = 0$, i.e. a 'heap' of water, so that sizeable pulses are generated and propagate in both directions. The two-way propagation is shown in Figure 2.9. Figure 2.10 shows a magnified profile of the surface elevation ζ as

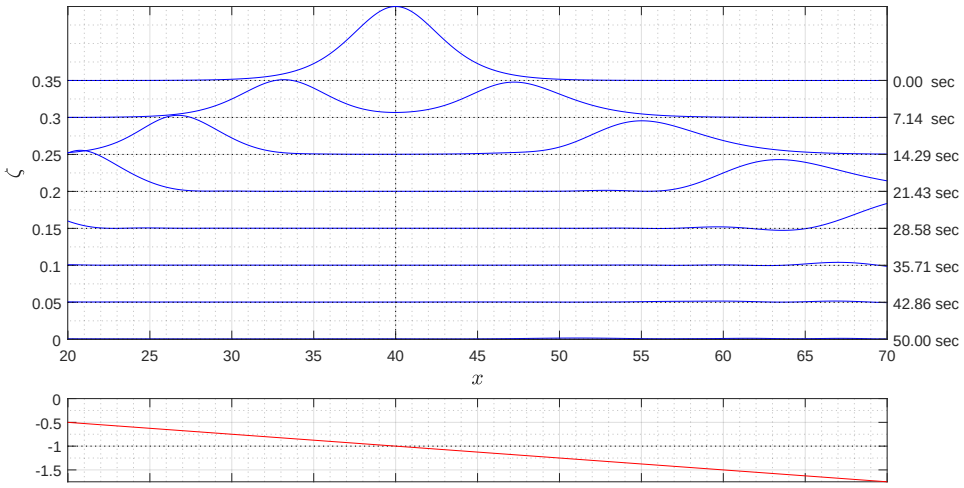


Figure 2.9: Two-way propagation of a 'heap' of water, (CBs), solitary wave, $a_0 = 0.1$, $x_0 = 40$, beach slope $= 1/40$. Bottom is depicted in the lower graph.

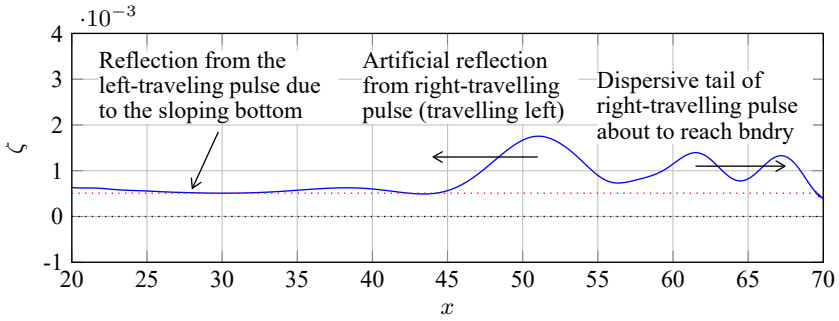


Figure 2.10: Magnification of ζ reflections near the right boundary, $t = 50$

a function of x in the interval $[20, 70]$ at $t = 50$, by which time the right-travelling pulse has left the domain. In the interval $[20, 45]$ we observe the small-amplitude (of height approximately 5×10^{-4}) reflection due to interaction of the left-travelling pulse with the sloping bottom. In the interval $[45, 60]$ we observe the artificial oscillations reflected from the right-hand boundary at $x = 70$ due to the approximate absorbing b.c. after the exit of the main right-travelling pulse. The ratio of the amplitude of the artificial reflection to that of the main pulse is about 4%. Finally, one

may also observe on the extreme right the dispersive-tail oscillations that follow the main right-travelling pulse as they exit the domain.

2.3.3.2 Transformation of a solitary wave propagating onto a shelf

We next consider in detail an example of the transformation that a solitary wave undergoes as it propagates over a bottom of shelf type like the one shown in Figure 2.11. This test problem was considered by Madsen and Mei in [MM69]. In this subsection we work in dimensionless, unscaled variables with $\varepsilon = \mu = 1$.

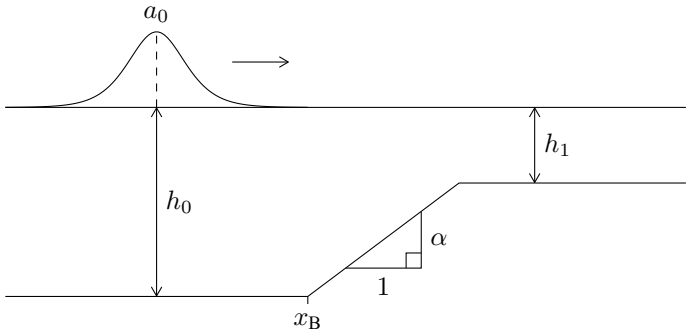


Figure 2.11: Solitary wave propagating onto a shelf

The initial elevation of the solitary wave is given again by (2.38), in which x_0 is taken far enough from the toe of the sloping part of the bottom at $x = x_{\rm B}$, so that $\zeta_0(x_{\rm B})/a_0 \ll 1$. The initial velocity is found again from the continuity equation but is now computed for a bottom of constant depth $h_0 = 1$, i.e. as

$$u_0(x) = \frac{\left(1 + \frac{1}{2}a_0\right)\zeta_0(x)}{1 + \zeta_0(x)}. \tag{2.41}$$

The solitary wave travels to the right, changes in amplitude and shape as it climbs the slope, and resolves itself into a sequence of solitary-wave pulses as it travels on the shelf of uniform depth $h_1 < 1$, cf. Figures 2.12, 2.15.

In [MM69] the pde model used was a Boussinesq system of KdV-BBM type with variable-bottom terms originally derived in [ML66], and which, in the case of horizontal bottom, is locally well-posed, cf. [BCS04]. The initial-value problem was integrated with a type of a method of 'characteristics'. In order to form some idea of the proximity of the model used in [MM69] to (CBs) we integrated both systems using our fully discrete scheme with cubic splines and RK4 time stepping over a variable bottom domain like that of Figure 2.11 with $0 \le x \le 150$, $x_B = 60$, $h_1 = 0.5$, $\alpha = 1/20$. As initial values we took solitary waves of the respective systems of the same amplitude $a_0 = 0.12$ and centered at $x_0 = 30$. (Their speeds are very close but the wavelength of the solitary wave of the system of [MM69] was about 22% larger. The difference of the two-solitary waves in L^2 was about 4.37×10^{-2}.) The evolution of both systems is shown in Figure 2.12 At the end

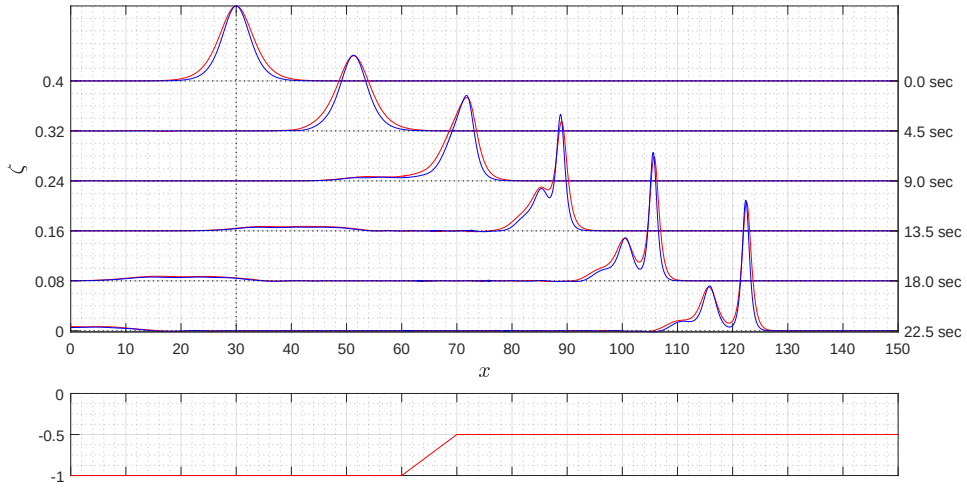


Figure 2.12: Comparison between Madsen & Mei system (red, the one with the larger wavelength) and (CBs) (blue). Propagation onto a shelf. Bottom is shown in the lower graph.

of the computational domain at $t = 22.5$, when both waves had climbed well onto the shelf and resolved themselves into two solitary waves plus dispersive tail, the two wavetrains had an L^2 distance of 5.53×10^{-2}, while the leading solitary waves had a difference in amplitude of about 3×10^{-3} and a phase difference (distance of positions of the crest) of 0.15. We conclude that in the time scales of this and similar experiments typical solutions of the two systems stay close to each other, so that it is fair to compare in a general way the results of numerical experiments in [MM69] with similar ones that we performed with (CBs) to be described in the sequel.

We first make some quantitative remarks on the transformation of the solitary wave as it climbs on the sloping part of the bottom in Figure 2.11. As observed in subsection 2.3.3.1, the amplitude of the solitary wave increases as the depth of the water decreases. In order to quantify this increase in the case of (CBs) and our numerical method, and motivated by analogous experiments in [MM69], we took $h_1 = 0.1$, $\alpha = 1/20$, $0 \leq x \leq 150$, $x_{\rm B} = 60$, and computed with cubic splines, $N = 3000$, $M = 2N$, the evolution (according to (CBs)) of a solitary wave of (CB) centered at $x = 30$. We recorded the variation of the normalized amplitude $\zeta_{\rm max}/a_0$ of the solitary wave as a function of the water depth η_b for various values of the initial amplitude a_0. In Figure 2.13 we show the outcome of these numerical experiments corresponding to solitary waves of initial amplitudes $a_0 = 0.1$, 0.15 and 0.2. (The graph starts when the crest of the solitary wave is at $x = x_{\rm B}$. At that point $\zeta_b = 1$, but the forward point of the solitary wave is already travelling on the sloping bed; hence, the corresponding value of $\zeta_{\rm max}/a_0$ is about 1.034 and not 1. For η_b larger than about 0.6 the three curves corresponding to the three amplitudes chosen are quite close to each other with the lowest initial amplitude

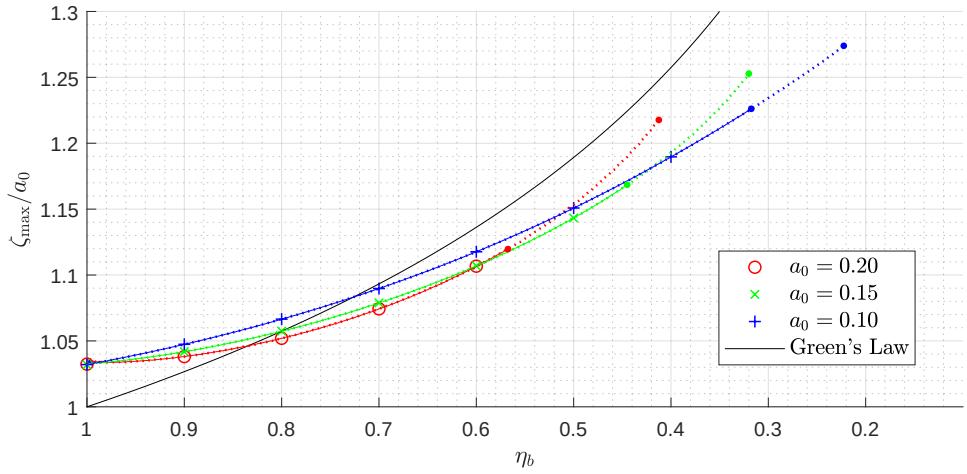


Figure 2.13: Amplitude variation with depth for beach slope $\alpha = 1/20$ for $a_0 = 0.2,\ 0.15,\ 0.1$. Computation stopping criteria: solid lines, $\max_x(\zeta(x,t)/\eta_b(x)) < 0.4$; dotted lines, $\max(\zeta(x,t)/\eta_b(x)) < 0.6$.

$a_0 = 0.10$ giving the highest values of $\zeta_{\max}/a_0$. For η_b smaller than about 0.5 the sequence is reversed with the highest $a_0 = 0.2$ giving the highest $\zeta_{\max}/a_0$ values. The initial solid-line part of the three curves represents the values of $\zeta_{\max}/a_0$ up to the point where $\max_x \left(\frac{\zeta(x,t)}{\eta_b(x)}\right) = 0.4$, which is probably a large upper bound of the range of validity of (CBs), while the dotted-line extensions of the curves go up to $\max_x \left(\frac{\zeta(x,t)}{\eta_b(x)}\right) = 0.6$, which is probably beyond that range. We also show the curve of Green's law given by $\zeta_{\max}/a_0 = \eta_b^{-1/4}$ for comparison purposes. It is to be noted that our results are in satisfactory agreement with those of the corresponding Fig. 3 of [MM69] for values of η_b in the range 1 to 0.75.

These results are supplemented by those of Figure 2.14 in which we record the variation of $\zeta_{\max}/a_0$ as a function of η_b for a solitary wave of fixed $a_0 = 0.1$ and slopes equal to 0.023, 0.05, and 0.065. For η_b larger than about 0.65 all curves are fairly close to each other with the steeper slopes giving slightly higher values of $\zeta_{\max}/a_0$. For values of η_b less than about 0.65 the smaller slope gives the highest ratio $\zeta_{\max}/a_0$ while the two other curves remain close together (stopping criteria as in Figure 2.11). A qualitatively similar behavior is observed in the analogous Figure 4 of [MM69].

The distortion the solitary wave suffers as it travels upslope causes the wave, when it reenters a horizontal-bottom region reaching the shelf, to resolve itself into a sequence of solitary waves followed by dispersive oscillations. This phenomenon was noticed in [MM69] for the model used in that paper, and is also present in our case of the (CBs) system as well. In Figure 2.15 we show this phenomenon, which may be viewed as a manifestation of the stability of solitary waves of (CB). We took a spatial interval $[0, 150]$, $h_1 = 0.5$, $x_B = 60$, $\alpha = 1/20$, and considered

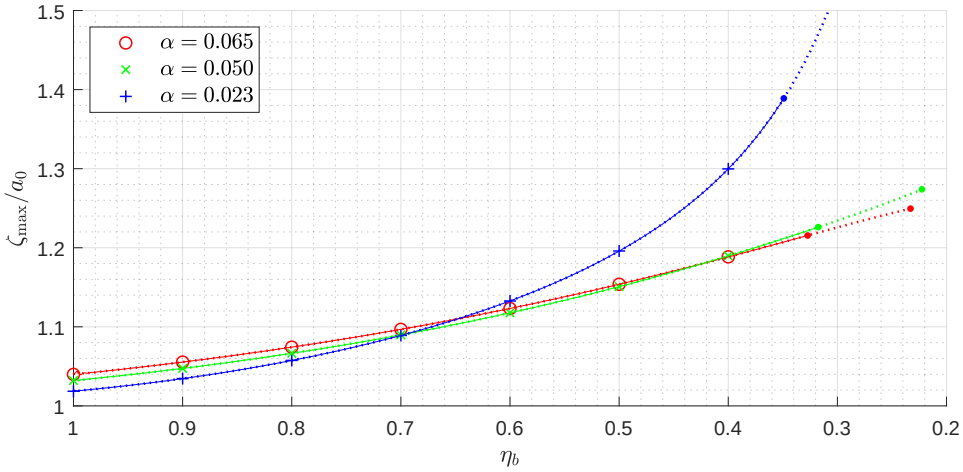


Figure 2.14: Amplitude variation with depth for initial amplitude $a_0 = 0.1$ for slopes $\alpha = 0.065,\ 0.05,\ 0.023$. Computation stopping criteria: solid lines, $\max_x(\zeta(x,t)/\eta_b(x)) < 0.4$; dotted lines, $\max_x(\zeta(x,t)/\eta_b(x)) < 0.6$.

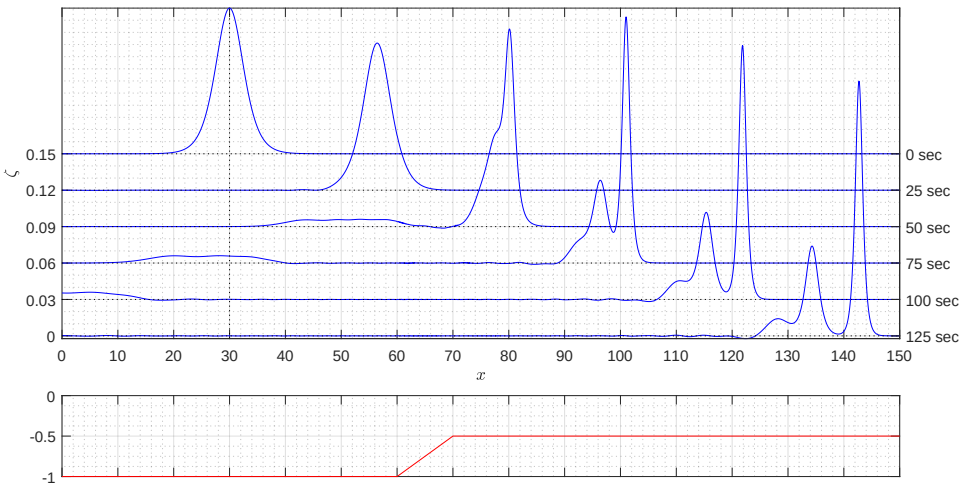


Figure 2.15: Transformation of a CB solitary wave ($a_0 = 0.12$) propagating up a slope of $\alpha = 1/20$, onto a shelf of smaller depth, $h_1 = 0.5h_0$. (CBs) computation. Bottom is shown in the lower graph.

the evolution of a solitary wave of initial amplitude $a_0 = 0.12$. The graphs in Figure 2.15 show the temporal evolution every 25 temporal units ("seconds"). The solitary wave distorts as it climbs the sloping part of the bottom (depicted in the lower part of he graph), increases in amplitude, and by $t = 125$ it has resolved itself into two solitary waves (a third is also possibly forming) plus a dispersive tail. The first solitary wave has an amplitude of about 0.2099 and travels at a speed of about 0.84. (We checked that it is indeed a CB-solitary wave.) This wavetrain is followed by the usual for upsloping environments flat reflection wave that travels to the left. The results of a similar experiment in [MM69] are qualitatively the same.

2.3.3.3 Reflection and dispersion from various types of variable bottom

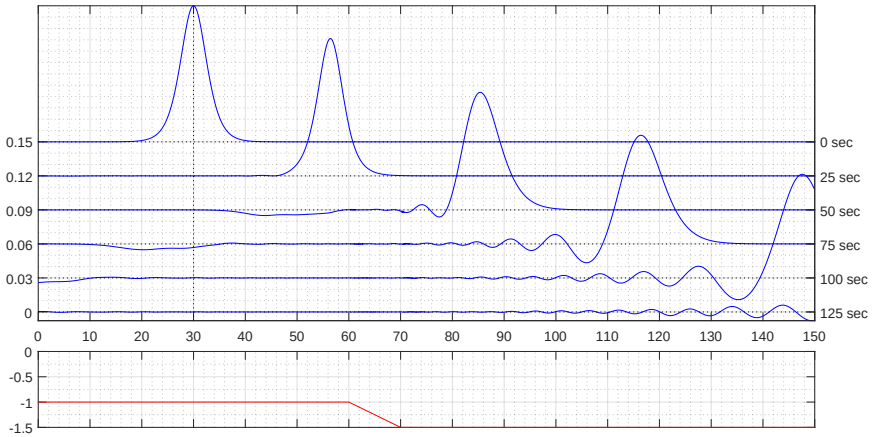
As already mentioned in subsection 2.3.3.1, when a solitary wave propagates up a sloping bottom, a small-amplitude, flat wave of elevation is generated by reflection from the uneven bottom and travels in the opposite direction. This phenomenon has been shown e.g. in Figs 2.10 and 2.15. (In this subsection we work again in dimensionless, unscaled variables with $\varepsilon = \mu = 1$.) Using characteristic variables theory for the linearized shallow water equations, in addition to the approximate formula (2.40) for the reflected wave, Peregrine predicted in [Per67] that the reflected wave will have a wavelength of about $2L$ if the slope occurs over a horizontal interval of length L. In order to check these results we integrated the (CBs) over the variable bottom shown in the lower graph of Figure 2.15 with an initial solitary wave of (CB), varying the slope and the initial amplitude a_0 of the wave; we present the results in Table 2.4 that shows the amplitudes and wavelengths of the reflected wave predicted in [Per67] and the numerical results given by our code. (Due to the shape of the reflected wave we measured its length by the formula $\frac{1}{|I|}\int_I \zeta \, \mathrm{d}x$, where $I = \{x : \zeta > 0.8\,\zeta_{\max}\}$, at a short time after the full reflected wave had formed. In the case $\alpha = 1/40$, $a_0 = 0.18$, we took $I = \{x : \zeta > 0.6\,\zeta_{\max}\}$.) We conclude that the predictions of [Per67] underestimate by a small amount the actual

α	a_0	L	refl. ampl. by (2.40)	reflected amplitude	reflected wavelength
1/20	.12	10	5.000e-3	5.578e-3	22.35
1/40	.12	20	2.500e-3	2.875e-3	43.00
1/20	.18	10	6.124e-3	6.880e-3	21.25
1/40	.18	20	3.062e-3	3.451e-3	41.65

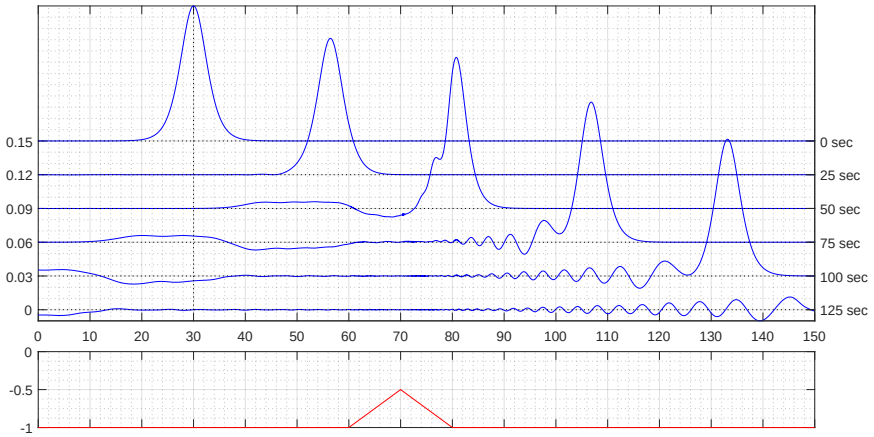
Table 2.4: Predicted and numerical values of amplitude and wavelength of reflected wave.

numerical results.

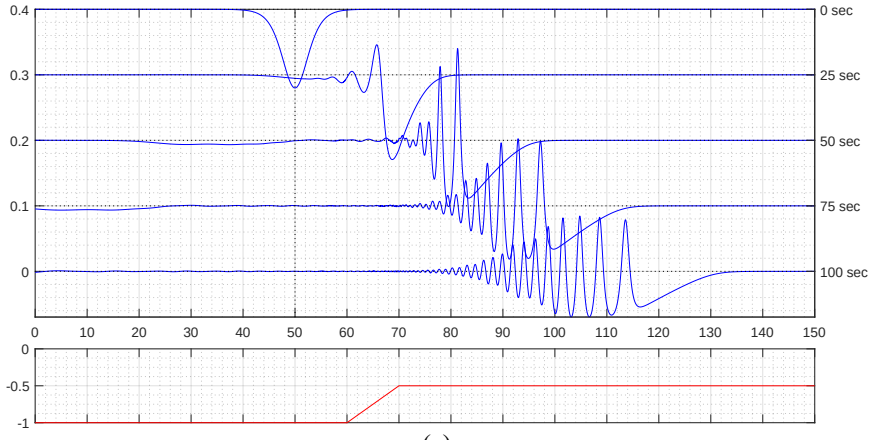
In [Per67] Peregrine also made some qualitative comments about the type of reflected waves generated by various kinds of uneven bottoms. We verified his general statements by performing various numerical experiments, the results of some of which appear in Figure 2.16. In each case an initial wave, originally on a horizontal



(a)



(b)



(c)

Figure 2.16: Reflection due to sloping bottom, various topographies. $\zeta(x,t)$ as a function of x at various t. (a): solitary wave travelling into deeper water, (b): solitary wave passing over a hump, (c): wave of depression travelling into shallower water. The various bottom topographies are shown in the lower graphs.

bottom, is let to evolve under (CBs) and travel over uneven bottoms of various simple topographies shown in the lower graphs in Figure 2.16. Fig. 2.16(a) shows a CB solitary wave of amplitude $a_0 = 0.12$ passing into shallower water. The resulting reflected wave is a wave of depression; this solitary wave seems to be dispersing as a result of its interaction with the bottom. In the case of a hump (Fig. 2.16(b)) the same initial wave gives rise first to a reflected wave of elevation followed by a reflected wave of depression as one would expect. This particular perturbation due to this bottom topography seems to lead to a solitary wave very close to the initial one plus a trailing dispersive tail. Finally, an initial wave of depression climbing upslope gives rise to a reflected wave of depression and large-amplitude dispersive oscillations as it travels on the shelf.

2.3.3.4 Comparison of (CBs) and (CBw) as the variation of the bottom increases

As was mentioned in Chapter 1 (CBs) is valid as a model for bottoms where topography, described by $\eta_b(x) = 1 - \beta b(x)$, may vary arbitrarily (so that $\eta_b > 0$ of course), i.e. where the parameter β can be taken as an $\mathcal{O}(1)$ quantity, while (CBw) was derived under the assumption that $\beta = \mathcal{O}(\varepsilon)$. In this subsection we suppose that the systems are written in scaled, dimensionless variables with $\mu = \varepsilon$ and we compare computationally the behavior of an initial CB solitary wave as it evolves according to each of the two systems travelling over a bottom of smooth topography with a fixed shelf-like function $b(x)$ and a parameter β that varies from $\mathcal{O}(\varepsilon)$ to $\mathcal{O}(1)$, so that the bottom becomes steeper.

For this purpose we solve both systems with our fully discrete scheme using cubic splines with uniform mesh, $N = 2000$ and the RK4 with $M = 2N$ on a spatial interval of $[0, 140]$ with a CB solitary wave of amplitude 0.5 as initial condition. (We experimented with several values of $\varepsilon = \mu$ but the results were qualitatively similar, so we show in Figure 2.17 below only the case $\varepsilon = \mu = 0.05$.)

As $b(x)$ we took a fixed profile given by

$$b(x) = \begin{cases} 0, & x \in \left[0, L - \frac{3}{2}\right], \\ \frac{1}{2}\left(1 + \sin\left(\frac{\pi}{3}(x - L)\right)\right), & x \in \left[L - \frac{3}{2}, L + \frac{3}{2}\right], \\ 1, & x \in \left[L + \frac{3}{2}, 140\right], \end{cases} \tag{2.42}$$

with $L = 70$. Thus b is a C^1 nonnegative function that bridges 0 and 1 over an interval of length 3. As a result, the undisturbed water depth η_b will vary from 1 to a shelf of depth $1 - \beta$ smoothly over this interval. We consider three cases: $\beta = \varepsilon = 0.05$, $\beta = 0.4$, $\beta = 0.6$, and present the results of the evolution for $0 \leq t \leq 89$ in Figure 2.17. In Fig. 2.17(a), where $\beta = \varepsilon = 0.05$, there is, as expected, practically no difference between the two solitary waves that suffer only a very small perturbation due to the bottom. But for $\beta = \mathcal{O}(1)$, i.e. when the bottom is steeper, we observe in Figure 2.17(b) ($\beta = 0.4$) and 2.17(c) ($\beta = 0.6$) large differences in the solutions of the two systems. As it travels on the shelf the

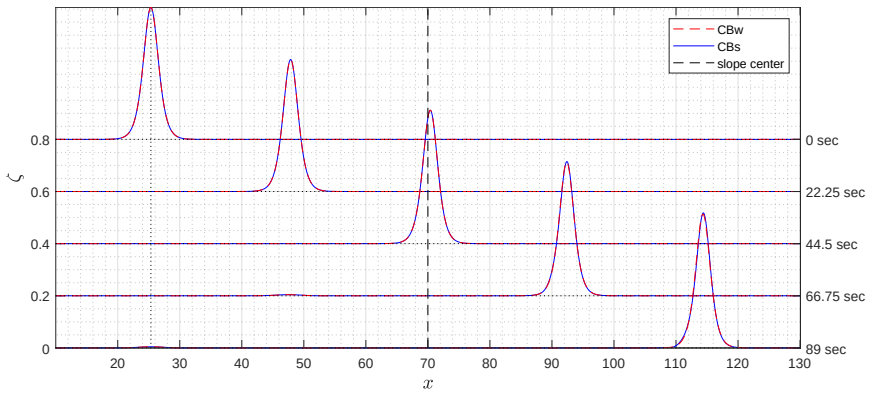
solitary wave evolves under (CBs) into a sequence of solitary waves as expected, whilst no such resolution is discernible in the case of the evolution under (CBw) at least for the time frame of this experiment. Both systems produce he same small-amplitude reflection waves. Our conclusion is that for $\beta = \mathcal{O}(1)$ (CBw) does not seem to give the correct longer-time behaviour of solutions in the case of strongly varying bottoms.

2.3.3.5 Comparison of (CBs) with the Serre-Green-Naghdi system

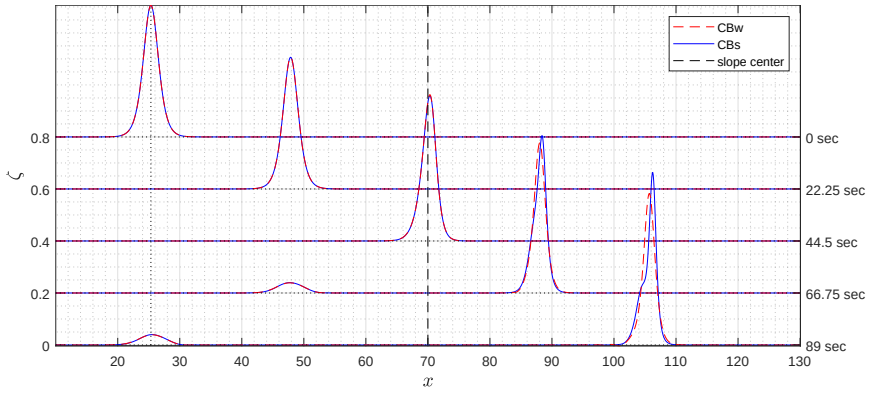
Finally, we compare by means of numerical experiment, the evolution of an initial solitary wave as it climbs a sloping bed, and as it is reflected by a vertical wall at the end of a slope. Recall from Chapter 1 that the system of Serre-Green-Naghdi (SGN) equations models two-way propagation of long dispersive waves (i.e. for which $\mu \ll 1$) without the assumption of small amplitude, i.e. with no restriction of ε, and that (CBs) is obtained from the (SGN) system with variable bottom under the Boussinesq scaling $\varepsilon = \mathcal{O}(\mu)$, [LB09]. The SGN system has been used in many computations, cf. e.g. [CBB07], [Bon+11], [MSM17], and their references, that agree quite well with experimental results of long-wave propagation over variable bottoms. In [ADM17], the authors analyzed Galerkin-finite element methods for (SGN) on a horizontal bottom (i.e. for the Serre equations) and shown optimal-order, L^2-error estimates in the case of periodic splines ($r \geq 3$) on uniform meshes.

Our aim in this subsection is to compare the results of numerical simulations of two test problems with (CBs), computed with our code, with numerical results for (SGN) obtained by Mitsotakis *et al*. in [MSM17]. The spatial semidiscretization used in [MSM17] is based on a modified Galerkin finite element scheme that uses a projection of a term containing a second-order derivative in SGN so that the scheme is also well defined for piecewise linear continuous elements (i.e. for $r = 2$) as well. In what follows we will solve numerically (CBs) using cubic splines on a uniform mesh with $N = 2000$ and RK4 time stepping with $M = 2N$. All variables for this experiment are nondimensional and unscaled with $\varepsilon = \mu = 1$.

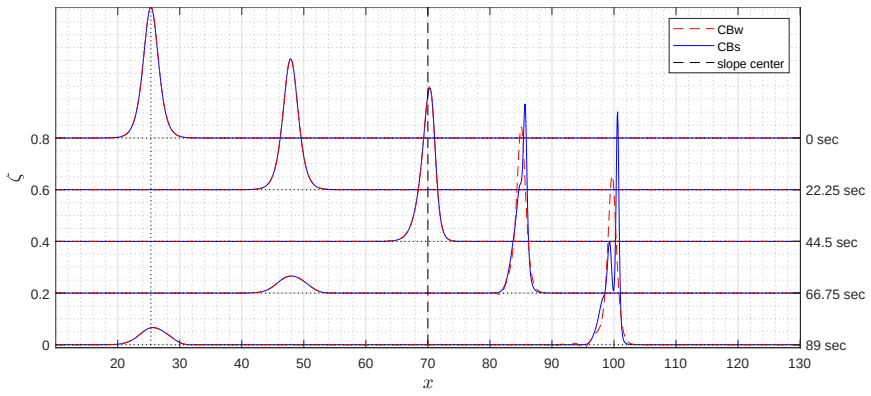
In the first experiment (shoaling of a solitary wave) we consider the variable-bottom example in §4.1 of [MSM17]. The geometry, in our notation, consists of a channel in the interval $[0, 84]$. The bottom is horizontal at a depth equal to -1 for $0 \leq x \leq x_{\rm B} = 50$, and upsloping with slope $\alpha = 1/35$ up to $x = 84$ where the water depth is equal to $1/35$. The initial condition is a solitary wave of the form (2.38), (2.41) of amplitude $a_0 = 0.2$ with crest at $x_0 = 29.8829$. The evolution of the numerical solution is monitored at ten gauges numbered 0, 1, …, 9, and located, respectively, at $x = 45,\ 70.96,\ 72.55,\ 73.68,\ 74.68$, and 76.91. In this experiment the variables are dimensionless and unscaled with $\varepsilon = \mu = 1$. In the experimental data and the (SGN) computations g was equal to 1. The temporal evolution under (CBs) is shown in Figure 2.18. Beyond gauge No. 9 the water becomes very shallow and the (CBs) model is certainly invalid. In Figure 2.19 we show the elevation of the wave at gauge 0 (at $x = x_{\rm B} - 5 = 45$, i.e. on the left of the toe of the slope), as a function of t. The three graphs shown correspond to the



(a) $\beta = \varepsilon = 0.05$



(b) $\beta = 0.4$



(c) $\beta = 0.6$

Figure 2.17: Comparison of evolution of a solitary wave under (CBw) and (CBs) over a bottom of varying steepness: $\eta_b = 1 - \beta b(x)$, $b(x)$ given by (2.42), $\varepsilon = \mu = 0.05$.

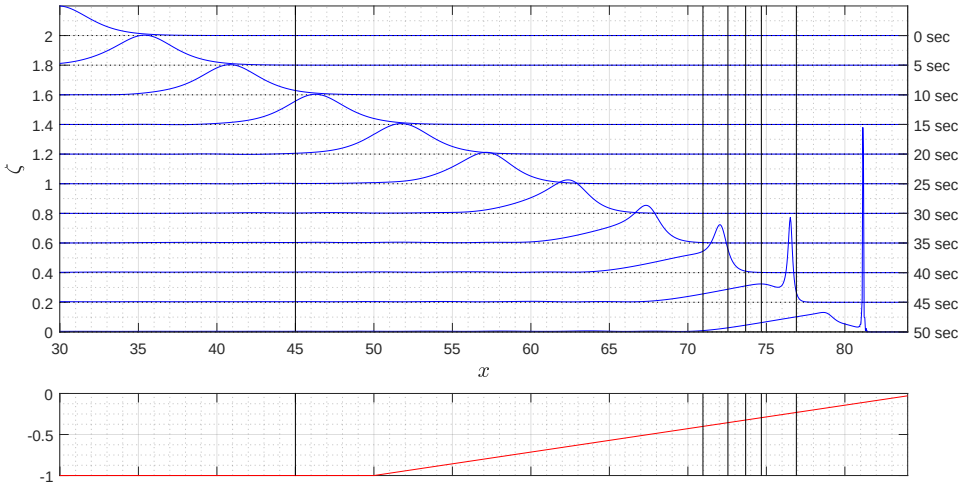


Figure 2.18: Initial condition and evolution of the solitary wave on a plain beach of slope $1 : 35$. Vertical lines depict the location of gauges 0, 1, 3, 5, 7, 9. Bottom is shown in the lower graph.

numerical solutions of (CBs) and (SGN), and to experimental data for this problem due to Grilli et al. [Gri+94], and are all in satisfactory agreement. Figure 2.20 shows the corresponding graphs of the elevation of the wave as a function of time recorded at gauges 1, 3, 5, 7, and 9 on the sloping bed. The numerical solution of (SGN) is in good agreement with the experimental data of [Gri+94]. As the wave climbs up the slope the (CBs) solution grows to a higher amplitude, whose ratio to the amplitude of the (SGN) wave increases monotonically from 1.14 for gauge 1 to 1.49 for gauge 9.

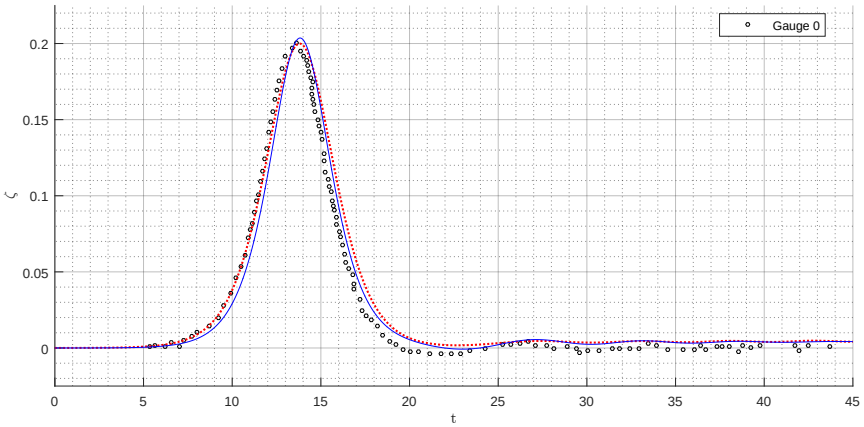


Figure 2.19: Elevation of wave at $x = x_{\mathrm{B}} - 5 = 45$ as a function of time. Markers show the experimental data, [Gri+94], dotted lines the numerical solution of (SGN), [MSM17], and solid lines the numerical solution of (CBs).

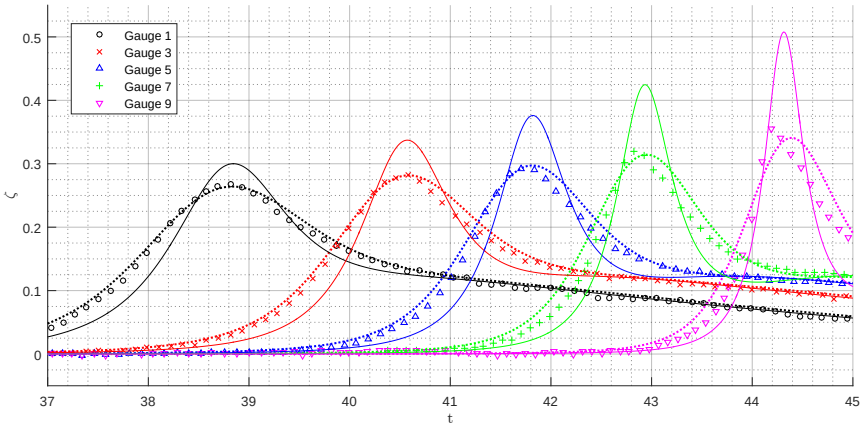


Figure 2.20: Elevation of wave at various gauges as a function of time for the shoaling on a beach of slope $1:35$ of a solitary wave with $a_0 = 0.12$. Markers show experimental data, [Gri+94], dotted lines the numerical solution of (SGN), [MSM17], and solid lines the numerical solution of (CBs).

For the second numerical experiment (shoaling and reflection of a solitary wave from a vertical wall at the end of the sloping beach), we consider a benchmark problem, cf. e.g. [MSM17], [WB99], [Dod98], [CBB07], [Bon+11], among other, that we solve numerically with our code of (CBs) and compare the results with those found by the numerical integration of (SGN) in Section 4.3 of [MSM17], and with experimental data due to Dodd, [Dod98]. The setup consists of a channel of length $[0, 70]$, initially horizontal at a depth of $h_0 = 0.7$, a sloping bed of uniform slope $1:50$ that starts rising at $x_{\rm B} = 50$ and ends at $x = 70$, where a vertical wall is placed. (This is shown in the lower graph of Figure 2.21.) We consider two solitary waves of the form (2.38), (2.41) (suitably modified so that the horizontal part of the waveguide has now a depth of $h_0 = 0.7$) with amplitudes 0.07 and 0.12 and crest initially located at $x = 20$. We solve the problem numerically with our code for (CBs) with a wall boundary condition $u = 0$ using cubic splines, $N = 2000$, $M = 2N$. All variables for this experiment are dimensional, x and η are measured in meters and t in seconds. The parameters ε and μ are equal to 1. The value of the gravitational acceleration constant is $g = 9.80665\,\mathrm{m/s^2}$ (standard gravity).

In Figure 2.21 we show snapshots every 3 secs of the (CBs)-free surface elevation as a function of x as the wave (of initial amplitude $a_0 = 0.07$) climbs up the slope and is reflected by the wall at $x = 70$ between $t = 15$ and $t = 18$. The reflected pulse apparently consists of a leading pulse followed by a dispersive tail. This wave travels downslope, and by $t = 30$ the leading pulse is located well within the horizontal-bottom region. The maximum runup at the wall was recorded to be equal to .1899.

In the (related) Figure 2.22 we show the temporal histories of the wave elevation $\zeta(x,t)$, generated by the solitary wave of amplitude $a_0 = 0.07$, at three gauges g_1,

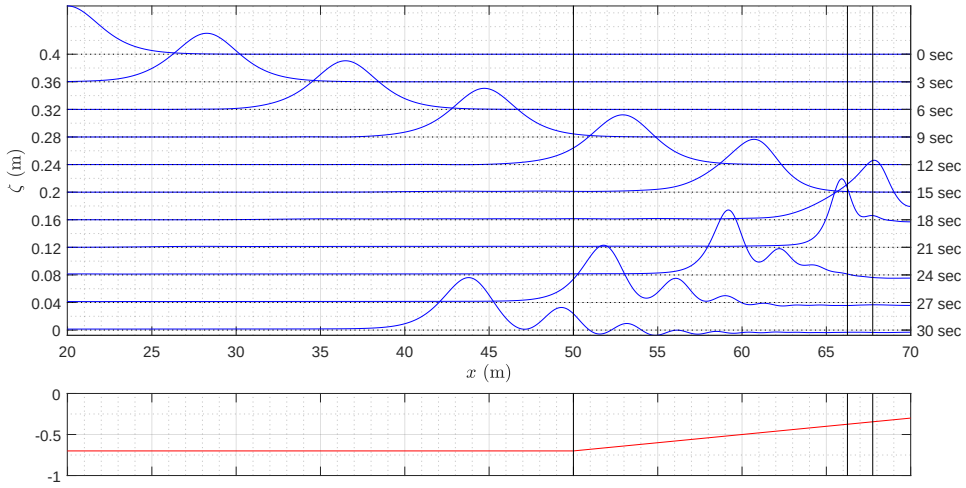


Figure 2.21: Evolution of the solitary wave of amplitude $a_0 = 0.07$ according to (CBs) on a beach of slope $1 : 50$, reflected on a vertical wall at $x = 70$. Vertical lines depict the location of gauges 1, 2 and 3. Bottom is shown in the lower graph.

g_2, g_3, located at $x = 50$, $x = 66.25$, and $x = 67.75$ (very close to the wall), respectively, computed by (CBs) and (SGN) (code of [MSM17]), in comparison with the experimental data of [Dod98] for this problem.

We observe that there is quite a good agreement between the three curves. The maximum amplitude of the reflected wave at gauge g_3 is found to be equal to .11080 for (CBs) and to .10280 for (SGN), giving a ratio of about 1.08.

Figure 2.23 depicts the analogous graphs in the case of the initial solitary wave of amplitude $a_0 = 0.12$. (Note the different scale of the ζ-axis.) This wave becomes steeper as it climbs up the slope; the reflected wave is of higher amplitude as well. The incident waves computed by the two models are quite close to each other and to the experimental data but the short-time behavior of the reflected pulse is somewhat different. For example, at g_3 the amplitude of the reflected (CBs) pulse is now equal to .2285 while the amplitude of the (SGN) reflected pulse is .1838 (giving a ratio of about 1.24), and there are phase and amplitude differences in the leading trailing oscillations. When the reflected wave has returned to the horizontal part of the channel (i.e. at g_1 in Figure 2.23 for $t \geq 25$) the agreement is much better and the ratio is now 0.98. The leading reflected pulse of the (SGN) solution is in satisfactory agreement with the data at all three gauges. The maximum runup at the wall of (CBs) for this amplitude was equal to .4012.

Our conclusion from the two numerical experiments in this subsection is that when the elevation wave steepens either while climbing up a sloping beach or after reflection from a vertical wall and close to the wall, the (CBs) solution overestimates that of the (SGN); the latter stays quite close to the available experimental data in the cases that we tried.

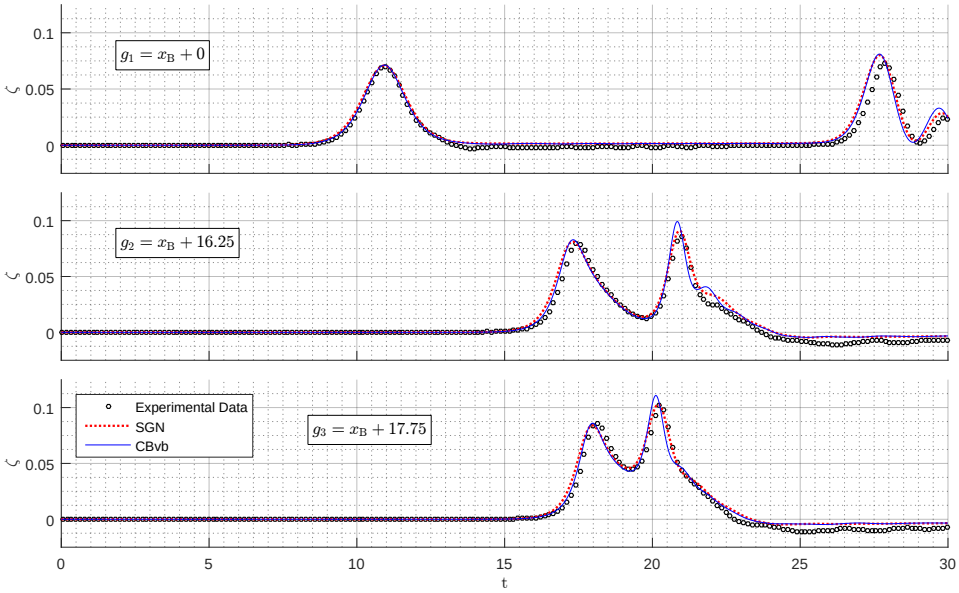


Figure 2.22: Reflection at a vertical wall located at $x = 70$ of a shoaling wave over a beach of slope $1 : 50$, with toe at $x_B = 50$. Initial solitary wave amplitude $a_0 = 0.07$.

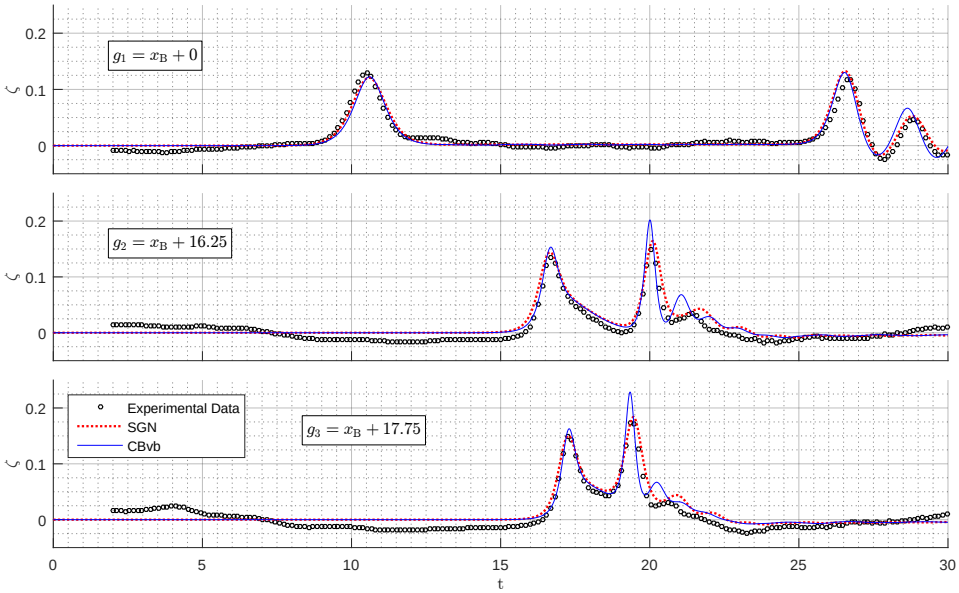


Figure 2.23: Reflection at a vertical wall located at $x = 70$ of a shoaling wave over a beach of slope $1 : 50$, with toe at $x_B = 50$. Initial solitary wave amplitude $a_0 = 0.12$.

Chapter 3

Standard Galerkin finite element methods for the numerical solution of the Shallow Water equations over variable bottom

3.1 Introduction

In this chapter we will consider standard Galerkin finite element approximations of the one-dimensional system of shallow water equations over a variable bottom that we write following [Per72], as

$$\begin{aligned} &\eta_t + (\eta u)_x + (\beta u)_x = 0, \\ &u_t + \eta_x + u u_x = 0. \end{aligned} \tag{SW}$$

As we saw in Chapter 1 the system (SW) approximates the two-dimensional Euler equations of water wave theory and models two-way propagation of long waves of finite amplitude on the surface of an ideal fluid in a channel with a variable bottom. The variables in (SW) are nondimensional and unscaled; $x \in \mathbb{R}$ and $t \geq 0$ are proportional to position along the channel and time, respectively. With the depth variable z taken to be positive upwards, the function $\eta = \eta(x,t)$ is proportional to the elevation of the free surface from a level of rest corresponding to $z = 0$ and $u = u(x,t)$ is proportional to the horizontal velocity of the fluid at the free surface. The bottom of the channel is defined by the function $z = -\beta(x)$; it will be assumed that $\beta(x) > 0$, $x \in \mathbb{R}$, and that the water depth $\eta(x,t) + \beta(x)$ is positive for all x, t. It should be noted that there are several equivalent formulations of the system represented by (SW), some of which will be considered in section 3.3.

It is well known that given smooth initial conditions $\eta(x,0) = \eta^0(x)$, $u(x,0) = u^0(x)$, $x \in \mathbb{R}$, and smooth bottom topography, the Cauchy problem for (SW) has smooth solutions, in general only locally in t. Here we will be concerned with

numerical approximations of (SW) and suppose that its solution is sufficiently smooth so that the error estimates of section 3.2 hold. We will specifically consider three initial-boundary-value problems (ibvp's) for (SW), posed on the spatial interval $[0,1]$: A simple ibvp with vanishing fluid velocity at the endpoints and two ibvp's with transparent (characteristic) boundary conditions, in the supercritical and subcritical flow cases, respectively. For these types of ibvp's there exists a well-posedness theory locally in t, cf. e.g. [PT11], [HPT11], [PT13]. For the formulation and numerical solution of ibvp's with transparent boundary conditions see also [Shi+11], [NHF08]. In section 3.2 we will specify in detail these ibvp's and summarize their well-posedness theory.

The literature on the numerical solution of the shallow water equations is vast. In recent years there has been considerable interest in solving them numerically by Discontinuous Galerkin finite element methods and refer the reader to chapter 4 of this thesis and [XZS10] and the recent surveys [QZ16], [Xin17], for an overview of issues related to the implementation of such methods in the presence of discontinuities.

In section 3.2 we consider the ibvp's previously mentioned, discretize them in space by the standard Galerkin finite element method, and prove L^2-error estimates for the semidiscrete approximations assuming smooth solutions of the equations and extending results of [AD16], [AD17], to the variable bottom case. In section 3.3 we discretize the semidiscrete problem in the temporal variable using the classical fourth-order accurate, four-stage explicit Runge-Kutta method. The resulting fully discrete scheme is stable under a Courant number stability condition and its convergence has been analyzed for (SW) in the case of a horizontal bottom in [ADK19]. We use this scheme in a series of numerical experiments simulating shallow water wave propagation over variable bottom topography and in the presence of absorbing (characteristic) boundary conditions up to the attainment of steady-state solutions. We also discuss issues of good balance, cf. [BV94], [XZS10], of the standard Galerkin method applied to the shallow water equations written in balance-law form.

A revised version of this chapter has appeared in the paper [KD19] written jointly with V. Dougalis. In this chapter we denote, for integer $m \geq 0$, by $H^m = H^m(0,1)$ the usual L^2-based real Sobolev spaces of order m, and by $\|\cdot\|_m$ their norm. The space $H_0^1 = H_0^1(0,1)$ will consist of the H^1 functions that vanish at $x = 0, 1$. The inner product and norm on $L^2 = L^2(0,1)$ will be denoted by $(\cdot,\cdot)$, $\|\cdot\|$, respectively, while C^m will be the m times continuously differentiable functions on $[0,1]$. The norms of L^∞ and of the L^∞-based Sobolev space $W^{1,\infty}$ on $(0,1)$ will be denoted by $\|\cdot\|_\infty$, $\|\cdot\|_{1,\infty}$, respectively. $\mathbb{P}_r$ will be the space of polynomials of degree at most r.

3.2 Initial-boundary-value problems and error estimates

In this section we will specify the initial-boundary-value problems (ibvp's) for the shallow water equations to be analyzed numerically, their Galerkin-finite element space discretizations and the properties of the attendant finite element spaces. We will then prove L^2-error estimates for these discretizations assuming that the data and the solutions of the ibvp's are smooth enough for the purposes of the error estimation.

3.2.1 Semidiscretization of a simple ibvp with vanishing fluid velocity at the endpoints

We consider first a simple ibvp for (SW) posed in the finite channel $[0,1]$. Let $T>0$ be given. We seek $\eta=\eta(x,t)$, $u=u(x,t)$, for $0\le x\le 1$, $0\le t\le T$, satisfying

$$\begin{aligned}&\eta_t+(\eta u)_x+(\beta u)_x=0,\\ &u_t+\eta_x+uu_x=0,\end{aligned}\qquad 0\le x\le 1,\ \ 0\le t\le T, \tag{3.1}$$
$$\eta(x,0)=\eta^0(x),\ \ u(x,0)=u^0(x),\ \ 0\le x\le 1,$$
$$u(0,t)=u(1,t)=0,\ \ 0\le t\le T.$$

In [PT11] Petcu and Temam, using an equivalent form of (3.1), established the existence-uniqueness of solutions (η,u) of (3.1) in $H^2\times H^2\cap H_0^1$ for some $T=T(\|\eta^0\|_2,\|u^0\|_2)$ under the hypotheses that $\eta^0\in H^2$, and, say, $\beta\in H^2$, such that $\eta^0(x)+\beta(x)>0$, $x\in[0,1]$, and $u^0\in H^2\cap H_0^1$. Moreover, they proved that $\eta(x,t)+\beta(x)>0$ for $(x,t)\in[0,1]\times[0,T]$, i.e. that the water depth is always positive. (This property will be assumed in all the error estimates to follow in addition to the sufficient smoothness of η and u.)

In order to solve (3.1) numerically let $0=x_1<x_2<\ldots<x_{N+1}=1$ be a quasiuniform partition of $[0,1]$ with $h:=\max_i(x_{i+1}-x_i)$, and for integers k, r such that $r\ge 2$, $0\le k\le r-2$, consider the finite element spaces $S_h=\{\varphi\in C^k:\ \varphi\big|_{[x_j,x_{j+1}]}\in\mathbb{P}_{r-1},\ 1\le j\le N\}$ and $S_{h,0}=\{\varphi\in S_h:\varphi(0)=\varphi(1)=0\}$. It is well known, see [Cia78], that given $w\in H^r$, there exists $\chi\in S_h$ such that

$$\|w-\chi\|+h\|w'-\chi'\|\le Ch^r\|w^{(r)}\|, \tag{3.2a}$$

and, in addition, if $r\ge 3$, such that

$$\|w-\chi\|_2\le Ch^{r-2}\|w^{(r)}\|, \tag{3.2b}$$

where C is a constant independent of h and w; a similar property holds in $S_{h,0}$ provided $w\in H^r\cap H_0^1$. It follows from (3.2a), cf. [DDW75], that if P is the L^2-projection operator onto S_h, then

$$\|\mathrm{P}\,w\|_1\le C\|w\|_1,\quad \forall w\in H^1, \tag{3.3a}$$
$$\|\mathrm{P}\,w\|_\infty\le C\|w\|_\infty,\quad \forall w\in C^0, \tag{3.3b}$$
$$\|\mathrm{P}\,w-w\|_{L^\infty}\le Ch^r\|w^{(r)}\|_\infty,\quad \forall w\in C^r, \tag{3.3c}$$

and that the analogous properties also hold for P_0, the L^2-projection operator onto $S_{h,0}$. In addition, as a consequence of the quasiuniformity of the mesh, cf. [Cia78], the inverse properties

$$\|\chi\|_1 \le Ch^{-1}\|\chi\|, \quad \|\chi\|_{j,\infty} \le Ch^{-(j+1/2)}\|\chi\|, \quad j = 0, 1, \tag{3.4}$$

hold for $\chi \in S_h$ or $\chi \in S_{h,0}$.

The standard Galerkin semidiscretization of (3.1) is defined as follows: Seek $\eta_h : [0,T] \to S_h$, $u_h : [0,T] \to S_{h,0}$, such that for $t \in [0,T]$

$$\begin{aligned} &(\eta_{ht}, \varphi) + ((\eta_h u_h)_x, \varphi) + ((\beta u_h)_x, \varphi) = 0, \quad \forall \varphi \in S_h, \\ &(u_{ht}, \chi) + (\eta_{hx}, \chi) + (u_h u_{hx}, \chi) = 0, \quad \forall \chi \in S_{h,0}, \end{aligned} \tag{3.5}$$

with initial conditions

$$\eta_h(0) = \mathrm{P}\,\eta_0, \quad u_h(0) = \mathrm{P}_0\, u_0. \tag{3.6}$$

We will prove below that the semidiscrete approximations (η_h, u_h) satisfy an L^2-error bound of $\mathcal{O}(h^{r-1})$. It is well known that this order of accuracy cannot be improved in the case of the standard Galerkin finite element method for first-order hyperbolic problems in the presence of general nonuniform meshes, [Dup73], [AD16]; for uniform meshes better results are possible, cf. [AD16] and the numerical experiments of section 3.3.

Proposition 3.1. *Let (η, u) be the solution of* (3.1)*, assumed to be sufficiently smooth and satisfying $\beta + \eta > 0$ in $[0,1] \times [0,T]$, where $\beta \in C^1$, $\beta > 0$. Let $r \ge 3$ and h be sufficiently small. Then, the semidiscrete ivp* (3.5)–(3.6) *has a unique solution (η_h, u_h) for $t \in [0,T]$, such that*

$$\max_{0\le t\le T} (\|\eta - \eta_h\| + \|u - u_h\|) \le Ch^{r-1}, \tag{3.7}$$

where, here and in the sequel, C will denote a generic constant independent of h.

Proof. As the proof is similar to that of Proposition 2.2 in [AD16], which is valid in the case of horizontal bottom ($\beta(x) = 1$), we will only indicate the steps where the two proofs differ. We let $\rho := \eta - \mathrm{P}\,\eta$, $\theta := \mathrm{P}\,\eta - \eta_h$, $\sigma := u - \mathrm{P}_0\, u$, $\xi := \mathrm{P}_0\, u - u_h$. While the solution exists we have

$$(\theta_t, \phi) + (\beta(\xi_x + \sigma_x), \phi) + (\beta_x(\xi + \sigma), \phi) + ((\eta u)_x - (\eta_h u_h)_x, \phi) = 0, \quad \forall \phi \in S_h, \tag{3.8}$$

$$(\xi_t, \chi) + (\theta_x + \rho_x, \chi) + (u u_x - u_h u_{hx}, \chi) = 0, \quad \forall \chi \in S_{h,0}. \tag{3.9}$$

Taking $\phi = \theta$ in (3.8) and integrating by parts we have

$$\begin{aligned} \tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}\|\theta\|^2 + ([(\beta+\eta)\xi]_x, \theta) = &-(\beta\sigma_x, \theta) - (\beta_x\sigma, \theta) - ((\eta\sigma)_x, \theta) - ((u\rho)_x, \theta) \\ &- ((u\theta)_x, \theta) + ((\rho\sigma)_x, \theta) + ((\theta\sigma)_x, \theta) + ((\rho\xi)_x, \theta) + ((\theta\xi)_x, \theta). \end{aligned} \tag{3.10}$$

In view of (3.6), we conclude by continuity that there exists a maximal temporal instance $t_h > 0$ such that (η_h, u_h) exist and $\|\xi_x\|_\infty \leq 1$ for $t \leq t_h$. Suppose that $t_h < T$. Using the approximation and inverse properties of S_h and $S_{h,0}$, we may then estimate the various terms on the r.h.s. of (3.10) for $t \in [0, t_h]$ in a similar way as in [AD16], since $\beta \in C^1$, and conclude that for $t \in [0, t_h]$

$$\tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}\|\theta\|^2 - (\gamma, \theta_x) \leq C(h^{r-1}\|\theta\| + \|\theta\|^2 + \|\xi\|^2), \tag{3.11}$$

where we have put $\gamma := (\beta + \eta)\xi$.

We turn now to (3.9) in which we take $\chi = \mathrm{P}_0\,\gamma = \mathrm{P}_0[(\beta+\eta)\xi]$. For $0 \leq t \leq t_h$ it follows that

$$\begin{aligned}(\xi_t, \gamma) + (\theta_x, \mathrm{P}_0\,\gamma) = &-(\rho_x, \mathrm{P}_0\,\gamma) - ((u\xi)_x, \mathrm{P}_0\,\gamma) - ((u\sigma)_x, \mathrm{P}_0\,\gamma) \\ &+ ((\sigma\xi)_x, \mathrm{P}_0\,\gamma) + (\sigma\sigma_x, \mathrm{P}_0\,\gamma) + (\xi\xi_x, \mathrm{P}_0\,\gamma). \end{aligned} \tag{3.12}$$

Arguing now as in [AD16], since $\beta \in C^1$, noting that

$$((u\xi)_x, \mathrm{P}_0\,\gamma) = ((u\xi)_x, \mathrm{P}_0\,\gamma - \gamma) + (u_x(\beta + \eta), \xi^2) - \tfrac{1}{2}([(\beta + \eta)u]_x, \xi^2),$$

and using a well-known *superapproximation* property of $S_{h,0}$ to estimate the term $\mathrm{P}_0\,\gamma - \gamma$:

$$\|\,\mathrm{P}_0\,\gamma - \gamma\| = \|\,\mathrm{P}_0[(\beta + \eta)\xi] - (\beta + \eta)\xi\| \leq Ch\|\xi\|,$$

we get

$$|((u\xi)_x, \mathrm{P}_0\,\gamma)| \leq Ch\|\xi\|_1\|\xi\| + C\|\xi\|^2 \leq C\|\xi\|^2.$$

With similar estimates as in [AD16], using the hypothesis that $\|\xi_x\|_\infty \leq 1$ for $0 \leq t \leq t_h$, we conclude from this inequality and (3.12) that for $0 \leq t \leq t_h$

$$(\xi_t, (\beta + \eta)\xi) + (\theta_x, \mathrm{P}_0\,\gamma) \leq C(h^{r-1}\|\xi\| + \|\xi\|^2). \tag{3.13}$$

Adding now (3.12) and (3.13) we obtain

$$\tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}\|\theta\|^2 + (\xi_t, (\beta + \eta)\xi) + (\theta_x, \mathrm{P}_0\,\gamma - \gamma) \leq C[h^{r-1}(\|\theta\| + \|\xi\|) + \|\theta\|^2 + \|\xi\|^2].$$

But, since $\beta = \beta(x)$, we have $(\xi_t, (\beta + \eta)\xi) = \tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}((\beta + \eta)\xi, \xi) - \tfrac{1}{2}(\eta_t\xi, \xi)$. Therefore, for $0 \leq t \leq t_h$

$$\tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}[\|\theta\|^2 + ((\beta + \eta)\xi, \xi)] \leq C[h^{r-1}(\|\theta\| + \|\xi\|) + \|\theta\|^2 + \|\xi\|^2],$$

for a constant C independent of h and t_h. Since $\beta + \eta > 0$, the norm $((\beta+\eta)\,\cdot, \cdot)^{1/2}$ is equivalent to that of L^2 uniformly for $t \in [0, T]$. Hence, Gronwall's inequality and (3.6) yield for a constant $C = C(T)$

$$\|\theta\| + \|\xi\| \leq Ch^{r-1} \quad \text{for} \quad 0 \leq t \leq t_h. \tag{3.14}$$

We conclude from (3.14), using inverse properties, that $\|\xi_x\|_\infty \leq Ch^{r-5/2}$ for $0 \leq t \leq t_h$, and, since $r \geq 3$, if h is taken sufficiently small, we see that t_h is not maximal. Hence we may take $t_h = T$ and (3.7) follows from (3.14). □

The hypothesis that $r \geq 3$ seems to be technical, as numerical experiments indicate that (3.7) apparently holds for $r = 2$ as well, cf. [AD16].

3.2.2 Semidiscretization of an ibvp with absorbing (characteristic) boundary conditions in the supercritical case

We consider now the shallow water equations with variable bottom with transparent (characteristic) boundary conditions. First we examine the *supercritical* case: For $(x,t) \in [0,1] \times [0,T]$ we seek $\eta = \eta(x,t)$ and $u = u(x,t)$ satisfying the ibvp

$$\begin{aligned} &\eta_t + (\beta u)_x + (\eta u)_x = 0, \\ &u_t + \eta_x + u u_x = 0, \end{aligned} \qquad 0 \le x \le 1, \quad 0 \le t \le T, \tag{3.15}$$

$$\begin{aligned} &\eta(x,0) = \eta^0(x), \quad u(x,0) = u^0(x), \quad 0 \le x \le 1, \\ &\eta(0,t) = \eta_0, \quad u(0,t) = u_0, \quad 0 \le t \le T, \end{aligned}$$

where $\beta \in C^1$, η^0, u^0 are given functions on $[0,1]$ and η_0, u_0 constants such that $\beta(x) + \eta_0 > 0$, $u_0 > 0$, $u_0 > \sqrt{\beta(x) + \eta_0}$, $x \in [0,1]$.

The ibvp (3.15) was studied by Huang et al., [HPT11], in the more general case of the presence of a lateral component of the horizontal velocity depending on x only (nonzero Coriolis parameter). In the simpler case of (3.15), we assume that (η_0, u_0) is a suitable constant solution of (3.15) and that $\eta^0(x)$, $u^0(x)$ are sufficiently smooth initial conditions close to (η_0, u_0) and satisfying appropriate compatibility relations at $x = 0$. Then, as is proved in [HPT11], given positive constants c_0, α_0, $\underline{\zeta}_0$, and $\overline{\zeta}_0$, there exist a $T > 0$ and a sufficiently smooth solution (η, u) of (3.15) satisfying for $(x,t) \in [0,1] \times [0,T]$ the strong supercriticality properties

$$u^2 - (\beta + \eta) \ge c_0^2, \tag{3.16a}$$

$$u \ge \alpha_0, \tag{3.16b}$$

$$\underline{\zeta}_0 \le (\beta + \eta) \le \overline{\zeta}_0. \tag{3.16c}$$

For the purposes of the error estimation to follow we will assume in addition that the solution of (3.15) satisfies a strengthened supercriticality condition of the following form: There exist positive constants a, and b, such that for $(x,t) \in [0,1] \times [0,T]$

$$\beta + \eta \ge b, \tag{3.17a}$$

$$u \ge 2a, \tag{3.17b}$$

$$\beta + \eta \le (u - a)(u - \tfrac{2a}{3}). \tag{3.17c}$$

Obviously (3.17a), (3.17b) and (3.17c) imply that $u \ge \sqrt{\beta + \eta}$. It is not hard to see that (3.17c) follows from (3.16a)–(3.16c) if e.g. α_0 is taken sufficiently small and c_0 sufficiently large. We also remark here that in the error estimates to follow, (3.17c) will be needed only at $x = 1$ for $t \in [0,T]$.

We will approximate the solution of (3.15) in a slightly transformed form. We let $\tilde{\eta} = \eta - \eta_0$, $\tilde{u} = u - u_0$ and rewrite (3.15) as an ibvp for $\tilde{\eta}$ and $\tilde{u}$ with homo-

geneous boundary conditions. Dropping the tildes we obtain the system

$$\begin{aligned} &\eta_t + u_0\eta_x + (\beta+\eta_0)u_x + (\eta u)_x + (u+u_0)\beta_x = 0, \\ &u_t + \eta_x + u_0 u_x + u u_x = 0, \end{aligned} \qquad 0 \le x \le 1, \quad 0 \le t \le T, \tag{3.18}$$

$$\begin{aligned} &\eta(x,0) = \eta^0(x) - \eta_0, \quad u(x,0) = u^0(x) - u_0, \quad 0 \le x \le 1, \\ &\eta(0,t) = 0, \quad u(0,t) = 0, \quad 0 \le t \le T. \end{aligned}$$

In terms of the new variables (3.17a)–(3.17c) become

$$\beta + \eta + \eta_0 \ge b, \tag{3.19a}$$

$$u + u_0 \ge 2a, \tag{3.19b}$$

$$\beta + \eta + \eta_0 \le (u + u_0 - a)(u + u_0 - \tfrac{2a}{3}). \tag{3.19c}$$

In the rest of this subsection, for integer $k \ge 0$, let $\overset{0}{C}{}^k = \{v \in C^k[0,1] : v(0) = 0\}$, and $\overset{0}{H}{}^{k+1} = \{v \in H^{k+1}(0,1) : v(0) = 0\}$. Using the hypotheses of section 3.2.1 on the finite element space discretization we define $\overset{0}{S}_h = \{\phi \in \overset{0}{C}{}^{r-2} : \phi\big|_{[x_j,x_{j+1}]} \in \mathbb{P}_{r-1}, 1 \le j \le N\}$ and P^0 the L^2 projection operator onto $\overset{0}{S}_h$. Note that (3.2)–(3.4) also hold on $\overset{0}{S}_h$ *mutatis mutandis*.

The standard Galerkin semidiscretization of (3.18) is defined as follows: We seek $\eta_h,\ u_h,\ : [0,T] \to \overset{0}{S}_h$ such that for $0 \le t \le T$

$$(\eta_{ht}, \phi) + (u_0\eta_{hx}, \phi) + ((\beta+\eta_0)u_{hx}, \phi) + ((\eta_h u_h)_x, \phi) + ((u_h+u_0)\beta_x, \phi) = 0, \quad \forall \phi \in \overset{0}{S}_h, \tag{3.20}$$

$$(u_{ht}, \phi) + (\eta_{hx}, \phi) + (u_0 u_{hx}, \phi) + (u_h u_{hx}, \phi) = 0, \quad \forall \phi \in \overset{0}{S}_h, \tag{3.21}$$

with

$$\eta_h(0) = \mathrm{P}^0(\eta^0(\cdot) - \eta_0), \quad u_h(0) = \mathrm{P}^0(u^0(\cdot) - u_0). \tag{3.22}$$

The boundary conditions implied by the choice of $\overset{0}{S}_h$ are no longer exactly transparent, but they are highly absorbing as will be seen in the numerical experiments of Section 3.3.

Proposition 3.2. *Let* (η, u) *be the solution of* (3.18)*, and assume that the hypotheses* (3.19a)–(3.19c) *hold, that* $r \ge 3$*, and* h *is sufficiently small. Then the semidiscrete ivp* (3.20)–(3.22) *has a unique solution* (η_h, u_h) *for* $0 \le t \le T$ *satisfying*

$$\max_{0\le t\le T}(\|\eta(t) - \eta_h(t)\| + \|u(t) - u_h(t)\|) \le Ch^{r-1}. \tag{3.23}$$

Proof. Let $\rho = \eta - \mathrm{P}^0\,\eta$, $\theta = \mathrm{P}^0\,\eta - \eta_h$, $\sigma = u - \mathrm{P}^0\,u$, $\xi = \mathrm{P}^0\,u - u_h$. After choosing a basis for $\overset{0}{S}_h$, it is straightforward to see that the semidiscrete problem represents an ivp for an ode system which has a unique solution locally in time.

While this solution exists, it follows from (3.20)–(3.22) and the pde's in (3.18), that

$$(\theta_t,\phi)+(u_0(\rho_x+\theta_x),\phi)+((\beta+\eta_0)(\sigma_x+\xi_x),\phi)+((\eta u-\eta_h u_h)_x,\phi)+ ((\sigma+\xi)\beta_x,\phi)=0,\quad \forall\phi\in \overset{0}{S}_h,$$
$$(\xi_t,\phi)+(\rho_x+\theta_x,\phi)+(u_0(\sigma_x+\xi_x),\phi)+(uu_x-u_hu_{hx},\phi)=0,\quad \forall\phi\in \overset{0}{S}_h$$

Proceeding as in the proof of Proposition 2.1 of [AD17], which is valid for a horizontal bottom, we obtain from the above in the case of variable bottom that

$$(\theta_t,\phi)+(u_0\theta_x,\phi)+(\gamma_x,\phi)+((u\theta)_x,\phi)-((\theta\xi)_x,\phi)=-(R_1,\phi),\quad \forall\phi\in \overset{0}{S}_h, \tag{3.24}$$
$$(\xi_t,\phi)+(\theta_x,\phi)+(u_0\xi_x,\phi)+((u\xi)_x,\phi)-(\xi\xi_x,\phi)=-(R_2,\phi),\quad \forall\phi\in \overset{0}{S}_h, \tag{3.25}$$

where $\gamma=(\beta+\eta_0+\eta)\xi$ and

$$R_1=u_0\rho_x+(\beta+\eta_0)\sigma_x+\sigma\beta_x+(\eta\sigma)_x+(u\rho)_x-(\rho\sigma)_x-(\rho\xi)_x-(\theta\sigma)_x, \tag{3.26}$$
$$R_2=\rho_x+u_0\sigma_x+(u\sigma)_x-(\sigma\xi)_x-\sigma\sigma_x. \tag{3.27}$$

Putting $\phi=\theta$ in (3.24), using integration by parts, and suppressing the dependence on t we have

$$\tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}\|\theta\|^2-(\gamma,\theta_x)+\tfrac{1}{2}(u_0+u(1))\theta^2(1)+(\beta(1)+\eta_0+\eta(1))\xi(1)\theta(1) -\tfrac{1}{2}\xi(1)\theta^2(1)=-\tfrac{1}{2}(u_x\theta,\theta)+\tfrac{1}{2}(\xi_x\theta,\theta)-(R_1,\theta). \tag{3.28}$$

Take now $\phi=\mathrm{P}^0\gamma=\mathrm{P}^0[(\beta+\eta_0+\eta)\xi]$ in (3.25) and get

$$(\xi_t,\gamma)+(\theta_x,\gamma)+(u_0\xi_x,\gamma)+((u\xi)_x,\gamma)-(\xi\xi_x,\gamma)=-(R_3,\mathrm{P}^0\gamma-\gamma)-(R_2,\mathrm{P}^0\gamma), \tag{3.29}$$

where

$$R_3=\theta_x+u_0\xi_x+(u\xi)_x-\xi\xi_x. \tag{3.30}$$

Integration by parts in various terms in (3.29) gives

$$(\xi_t,\gamma)+(\theta_x,\gamma)+\tfrac{1}{2}(u_0+u(1))(\beta(1)+\eta_0+\eta(1))\xi^2(1)-\tfrac{1}{3}(\beta(1)+\eta_0+\eta(1))\xi^3(1) =(R_4,\xi)-(R_3,\mathrm{P}^0\gamma-\gamma)-(R_2,\mathrm{P}^0\gamma), \tag{3.31}$$

where

$$R_4=\tfrac{1}{2}u_0(\beta_x+\eta_x)\xi-\tfrac{1}{2}u_x(\beta+\eta_0+\eta)\xi+\tfrac{1}{2}u(\beta_x+\eta_x)\xi-\tfrac{1}{3}(\beta_x+\eta_x)\xi^2. \tag{3.32}$$

Adding now (3.28) and (3.31) we obtain

$$\tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}\left[\|\theta\|^2+((\beta+\eta_0+\eta)\xi,\xi)\right]+\omega=\tfrac{1}{2}(\eta_t\xi,\xi)-\tfrac{1}{2}(u_x\theta,\theta) +\tfrac{1}{2}(\xi_x\theta,\theta)-(R_1,\theta)+(R_4,\xi)-(R_3,\mathrm{P}^0\gamma-\gamma)-(R_2,\mathrm{P}^0\gamma), \tag{3.33}$$

where

$$\omega = \tfrac{1}{2}(u_0 + u(1))\theta^2(1) + \tfrac{1}{2}(u_0 + u(1))(\beta(1) + \eta_0 + \eta(1))\xi^2(1) + (\beta(1) + \eta_0 + \eta(1))\xi(1)\theta(1) - \tfrac{1}{2}\xi(1)\theta^2(1) - \tfrac{1}{3}(\beta(1) + \eta_0 + \eta(1))\xi^3(1). \tag{3.34}$$

In view of (3.22), by continuity we conclude that there exists a maximal temporal instance $t_h > 0$ such that (η_h, u_h) exist and $\|\xi_x\|_\infty \le a$ for $t \le t_h$. Suppose that $t_h < T$. Then, since $\|\xi\|_\infty \le \|\xi_x\|_\infty$, it follows from (3.34) that for $t \in [0, t_h]$

$$\omega \ge \tfrac{1}{2}(u_0 + u(1) - a)\theta^2(1) + \tfrac{1}{2}(\beta(1) + \eta_0 + \eta(1))\left(u_0 + u(1) - \tfrac{2a}{3}\right)\xi^2(1) + (\beta(1) + \eta_0 + \eta(1))\xi(1)\theta(1) = \tfrac{1}{2}(\theta(1), \xi(1))^T \begin{pmatrix} \mu & \lambda \\ \lambda & \lambda\nu \end{pmatrix} \begin{pmatrix} \theta(1) \\ \xi(1) \end{pmatrix}, \tag{3.35}$$

where $\mu = u_0 + u(1) - a$, $\lambda = \beta(1) + \eta_0 + \eta(1)$, $\nu = u_0 + u(1) - \frac{2a}{3}$. The hypotheses (3.19a)–(3.19b) give that $0 < \mu < \nu$, $\lambda > 0$. It is easy to see then that the matrix in (3.35) will be positive semidefinite precisely when (3.19c) holds. Hence, (3.35) implies that $\omega \ge 0$.

We now estimate the various terms on the right-hand side of (3.33) for $0 \le t \le t_h$. As in the proof of Proposition 2.1 of [AD17] adapted in the case of a variable $\beta(x) \in C^1$ and using an appropriate variable-β superapproximation property to estimate $\|\,\mathrm{P}^0\,\gamma - \gamma\|$, we finally obtain from (3.33) and the fact that $\omega \ge 0$, that for $0 \le t \le t_h$ it holds that

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\left[\|\theta\|^2 + ((\beta + \eta_0 + \eta)\xi, \xi)\right] \le Ch^{r-1}(\|\theta\| + \|\xi\|) + C(\|\theta\|^2 + \|\xi\|^2),$$

where C is a constant independent of h and t_h. By (3.19a) the norm $((\beta + \eta_0 + \eta)\cdot, \cdot)^{1/2}$ is equivalent to that of L^2 uniformly for $t \in [0, T]$. Hence, Gronwall's inequality and the fact that $\theta(0) = \xi(0) = 0$ yield for a constant $C = C(T)$

$$\|\theta\| + \|\xi\| \le Ch^{r-1} \quad \text{for} \quad 0 \le t \le t_h. \tag{3.36}$$

We conclude from the inverse properties that $\|\xi_x\|_\infty \le Ch^{r-5/2}$ for $0 \le t \le t_h$, and, since $r \ge 3$, if h is taken sufficiently small, t_h is not maximal. Hence we may take $t_h = T$ and (3.23) follows from (3.36). □

3.2.3 Semidiscretization in the case of absorbing (characteristic) boundary conditions in the subcritical case

We finally consider the shallow water equations with variable bottom in the presence of transparent (characteristic) boundary conditions in the *subcritical case*. In this case, instead of the variable η, we will use the *total height* of the water, $H = \beta + \eta$. For $(x, t) \in [0, 1] \times [0, T]$ we seek $H = H(x, t)$ and $u = u(x, t)$

satisfying the ibvp

$$\begin{aligned} &H_t + (Hu)_x = 0, \\ &u_t + H_x + uu_x = \beta_x, \end{aligned} \quad 0 \le x \le 1, \quad 0 \le t \le T, \tag{3.37}$$

$$\begin{aligned} &H(x,0) = H^0(x), \quad u(x,0) = u^0(x), \quad 0 \le x \le 1, \\ &u(0,t) + 2\sqrt{H(0,t)} = u_0 + 2\sqrt{H_0}, \quad 0 \le t \le T, \\ &u(1,t) - 2\sqrt{H(1,t)} = u_0 - 2\sqrt{H_0}, \quad 0 \le t \le T, \end{aligned}$$

where H^0, u^0 are given functions on $[0,1]$ and H_0, u_0 constants such that $H_0 > 0$ and $u_0^2 < H_0$.

Implicit in the formulation of the boundary conditions in (3.37) is that outside the spatial domain $[0,1]$ u and H are equal to constants u_0, H_0, respectively. The ibvp (3.37) in a slightly different but equivalent form was studied by Petcu and Temam, [PT13], under the hypotheses that for some constant $c_0 > 0$ it holds that $u_0^2 - H_0 \le -c_0^2$ and that the initial conditions $H^0(x)$ and $u^0(x)$ are sufficiently smooth and satisfy the condition $(u^0(x))^2 - H^0(x) \le -c_0^2$ and suitable compatibility relations at $x = 0$ and $x = 1$. Under these assumptions one may infer from the theory of [PT13] that there exists a $T > 0$ such that a sufficiently smooth solution (H, u) of (3.37) exists for $(x,t) \in [0,1] \times [0,T]$ with the properties that H is positive and the strong supercriticality condition

$$u^2 - H \le -c_0^2, \tag{3.38}$$

holds for $(x,t) \in [0,1] \times [0,T]$. Here we will assume that the solution satisfies a stronger subcriticality solution; specifically that for some constant $c_0 > 0$ it holds that

$$u_0 + \sqrt{H_0} \ge c_0, \quad u_0 - \sqrt{H_0} \le -c_0, \tag{3.39a}$$

and for $(x,t) \in [0,1] \times [0,T]$ that

$$u + \sqrt{H} \ge c_0, \quad u - \sqrt{H} \le -c_0. \tag{3.39b}$$

In this section we will approximate the solution of (3.37) after transforming the system in diagonal form. We write the system of pde's in (3.37) as

$$\begin{pmatrix} H_t \\ u_t \end{pmatrix} + A \begin{pmatrix} H_x \\ u_x \end{pmatrix} = \begin{pmatrix} 0 \\ \beta_x \end{pmatrix} \tag{3.40}$$

where $A = \begin{pmatrix} u & H \\ 1 & u \end{pmatrix}$. The matrix A has eigenvalues $\lambda_1 = u + \sqrt{H}$, $\lambda_2 = u - \sqrt{H}$, (note that (3.39b) implies that $\lambda_1 \ge c_0$ and $\lambda_2 \le -c_0$ in $[0,1] \times [0,T]$), with associated eigenvectors $X_1 = \left(\sqrt{H}, 1\right)^{\mathrm{T}}$, $X_2 = \left(-\sqrt{H}, 1\right)^{\mathrm{T}}$. If S is the matrix with columns $X_1,\ X_2$ it follows from (3.40) that

$$S^{-1} \begin{pmatrix} H_t \\ u_t \end{pmatrix} + \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} S^{-1} \begin{pmatrix} H_x \\ u_x \end{pmatrix} = S^{-1} \begin{pmatrix} 0 \\ \beta_x \end{pmatrix}. \tag{3.41}$$

If we try to define now functions $v,\ w$ on $[0,1]\times[0,T]$ by the equations $S^{-1}\begin{pmatrix}H_t\\u_t\end{pmatrix}=\begin{pmatrix}v_t\\w_t\end{pmatrix}$, $S^{-1}\begin{pmatrix}H_x\\u_x\end{pmatrix}=\begin{pmatrix}v_x\\w_x\end{pmatrix}$, we see that these equations are consistent and their solutions are given by $v=\frac{1}{2}u+\sqrt{H}+c_v$, $w=\frac{1}{2}u-\sqrt{H}+c_w$, for arbitrary constants $c_v,\ c_w$. Choosing the constants $c_v,\ c_w$ so that $v(0,t)=0$, $w(1,t)=0$, and using the boundary conditions in (3.37) we get

$$v=\tfrac{1}{2}\big[u-u_0+2\big(\sqrt{H}-\delta_0\big)\big],\quad w=\tfrac{1}{2}\big[u-u_0-2\big(\sqrt{H}-\delta_0\big)\big] \tag{3.42}$$

where $\delta_0=\sqrt{H_0}$. The original variables $H,\ u$ are given in terms of v and w by the formulas

$$H=(\tfrac{1}{2}(v-w)+\delta_0)^2,\quad u=v+w+u_0 \tag{3.43}$$

Since

$$\lambda_1=u+\sqrt{H}=u_0+\delta_0+\frac{3v+w}{2},\quad \lambda_2=u-\sqrt{H}=u_0-\delta_0+\frac{v+3w}{2} \tag{3.44}$$

we see that the ibvp (3.37) becomes

$$\begin{pmatrix}v_t\\w_t\end{pmatrix}+\begin{pmatrix}u_0+\delta_0+\frac{3v+w}{2}&0\\0&u_0-\delta_0+\frac{v+3w}{2}\end{pmatrix}\begin{pmatrix}v_x\\w_x\end{pmatrix}=\tfrac{1}{2}\beta_x\begin{pmatrix}1\\1\end{pmatrix},$$
$$0\le x\le 1,\ \ 0\le t\le T. \tag{3.45}$$
$$v(x,0)=v^0(x),\quad w(x,0)=w^0(x),\quad 0\le x\le 1,$$
$$v(0,t)=0,\quad w(1,t)=0,\quad 0\le t\le T,$$

where $v^0(x)=\frac{1}{2}[u^0(x)-u_0+2(\sqrt{H^0(x)}-\delta_0)]$, $w^0(x)=\frac{1}{2}[u^0(x)-u_0-2(\sqrt{H^0(x)}-\delta_0)]$. Under our hypotheses (3.45) has a unique solution in $[0,1]\times[0,T]$ which will be assumed to be smooth enough for the purposes of the error estimation that follows.

Given a quasiuniform partition of $[0,1]$ as in section 3.2.1, in addition to the spaces defined there, let for integer $k\ge 0$ $\overset{0}{\mathcal{C}}{}^k=\{f\in C^k[0,1]: f(1)=0\}$, $\overset{0}{\mathcal{H}}{}^{k+1}=\{f\in H^{k+1}(0,1), f(1)=0\}$, and, for integer $r\ge 2$, $\overset{0}{\mathcal{S}}{}^0_h=\{\phi\in\overset{0}{\mathcal{C}}{}^{r-2}: \phi\big|_{[x_j,x_{j+1}]}\in\mathbb{P}_{r-1}, 1\le j\le N\}$. Note that the analogs of the approximation and inverse properties (3.2), (3.4) hold for $\overset{0}{\mathcal{S}}_h$ as well, and that the estimates in (3.3) are also valid for the L^2 projection P^1 onto $\overset{0}{\mathcal{S}}_h$, *mutatis mutandis*. The (standard) Galerkin semidiscretization of (3.45) is then defined as follows: Seek $v_h:[0,T]\to\overset{0}{S}_h,\ w_h:[0,T]\to\overset{0}{\mathcal{S}}_h$, such that for $t\in[0,T]$

$$(v_{ht},\phi)+((u_0+\delta_0)v_{hx},\phi)+\tfrac{3}{2}(v_hv_{hx},\phi)+\tfrac{1}{2}(w_hv_{hx},\phi)=\tfrac{1}{2}(\beta_x,\phi),\ \forall\phi\in\overset{0}{S}_h, \tag{3.46}$$

$$(w_{ht},\chi)+((u_0-\delta_0)w_{hx},\chi)+\tfrac{3}{2}(w_hw_{hx},\chi)+\tfrac{1}{2}(v_hw_{hx},\chi)=\tfrac{1}{2}(\beta_x,\chi),\ \forall\chi\in\overset{0}{\mathcal{S}}_h, \tag{3.47}$$

with

$$v_h(0)=\mathrm{P}^0(v^0),\quad w_h(0)=\mathrm{P}^1(w^0). \tag{3.48}$$

The boundary conditions induced by the finite element spaces and the discrete variational formulation (3.46)–(3.48) are no longer exactly transparent; they are highly absorbent nevertheless as will be checked in numerical experiments in Section 3.3. The main result of this section is

Proposition 3.3. *Let* (v, w) *be the solution of* (3.45) *and assume that the hypotheses* (3.39a)–(3.39b) *hold, that* $r \geq 3$*, and that* h *is sufficiently small. Then the semidiscrete ivp* (3.46)–(3.48) *has a unique solution* (v_h, w_h) *for* $0 \leq t \leq T$ *that satisfies*

$$\max_{0\leq t\leq T} (\|v - v_h\| + \|w - w_h\|) \leq Ch^{r-1}. \tag{3.49}$$

If (H, u) *is the solution of* (3.39) *and we define*

$$H_h = [\tfrac{1}{2}(v_h - w_h) + \delta_0]^2, \quad u_h = v_h + w_h + u_0, \tag{3.50}$$

then

$$\max_{0\leq t\leq T} (\|H - H_h\| + \|u - u_h\|) \leq Ch^{r-1}. \tag{3.51}$$

Proof. Let $\rho = v - \mathrm{P}^0 v$, $\theta = \mathrm{P}^0 v - v_h$, $\sigma = w - \mathrm{P}^1 w$, $\xi = \mathrm{P}^1 w - w_h$. After choosing bases for $\overset{0}{S}_h$ and $\overset{0}{S}_h$ we see that the ode ivp (3.46)–(3.48) has a unique solution locally in time. From (3.45) and (3.46), (3.47) we obtain, as long as the solution exists,

$$\begin{aligned}(\theta_t, \phi) + ((u_0 + \delta_0)(\theta_x + \rho_x), \phi) + \tfrac{3}{2}(vv_x - v_h v_{hx}, \phi) \\ + \tfrac{1}{2}(wv_x - w_h v_{hx}, \phi) = 0, \quad \forall \phi \in \overset{0}{S}_h,\end{aligned} \tag{3.52}$$

$$\begin{aligned}(\xi_t, \chi) + ((u_0 - \delta_0)(\sigma_x + \xi_x), \chi) + \tfrac{3}{2}(ww_x - w_h w_{hx}, \chi) \\ + \tfrac{1}{2}(vw_x - v_h w_{hx}, \chi) = 0, \quad \forall \chi \in \overset{0}{S}_h,\end{aligned} \tag{3.53}$$

Now, since

$$\begin{aligned}vv_x - v_h v_{hx} &= (v\rho)_x + (v\theta)_x - (\rho\theta)_x - \rho\rho_x - \theta\theta_x, \\ wv_x - w_h v_{hx} &= w(\rho_x + \theta_x) + vx(\sigma + \xi) - (\rho_x + \theta_x)(\sigma + \xi), \\ ww_x - w_h w_{hx} &= (w\sigma)_x + (w\xi)_x - \sigma\sigma_x - \xi\xi_x, \\ vw_x - v_h w_{hx} &= v(\sigma_x - \xi_x) + w_x(\rho + \theta) - (\sigma_x + \xi_x)(\rho + \theta).\end{aligned}$$

it follows that

$$vv_x - v_h v_{hx} = (v\theta)_x - (\theta\theta)_x + R_{11}, \quad wv_x - w_h v_{hx} = -\theta_x \xi + R_{12}, \tag{3.54}$$

$$ww_x - w_h w_{hx} = (w\xi)_x - \xi\xi_x + R_{21}, \quad vw_x - v_h w_{hx} = -\xi_x \theta + R_{22}, \tag{3.55}$$

where

$$R_{11} = (v\rho)_x - (\rho\theta)_x - \rho\rho_x, \quad R_{12} = w\rho_x + w\theta_x + v_x\sigma + v_x\xi - \rho_x\sigma - \rho_x\xi - \theta_x\sigma, \tag{3.56}$$

$$R_{21} = (w\sigma)_x - (\sigma\xi)_x - \sigma\sigma_x, \; R_{22} = v\sigma_x + v\xi_x + w_x\rho + w_x\theta - \sigma_x\rho - \sigma_x\theta - \xi_x\rho. \tag{3.57}$$

Putting now $\phi = \theta$ in (3.52) and $\chi = \xi$ in (3.53) we obtain

$$\begin{aligned}&\tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}\|\theta\|^2 + ((u_0+\delta_0)\theta_x,\theta) + \tfrac{3}{2}((v\theta)_x,\theta) - \tfrac{3}{2}(\theta\theta_x,\theta)\\ &\qquad = -((u_0+\delta_0)\rho_x,\theta) - \tfrac{3}{2}(R_{11},\theta) + \tfrac{1}{2}(\theta_x\xi,\theta) - \tfrac{1}{2}(R_{12},\theta)\end{aligned} \tag{3.58}$$

$$\begin{aligned}&\tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}\|\xi\|^2 + ((u_0-\delta_0)\xi_x,\xi) + \tfrac{3}{2}((w\xi)_x,\xi) - \tfrac{3}{2}(\xi\xi_x,\xi),\\ &\qquad = -((u_0-\delta_0)\sigma_x,\xi) - \tfrac{3}{2}(R_{21},\xi) + \tfrac{1}{2}(\xi_x\theta,\xi) - \tfrac{1}{2}(R_{22},\theta).\end{aligned} \tag{3.59}$$

Integration by parts yields (we suppress the t-dependence)

$$\begin{gathered}((u_0+\delta_0)\theta_x,\theta) = \frac{u_0+\delta_0}{2}\theta^2(1), \quad ((u\theta)_x,\theta) = \frac{1}{2}(v_x\theta,\theta) + \frac{1}{2}v(1)\theta^2(1),\\ (\theta\theta_x,\theta) = \frac{1}{3}\theta^2(1), \quad ((u_0-\delta_0)\xi_x,\xi) = -\frac{u_0-\delta_0}{2}\xi^2(0),\\ ((w\xi)_x,\xi) = \frac{1}{2}(w_x\xi,\xi) - \frac{1}{2}w(0)\xi^2(0), \quad (\xi\xi_x,\xi) = -\frac{1}{3}\xi^3(0).\end{gathered}$$

Hence, (3.58) becomes

$$\begin{aligned}&\tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}\|\theta\|^2 + \tfrac{1}{2}\theta^2(1)(u_0+\delta_0+\tfrac{3}{2}v(1)-\theta(1)) =\\ &\qquad -((u_0+\delta_0)\rho_x,\theta) - \tfrac{3}{4}(v_x\theta,\theta) + \tfrac{1}{2}(\theta_x\xi,\theta) - \tfrac{3}{2}(R_{11},\theta) - \tfrac{1}{2}(R_{12},\theta).\end{aligned}$$

By (3.39b) and (3.44) we see that $u_0+\delta_0+\frac{3}{2}v(1) \geq c_0 > 0$. Therefore the above equation gives

$$\begin{aligned}\tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}\|\theta\|^2 + \tfrac{1}{2}(c_0-\theta(1))\theta^2(1) &\leq -((u_0+\delta_0)\rho_x,\theta)\\ &\quad - \tfrac{3}{4}(v_x\theta,\theta) + \tfrac{1}{2}(\theta_x\xi,\theta) - \tfrac{3}{2}(R_{11},\theta) - \tfrac{1}{2}(R_{12},\theta).\end{aligned} \tag{3.60}$$

Similarly, for (3.59) we obtain

$$\begin{aligned}&\tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}\|\xi\|^2 + \tfrac{1}{2}\xi^2(0)(-(u_0-\delta_0+\tfrac{3}{2}w(0)) + \xi(0)) =\\ &\qquad -((u_0-\delta_0)\sigma_x,\xi) + \tfrac{1}{2}(\xi_x\theta,\xi) - \tfrac{3}{4}(w_x\xi,\xi) - \tfrac{3}{2}(R_{21},\xi) - \tfrac{1}{2}(R_{22},\xi).\end{aligned}$$

Again, by (3.39b) and (3.44) we get $u_0 - \delta_0 + \frac{3}{2}w(0) \leq -c_0 < 0$. We conclude that

$$\begin{aligned}\tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}\|\xi\|^2 + \tfrac{1}{2}(c_0+\xi(0))\xi^2(0) &\leq -((u_0-\delta_0)\sigma_x,\xi)\\ &\quad + \tfrac{1}{2}(\xi_x\theta,\xi) - \tfrac{3}{4}(w_x\xi,\xi) - \tfrac{3}{2}(R_{21},\xi) - \tfrac{1}{2}(R_{22},\xi).\end{aligned} \tag{3.61}$$

Finally, adding (3.60) and (3.61) we get, as long as the solution of (3.46)–(3.48) exists, that

$$\begin{aligned}&\tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}(\|\theta\|^2 + \|\xi\|^2) + \tfrac{1}{2}(c_0-\theta(1))\theta^2(1) + \tfrac{1}{2}(c_0+\xi(0))\xi^2(0)\\ &\qquad \leq -((u_0+\delta_0)\rho_x,\theta) - ((u_0-\delta_0)\sigma_x,\xi) - \tfrac{3}{4}(v_x\theta,\theta) - \tfrac{3}{4}(w_x\xi,\xi)\\ &\quad + \tfrac{1}{2}(\theta_x\xi,\xi) + \tfrac{1}{2}(\xi_x\theta,\xi) - \tfrac{3}{2}(R_{11},\theta) - \tfrac{1}{2}(R_{12},\theta) - \tfrac{3}{2}(R_{21},\xi) - \tfrac{1}{2}(R_{22},\xi).\end{aligned} \tag{3.62}$$

In view of (3.48), by continuity we conclude that there exists a maximal temporal instance $t_h > 0$ such that v_h, w_h exist for $t \leq t_h$ and

$$\|\theta(t)\|_{1,\infty} + \|\xi(t)\|_{1,\infty} \leq c_0, \quad t \in [0, t_h]. \tag{3.63}$$

Suppose that $t_h < T$. For $t \in [0, t_h]$ we have by (3.63)

$$\tfrac{1}{2}(c_0 - \theta(1))\theta^2(1) + \tfrac{1}{2}(c_0 + \xi(0))\xi^2(0) \geq 0, \tag{3.64}$$

and

$$\tfrac{1}{2}|(\theta_x\xi, \theta)| + \tfrac{1}{2}|(\xi_x\theta, \xi)| \leq \tfrac{1}{2}c_0\|\theta\|\|\xi\|. \tag{3.65}$$

We obviously have

$$|(v_x\theta, \theta)| + |(w_x\xi, \xi)| \leq C(\|\theta\|^2 + \|\xi\|^2). \tag{3.66}$$

Using now the approximation and inverse properties (3.2)–(3.4) for $\overset{0}{S}_h$ (and also for $\overset{0}{\mathcal{S}}_h$) we estimate the rest of the terms in the right-hand side of (3.62) as follows. We first clearly have

$$|((u_0 + \delta_0)\rho_x, \theta)| + |((u_0 - \delta_0)\sigma_x, \xi)| \leq Ch^{r-1}(\|\theta\| + \|\xi\|). \tag{3.67}$$

Integrating by parts we see by (3.56) that

$$\begin{aligned}(R_{11}, \theta) &= ((v\rho)_x, \theta) - ((\rho\theta)_x, \theta) - (\rho\rho_x, \theta)\\ &= v(1)\rho(1)\theta(1) - (v\rho, \theta_x) - \rho(1)\theta^2(1) + (\rho\theta, \theta_x) - (\rho\rho_x, \theta).\end{aligned}$$

Therefore

$$\begin{aligned}|(R_{11}, \theta)| &\leq C\|\rho\|_\infty\|\theta\|_\infty + C\|\rho\|_\infty\|\theta_x\|\\ &\quad + \|\rho\|_\infty\|\theta\|_\infty^2 + \|\rho\|_\infty\|\theta\|\|\theta_x\| + \|\rho\|_\infty\|\rho_x\|\|\theta\|\\ &\leq Ch^r\|\theta\|_\infty + Ch^r\|\theta_x\| + Ch^r\|\theta\|_\infty^2 + Ch^r\|\theta\|\|\theta_x\| + Ch^{2r-1}\|\theta\|\\ &\leq Ch^{r-1}(\|\theta\| + \|\theta\|^2).\end{aligned} \tag{3.68}$$

Integration by parts and (3.56) yield for the R_{12} term that

$$(R_{12}, \theta) = (w\rho_x, \theta) - \tfrac{1}{2}(w_x\theta, \theta) + (v_x\sigma, \theta) + (v_x\xi, \theta) - (\rho_x\sigma, \theta) - (\rho_x\xi, \theta) - (\theta_x\sigma, \theta).$$

Hence, similarly as above

$$\begin{aligned}|(R_{12}, \theta)| &\leq Ch^{r-1}\|\theta\| + C\|\theta\|^2 + Ch^r\|\theta\| + C\|\xi\|\|\theta\|\\ &\quad + Ch^{2r-1}\|\theta\| + Ch^{r-1}\|\xi\|_\infty\|\theta\| + Ch^r\|\theta\|_\infty\|\theta_x\|\\ &\leq Ch^{r-1}\|\theta\| + C\|\theta\|^2 + C\|\xi\|\|\theta\|.\end{aligned} \tag{3.69}$$

Again, using integration by parts and (3.56) for the R_{21} term, we obtain

$$(R_{21}, \xi) = -(w\sigma, \xi_x) - w(0)\sigma(0)\xi(0) + (\sigma\xi, \xi_x) + \sigma(0)\xi^2(0) - (\sigma\sigma_x, \xi).$$

Therefore

$$\begin{aligned}|(R_{21},\xi)| &\le C\|\sigma\|\|\xi_x\| + C\|\sigma\|_\infty\|\xi\|_\infty \\ &\quad + \|\sigma\|_\infty\|\xi\|\|\xi_x\| + \|\sigma\|_\infty\|\xi\|_\infty^2 + \|\sigma\|_\infty\|\sigma_x\|\|\xi\| \\ &\le Ch^r\|\xi_x\| + Ch^r\|\xi\|_\infty + Ch^r\|\xi\|\|\xi_x\| + Ch^r\|\xi\|_\infty^2 + Ch^{2r-1}\|\xi\| \\ &\le Ch^{r-1}(\|\xi\| + \|\xi\|^2).\end{aligned} \tag{3.70}$$

Finally, by (3.56) and integration by parts we have for the R_{22} term

$$(R_{22},\xi) = (v\sigma_x,\xi) - \tfrac{1}{2}(v_x\xi,\xi) + (w_x\rho,\xi) + (w_x\theta,\xi) - (\sigma_x\rho,\xi) - (\sigma_x\theta,\xi) - (\rho\xi_x,\xi).$$

Hence,

$$\begin{aligned}|(R_{22},\xi)| &\le C\|\sigma_x\|\|\xi\| + C\|\xi\|^2 + C\|\rho\|\|\xi\| + C\|\theta\|\|\xi\| \\ &\quad + \|\sigma_x\|\|\rho\|_\infty\|\xi\| + \|\sigma_x\|\|\theta\|_\infty\|\xi\| + \|\rho\|_\infty\|\xi_x\|\|\xi\| \\ &\le Ch^{r-1}\|\xi\| + C\|\xi\|^2 + Ch^r\|\xi\| + C\|\theta\|\|\xi\| + Ch^{2r-1}\|\xi\| \\ &\quad + Ch^{r-1}\|\theta\|_\infty\|\xi\| + Ch^r\|\xi_x\|\|\xi\| \\ &\le Ch^{r-1}\|\xi\| + C\|\xi\|^2 + C\|\theta\|\|\xi\|.\end{aligned} \tag{3.71}$$

By (3.62), taking into account (3.64)–(3.71) we see that

$$\tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}(\|\theta\|^2 + \|\xi\|^2) \le Ch^{r-1}(\|\theta\| + \|\xi\|) + C(\|\theta\|^2 + \|\xi\|^2), \quad t \in [0,t_h].$$

An application of Gronwall's Lemma and (3.48) yield

$$\|\theta(t)\| + \|\xi(t)\| \le Ch^{r-1}, \quad t \in [0,t_h], \tag{3.72}$$

from which by inverse assumptions it follows that $\|\theta\|_{1,\infty} + \|\xi\|_{1,\infty} \le Ch^{r-5/2}$ for $t \in [0,t_h]$. Since it was assumed that $r \ge 3$ this contradicts the maximality of t_h and (3.72) holds for $0 \le t \le T$. The estimate (3.49) follows. Since now $\|v - v_h\|_\infty \le \|\rho\|_\infty + \|\theta\|_\infty \le Ch^{r-3/2}$ and similarly $\|w - w_h\|_\infty \le Ch^{r-3/2}$, and since

$$H - H_h = [\delta_0 + \tfrac{1}{4}((v-w) + (v_h - w_h))][(v-w) - (v_h - w_h)],$$

we conclude that $\|H - H_h\| \le C(\|v - v_h\| + \|w - w_h\|) \le Ch^{r-1}$. Similarly $\|u - u_h\| \le \|v - v_h\| + \|w - w_h\| \le Ch^{r-1}$, and the proof of Proposition 3.3 is now complete. □

3.3 Numerical experiments

In this section we present results of numerical experiments that we performed solving numerically the shallow water equations using standard Galerkin finite element space discretizations like the ones analyzed in the previous section. The semidiscrete schemes were discretized in the temporal variable by the 'classical', explicit,

4-stage, 4$^{\text{th}}$-order Runge-Kutta scheme (RK4), unless otherwise indicated. The resulting fully discrete scheme is stable and fourth-order accurate in time provided a Courant-number stability condition of the form $\frac{k}{h} \leq \alpha$ is imposed; here k denotes the (uniform) time step. In the case of a horizontal bottom the convergence of this scheme for the ibvp (3.1) was analyzed in [ADK19] and used in numerical experiments for the absorbing b.c. ibvp's (3.15) and (3.37) in [AD17].

In section 3.3.1 below we use this fully discrete scheme to study computationally various issues related to the discretization of the ibvp's with absorbing (characteristic) b.c.'s considered in sections 3.2.2 and 3.2.3. In section 3.3.2 we write the shallow water equations in the form of a balance law and study various issues of the numerical solution of this model with Galerkin-finite element methods, including questions of 'good balance' of the schemes. Since the numerical method simulates only smooth solutions, initial conditions and bottom topographies were taken to be of small amplitude to ensure that no discontinuities developed within the time frame of the experiments.

3.3.1 Absorbing (characteristic) boundary conditions

In the numerical experiments of this section we use the standard Galerkin finite element method with continuous, piecewise linear functions for the space discretization of the numerical solution of the ibvp's with absorbing (characteristic) boundary conditions considered in sections 3.2.2 and 3.2.3. The theoretical error estimates in Propositions 3.2 and 3.3 require at least piecewise quadratic elements, i.e. $r \geq 3$, and predict L^2-error bounds of $\mathcal{O}(h^{r-1})$ for quasiuniform meshes. The results of numerical experiments shown in the sequel suggest that the method works with piecewise linear functions (i.e. $r = 2$) as well, and in this case the L^2 errors for a uniform mesh are of $\mathcal{O}(h^2)$.

In the *supercritical* case, in order to find the numerical convergence rates of the scheme (3.20)–(3.21) we consider an ibvp with $\eta_0 = 1$, $u_0 = 3$ and a bottom function and exact solution given for $x \in (0, 1)$ by

$$\begin{aligned} &\beta(x) = 1 - 0.04\exp(-100(x-0.5)^2), \\ &\eta(x,t) = x\exp(-xt) + \eta_0, \quad u(x,t) = (1 - x - \cos(\pi x))\exp(2t) + u_0. \end{aligned} \tag{3.73}$$

(The initial conditions and an appropriate right-hand side were computed from these formulas.) The problem was solved with a uniform mesh with $h = 1/N$ and $k = h/10$. The L^2 errors and rates of convergence at $T = 1$ are shown in Table 3.1.

In the case of a *subcritical* flow we consider an ibvp with $\eta_0 = 1$, $u_0 = 1$, and

	η		u	
N	L_2 error	rate	L_2 error	rate
40	1.3202e-03	-	6.1375e-03	-
80	3.2932e-04	2.003	1.5334e-03	2.001
160	8.2245e-05	2.001	3.8335e-04	2.000
320	2.0550e-05	2.001	9.5918e-05	1.999
640	5.1361e-06	2.000	2.4070e-05	1.995

Table 3.1: L^2 errors and rates of convergence at $T = 1$, $r = 2$, supercritical case, (3.73), $h = 1/N$, $k/h = 1/10$.

bottom function and exact solution given for $x \in (0, 1)$ by

$$\begin{aligned} &\beta(x) = 1 - 0.04\exp(-100(x - 0.5)^2), \\ &\eta(x, t) = (x + 1)\exp(-xt), \\ &u(x, t) = (2x + \cos(\pi x) - 1)\exp(t) + xA(t) + (1 - x)B(t), \end{aligned} \tag{3.74}$$

where

$$\begin{aligned} &A(t) = 2\sqrt{1 + \eta(1, t)} + u_0 - 2\sqrt{1 + \eta_0}, \\ &B(t) = -2\sqrt{1 + \eta(0, t)} + u_0 + 2\sqrt{1 + \eta_0}. \end{aligned}$$

(The initial conditions and an appropriate right-hand side were computed by these formulas). The problem was solved by the scheme (3.46)–(3.48), (3.50), with $h = 1/N$ and $k = h/10$. The L^2 errors and rates of convergence for the variables η and u at $T = 1$ are shown in Table 3.2.

	η		u	
N	L_2 error	rate	L_2 error	rate
40	7.8451e-03	-	4.7238e-03	-
80	1.9602e-03	2.001	1.2154e-03	1.959
160	4.8955e-04	2.001	3.0717e-04	1.984
320	1.2229e-04	2.001	7.7169e-05	1.993
640	3.0560e-05	2.001	1.9349e-05	1.996

Table 3.2: L^2 errors and rates of convergence at $T = 1$, $r = 2$, subcritical case, (3.74), $h = 1/N$, $k/h = 1/10$.

It is clear that Tables 3.1 and 3.2 suggest that the L^2 convergence rates are optimal in the case of piecewise linear elements on a uniform mesh.

In order to check further the accuracy of the numerical schemes we consider in the *supercritical* case a problem with a variable bottom having a single hump, and constant initial conditions on $(0, 1)$ given by

$$\begin{aligned} &\beta(x) = 1 - 0.4\exp(-100(x - 0.5)^2), \\ &\eta^0(x) = \eta_0 = 1, \quad u^0(x) = u_0 = 3, \end{aligned} \tag{3.75}$$

that we integrate numerically using $h = 1/400$, $k = h/3$. In Figure 3.1 we show some profiles of the temporal evolution of the numerical solution up to $t = 0.5$. The data given by (3.75) and the boundary conditions generate a wave moving to the right and sensing the effect of the variable bottom which is centered at $x = 0.5$. There are no spurious oscillations reflected from the boundary $x = 1$ as the wave exits. By $t = 0.5$ the solution has attained a *steady state* shown in (3.1(d)).

The steady state of such flows is straightforward to determine analytically. Its profile $\eta = \eta(x)$, $u = u(x)$ satisfies the equations

$$\begin{aligned} ((\beta + \eta)u)_x &= 0, \\ \left(\eta + \tfrac{1}{2}u^2\right)_x &= 0, \end{aligned} \tag{3.76}$$

from which using the boundary conditions at $x = 0$, we see that u is given in terms of η by

$$u = \frac{u_0\,(\eta_0 + \beta(0))}{\eta + \beta}, \tag{3.77a}$$

where η is the physically acceptable solution of the cubic equation

$$(\eta + \beta)^2\left(\eta - \eta_0 - \tfrac{1}{2}u_0^2\right) + \tfrac{1}{2}u_0^2\,(\eta_0 + \beta(0))^2 = 0. \tag{3.77b}$$

(For the analysis of the solutions of the steady-state problem, cf. [HK68]). We checked the ability of the code to preserve steady-state solutions by taking the profile computed analytically from (3.77) for this problem as initial condition and integrating up to $t = 0.6$. The difference between the final profile and the L^2 projection of the analytical initial condition was of $\mathcal{O}(10^{-9})$ in L^2 for both components when $h = 1/400$, $k = h/10$.

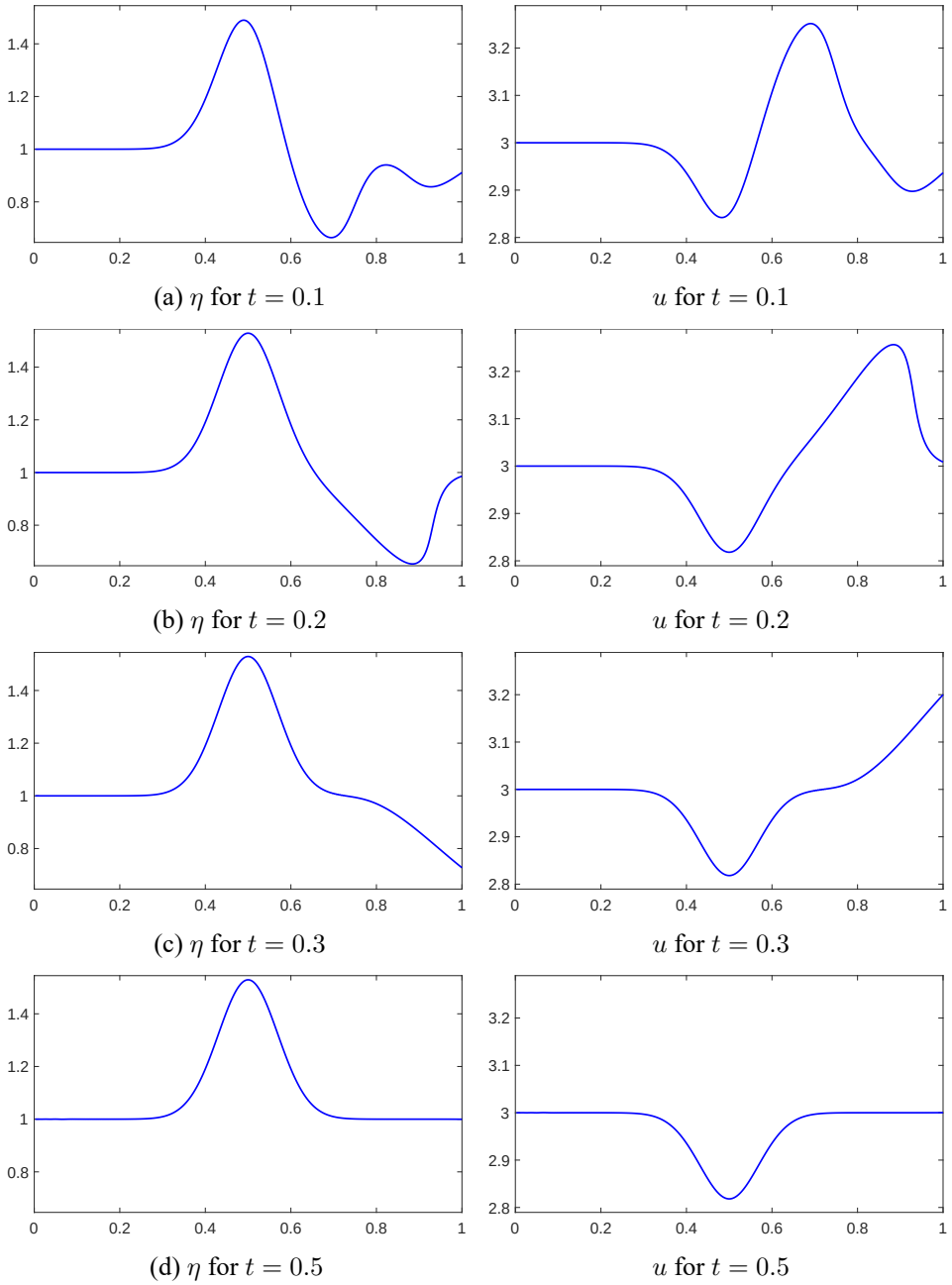
In Figure 3.2 we show instances of the temporal evolution up to the attainment of steady state (in (3.2(d))) of the supercritical flow generated with $h = 1/400$, $k = h/3$, by $\eta_0 = 1$, $u_0 = 3$ and bottom topography and initial conditions given on $[0, 1]$ by

$$\begin{aligned} \beta(x) &= 1 - 0.04\exp(-1000(x - 0.75)^2), \\ \eta^0(x) &= 0.05\exp(-400(x - 0.25)^2) + \eta_0, \\ u^0(x) &= 0.1\exp(-400(x - 0.25)^2) + u_0. \end{aligned} \tag{3.78}$$

The variable initial profile gives rise to a wavetrain that moves to the right, interacts with the bottom and exits without spurious oscillations leaving behind the steady state that depends only on η_0, u_0 and β.

We now present some analogous results in the *subcritical* case . We used the fully discrete scheme with spatial discretization given by (3.46)–(3.48), (3.50); the variables depicted in the figures are the approximations of η and u. The spatial discretization was effected on $[0, 1]$ with piecewise linear functions on a uniform mesh with $h = 1/2000$; the time-stepping procedure was RK4 as usual with $k = h/10$. In the first example we took $\eta_0 = 1$, $u_0 = 1$ and

$$\begin{aligned} \beta(x) &= 1 - 0.04\exp(-100(x - 0.5)^2), \\ \eta^0(x) &= \eta_0, \quad u^0(x) = u_0. \end{aligned} \tag{3.79}$$



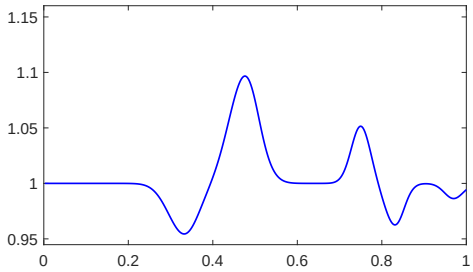
(a) η for $t = 0.1$ u for $t = 0.1$

(b) η for $t = 0.2$ u for $t = 0.2$

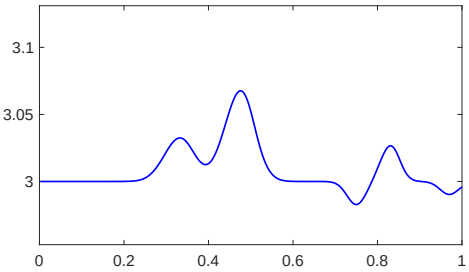
(c) η for $t = 0.3$ u for $t = 0.3$

(d) η for $t = 0.5$ u for $t = 0.5$

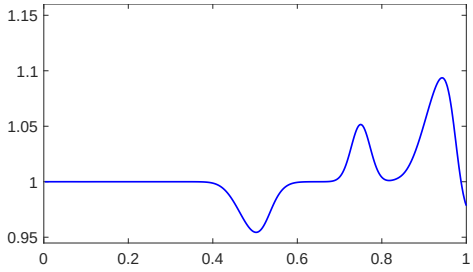
Figure 3.1: Evolution with data (3.75), supercritical case, $r = 2$, $h = 1/400$, $k = h/3$.



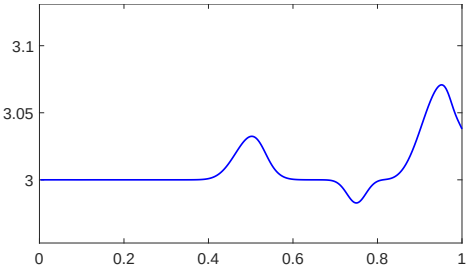
(a) η for $t = 0.5000$



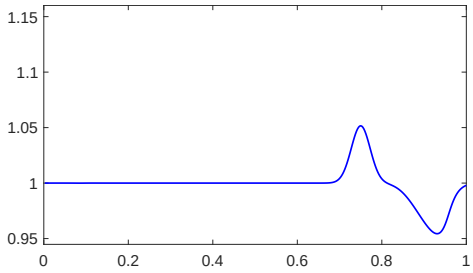
u for $t = 0.5000$



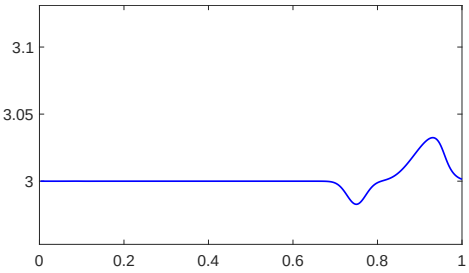
(b) η for $t = 0.1542$



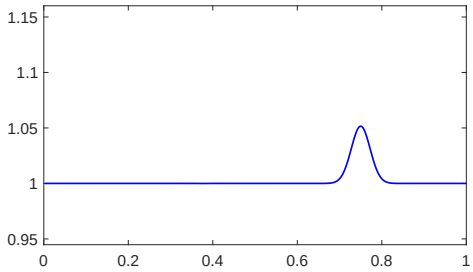
u for $t = 0.1542$



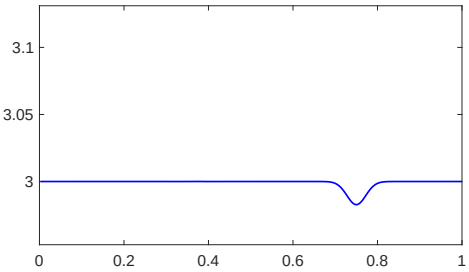
(c) η for $t = 0.4167$



u for $t = 0.4167$



(d) η for $t = 0.6000$



u for $t = 0.6000$

Figure 3.2: Evolution with data (3.78), supercritical case, $r = 2$, $h = 1/400$, $k = h/3$.

The ensuing evolution of the solution is shown in Figure 3.3. The generated wave interacts with the bottom and forms pulses that exit without artificial oscillations at both ends of the boundary; the steady-state solution may be found analytically as before. When used as initial condition, its L^2 projection differed from the numerical solution at $t = 2$ by an L^2-error of $\mathcal{O}(10^{-8})$ for this example.

An example of subcritical flow with variable initial conditions is shown in Figure 3.4, where we took $\eta_0 = u_0 = 1$, and

$$\begin{aligned} \beta(x) &= 1 - 0.04\exp(-100(x-0.75)^2), \\ \eta^0(x) &= 0.05\exp(-400(x-0.5)^2) + \eta_0, \\ u^0(x) &= 0.1\exp(-400(x-0.5)^2) + u_0 \end{aligned} \tag{3.80}$$

and integrated with $h = 1/2000$, $k = h/10$. A two-way wavetrain emerges and attains steady-state by $t = 3$.

We also tested the code in a few examples of the shallow water equations with absorbing (characteristic) boundary conditions, written in *dimensional* form, i.e. as

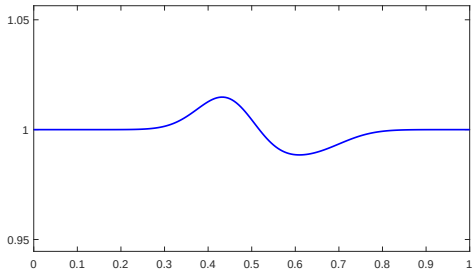
$$\begin{aligned} &\eta_t + ((\beta+\eta)u)_x = 0, \\ &u_t + g\eta_x + uu_x = 0, \end{aligned} \quad 0 \le x \le L, \;\; 0 \le t \le T, \tag{3.81}$$

with initial conditions $\eta(x,0) = \eta^0(x)$, $u(x,0) = u^0(x)$, $0 \le x \le L$, and analogous characteristic boundary conditions in the super- and subcritical cases. (The Riemann invariants are now $u \pm \sqrt{g(\beta+\eta)}$, g is the acceleration of gravity taken as $9.812\,\mathrm{m/s^2}$, and the bottom is at $z = -\beta(x)$. If the bottom is horizontal it is located at $z = -h_0$; in the general case h_0 will be a typical depth.)

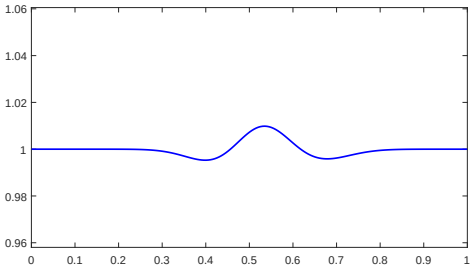
As an example of *supercritical* flow we considered a numerical experiment similar to the one described in Section 8.2 of [Shi+11]. Let $\widetilde{\beta}$ be the trapezoidal profile given by

$$\widetilde{\beta}(x) = \begin{cases} \dfrac{\delta_0}{c\kappa - \kappa/2}\left(x - \dfrac{L}{2} + c\kappa\right), & \text{if } -c\kappa \le x - L/2 \le -\kappa/2, \\ \delta_0, & \text{if } -\kappa/2 \le x - L/2 \le \kappa/2, \\ -\dfrac{\delta_0}{c\kappa - \kappa/2}\left(x - \dfrac{L}{2} - c\kappa\right), & \text{if } \kappa/2 \le x - L/2 \le c\kappa, \\ 0, & \text{otherwise,} \end{cases} \tag{3.82}$$

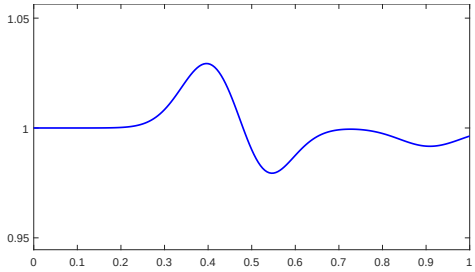
where $L = 10^6\,\mathrm{m}$, $\delta_0 = 500\,\mathrm{m}$, $k = L/10$. The bottom was located at $z = -\beta(x)$, where $\beta(x) = h_0 - \widetilde{\beta}(x)$, $h_0 = 1000\,\mathrm{m}$, and the problem (3.81) was solved with characteristic boundary conditions and initial conditions $\eta^0(x) = \eta_0 = 0$ and $u^0(x) = u_0$, where the constant u_0 was varied in order to give flows with different Froude numbers $Fr = u_0/\sqrt{gh_0}$. We solved (3.81)–(3.82) numerically with piecewise linear elements and RK4 on a uniform mesh with $h = 1000\,\mathrm{m}$, $k = 1\,\mathrm{s}$. Some profiles of the steady state of the free surface η and the associated bottom function $\beta(x)$ for various Froude numbers and values of the parameter c are



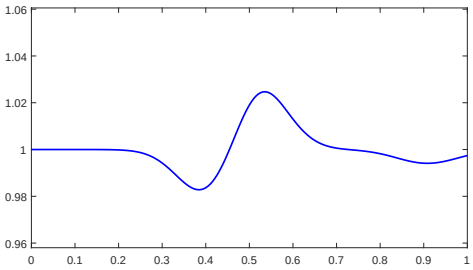
(a) η for $t = 0.06$



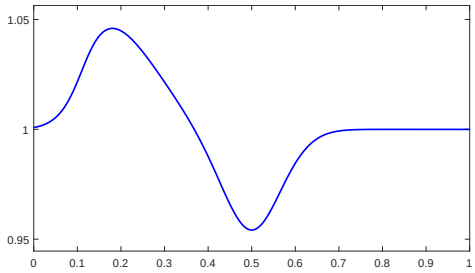
u for $t = 0.06$



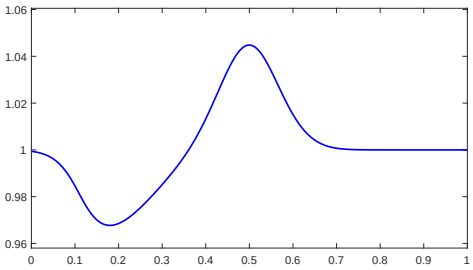
(b) η for $t = 0.17$



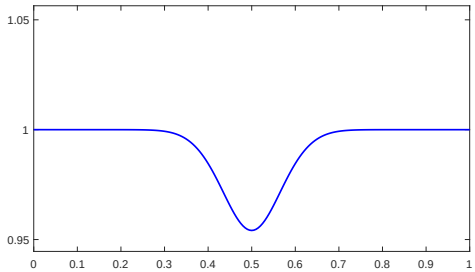
u for $t = 0.17$



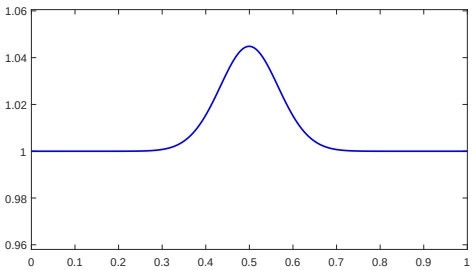
(c) η for $t = 0.72$



u for $t = 0.72$

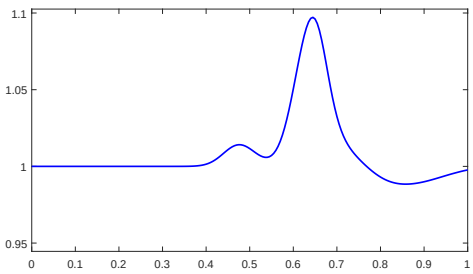


(d) η for $t = 2$

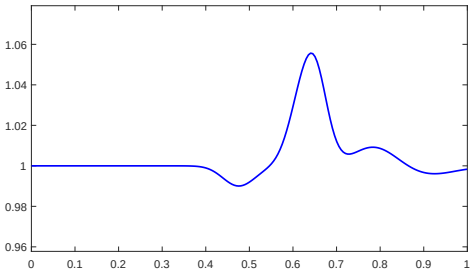


u for $t = 2$

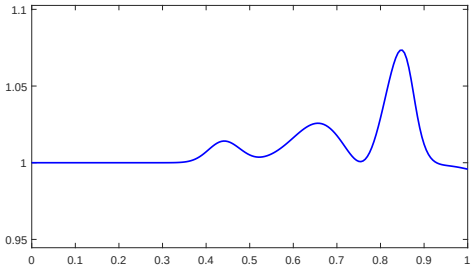
Figure 3.3: Evolution with data (3.79), subcritical case, $r = 2$, $h = 1/2000$, $k = h/10$.



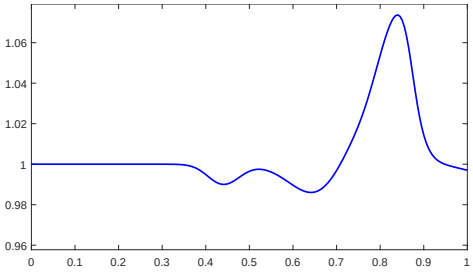
(a) η for $t = 0.057$



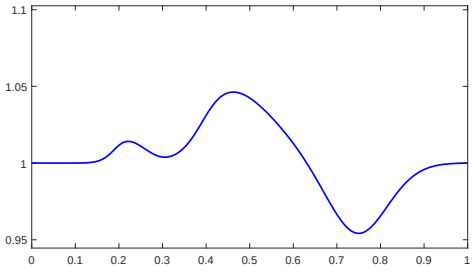
u for $t = 0.057$



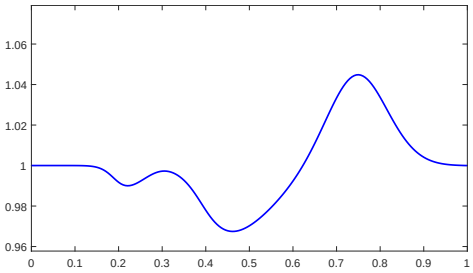
(b) η for $t = 0.138$



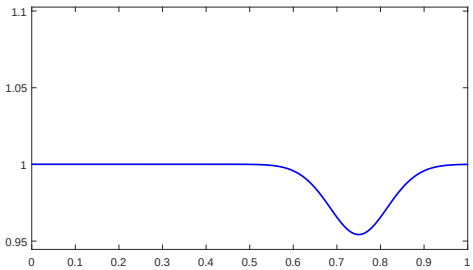
u for $t = 0.138$



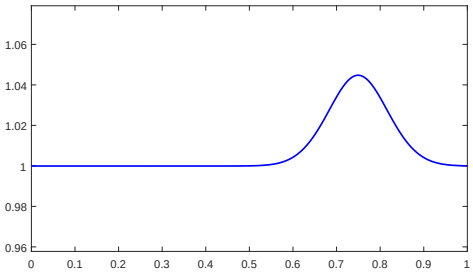
(c) η for $t = 0.651$



u for $t = 0.651$



(d) η for $t = 3$ (steady state)



u for $t = 3$ (steady state)

Figure 3.4: Evolution with data (3.80), subcritical case, $r = 2$, $h = 1/2000$, $k = h/10$.

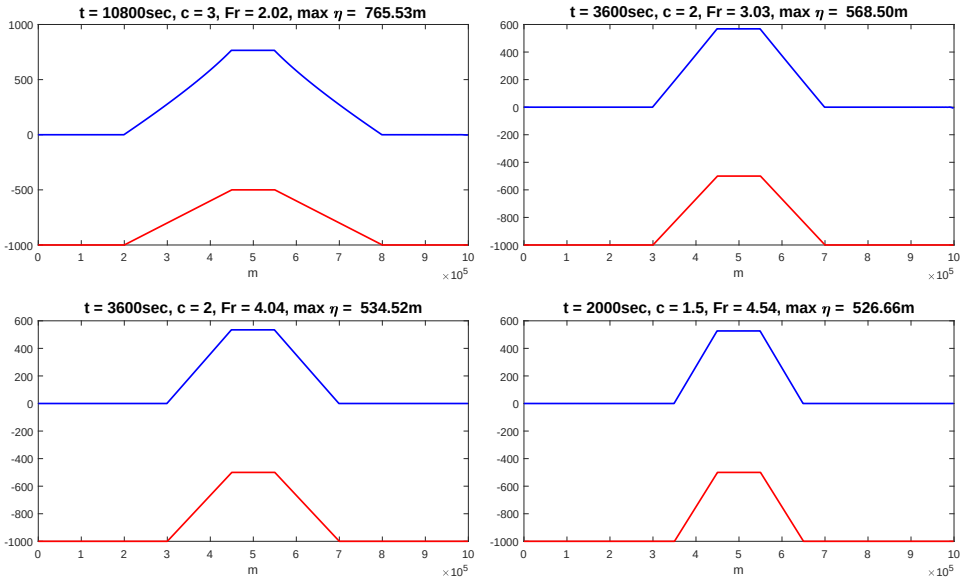


Figure 3.5: Supercritical flows over a trapezoidal bottom, (3.81)–(3.82), $r = 2$, $h = 1000\,\mathrm{m}$, $k = 1\,\mathrm{s}$. (Upper figures: steady-state $\eta(x)$; lower figures: $\beta(x)$.)

shown in Figure 3.5. As expected the eventual maximum value of η decreases as Fr increases; the results are consistent with those of [Shi+11].

In an example of a dimensional *subcritical* flow we modified the profile given in §5.1 of [Shi+11] in order to avoid discontinuity formation. Thus, the initial η-profile was rounded and its amplitude decreased. Let $\widetilde{\beta}$ be defined by

$$\widetilde{\beta}(x) = \begin{cases} \dfrac{\delta}{2} + \dfrac{\delta}{2}\cos\left[\dfrac{\pi(x - L/2)}{\kappa}\right], & \text{if } \left|x - \dfrac{L}{2}\right| < \kappa, \\ 0, & \text{otherwise,} \end{cases} \tag{3.83}$$

where $L = 10^6\,\mathrm{m}$, $\delta = 5000\,\mathrm{m}$, $k = L/10$. The bottom was taken at $z = -\beta(x)$, where $\beta(x) = h_0 - \widetilde{\beta}(x)$, $h_0 = 10^4\,\mathrm{m}$, and the problem (3.81) was solved with characteristic boundary conditions with $\eta_0 = u_0 = 0$ and $\eta^0(x) = 0.2\,\varepsilon\,h_0 \exp\left[-5{\cdot}10^{-8}\left((x - 3L/20)/10\right)^2\right]$, $0 \le x \le L$, where $\varepsilon = 0.2$.

The evolution of the η-profiles is shown in Figure 3.6 up to the attainment of the steady state $\eta = u = 0$. The results resemble qualitatively those of [Shi+11].

3.3.2 Shallow water equations in balance-law form

In this section we consider the numerical solution by the standard Galerkin finite element method of the shallow water equations written in *balance-law form* (i.e. in conservation-law form with a source term), as

$$\begin{aligned} &d_t + (du)_x = 0, \\ &(du)_t + \left(du^2 + \tfrac{1}{2}d^2\right)_x = \beta'(x)d, \end{aligned} \tag{3.84}$$

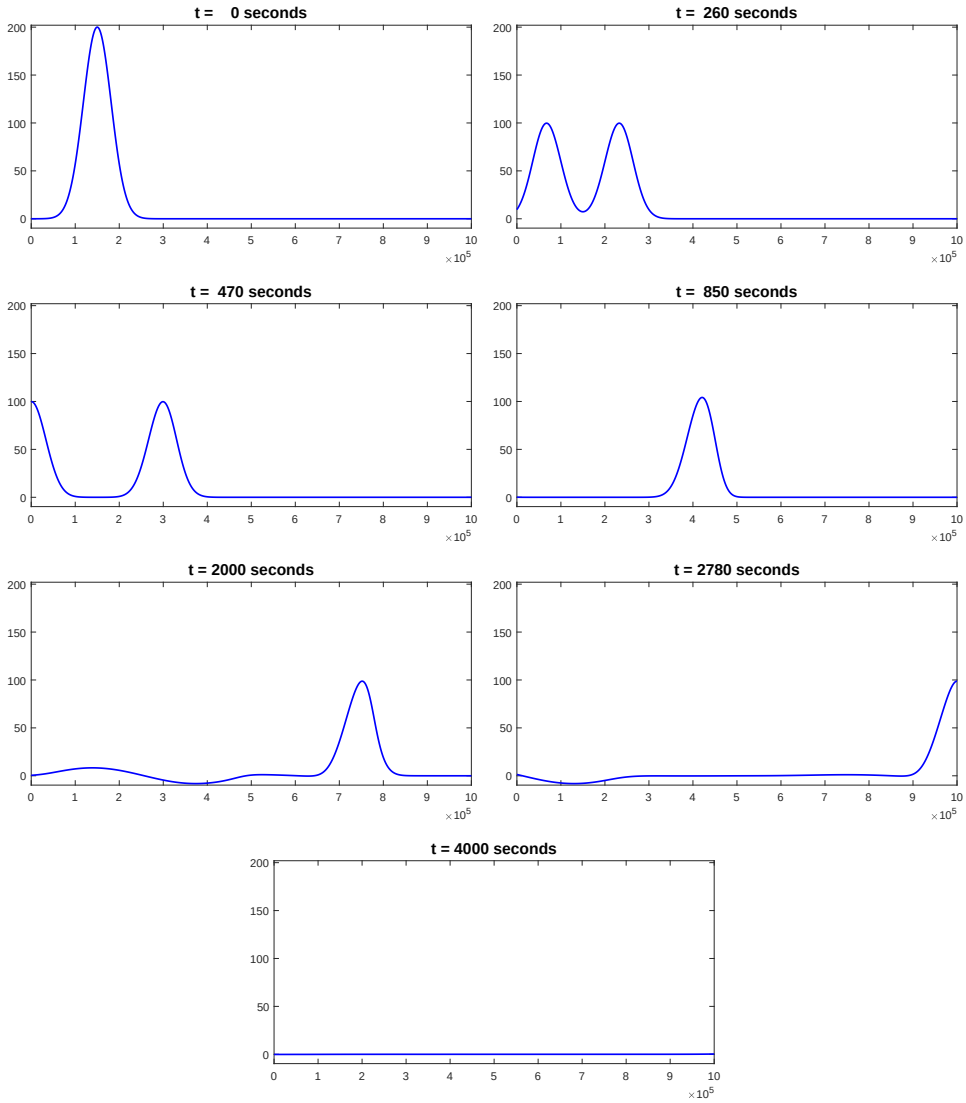


Figure 3.6: Subcritical flow over a hump, (3.81), (3.83), $r = 2$, $h = 1000\,\mathrm{m}$, $k = 1\,\mathrm{s}$, η profiles.

where $d = \eta + \beta$ is the water depth assumed as always to be positive; the variables in (3.84) are nondimensional. It is straightforward to see that the system (3.84) is equivalent to (SW) since $d \neq 0$. In the sequel we will consider the periodic initial-value problem for (3.84) on the spatial interval $[0, 1]$ and assume that it has sufficiently smooth solutions for $t \in [0, T]$, provided that β is smooth and 1-periodic. We will discretize the problem in space on a uniform or quasiuniform mesh $\{x_i\}$ in $[0, 1]$ and seek approximations d_h, u_h of d, u, respectively, in the finite element space $S_{h,p} = \{\phi \in C^k_p : \phi\big|_{[x_j,x_{j+1}]} \in \mathbb{P}_{r-1}, \text{all } j\}$, where as usual r, k are integers such that $r \geq 2$, $0 \leq k \leq r-2$, and C^k_p are the k times continuously differentiable, periodic functions on $[0, 1]$. The semidiscrete approximations satisfy

$$\begin{aligned} &(d_{ht}, \phi) + (d_h u_h, \phi) = 0, \\ &((d_h u_h)_t, \phi) + ((d_h u_h^2 + \tfrac{1}{2} d_h^2)_x, \phi) = (\beta' d_h, \phi), \end{aligned} \qquad \forall \phi \in S_{h,p},\ 0 \leq t \leq T, \quad (3.85)$$

$$d_h(0) = \mathrm{P}\, d^0, \quad u_h(0) = \mathrm{P}\, u^0, \qquad (3.86)$$

where d^0, u^0 are the initial conditions of d and u, and P is now the L^2 projection operator onto $S_{h,p}$. (The second equation in (3.85) is advanced in time for the variable $v_h = d_h u_h$ and u_h is recovered as v_h/d_h.) In the case of a uniform mesh it is expected that the L^2 errors of the semidiscrete solution will be of $\mathcal{O}(h^r)$ while, for a quasiuniform mesh, of $\mathcal{O}(h^{r-1})$, cf. [AD16]. We verified these rates of accuracy in numerical experiments using C^0 linear, C^2 cubic and C^4 quintic splines (i.e. spaces $S_{h,p}$ with $r = 2$, 4, and 6, respectively) on uniform and nonuniform spatial meshes, coupled with explicit Runge-Kutta schemes of third, fourth, and sixth order of accuracy, respectively. The fully discrete methods were stable under Courant number restrictions. We note that in order to preserve the optimal order of accuracy, say in the case of a uniform mesh, one has to compute the integrals that occur in the finite element equations using, on each subinterval $[x_i, x_{i+1}]$, an s-point Gauss quadrature rule with $s \geq r - 1$. For example, in the case of a cubic spline spatial discretization, a 3-point Gauss rule is sufficient.

It is interesting to examine whether the method (3.85)–(3.86) preserves the still water solution $\eta = 0$, $u = 0$, e.g. of the periodic ivp for the shallow water equations in the form (3.84). Discretizations that approximate accurately this solution are called 'well balanced', cf. e.g. [BV94], and [XZS10] and its references. (It is easy to check that the standard Galerkin semidiscretization of the periodic ivp for (SW), i.e. for the shallow water equations in their 'nonconservative' form, is trivially well-balanced, since it satisfies $\eta_h(x, t) = \alpha$, α constant, $u_h(x, t) = 0$ for all $t \geq 0$ and $x \in [0, 1]$, provided $\eta_h(x, 0) = \alpha$, $u_h(x, 0) = 0$. So, our attention is turned to the periodic ivp for (3.84) and its standard Galerkin semidiscretization (3.85)–(3.86).)

For this purpose, since $d = \eta + \beta$, assume that (suppressing the x-dependence in the variables), $d_h(0) = \mathrm{P}\, \beta$, $u_h(0) = 0$ in (3.86), and ask whether there exist time-independent solutions of (3.85)–(3.86) that approximate well the steady state solution $d = \beta$, $u = 0$ of the continuous problem. Taking $u_h = 0$ in (3.85) we see that a steady-state solution d_h must satisfy $(d_h d_{hx}, \phi) = (d_h \beta', \phi)$, for all ϕ in $S_{h,p}$, from which it is evident that the source term β should be replaced by

some approximation $\beta_h \in S_{h,p}$ thereof. Moreover for the equation $(d_h d_{h,x}, \phi) = (d_h \beta_h', \phi)$ to hold for $\phi \in S_{h,p}$, (this will imply that $d_h = \beta_h$, i.e. good balance), it is necessary that the integrals on each subinterval $[x_i, x_{i+1}]$ that contribute to these L^2 inner products should be evaluated *exactly*. Since both integrands are polynomials of degree at most $3r-4$ on each subinterval, if an s-point Gauss quadrature rule is used (recall that such a rule is exact for polynomials of degree at most $2s-1$), then it should hold that $s \geq \frac{3}{2}(r-1)$. For example, in the case of cubic splines $(r=4)$, a 5-point Gauss rule must be used. Therefore, although a 3-point Gauss is enough to preserve the optimal-order $\mathcal{O}(h^4)$ L^2-error estimate, good balance of the solution with cubic splines requires that a 5-point Gauss rule be used. This is confirmed by the results of the following experiment. We solve the periodic ivp for (3.84) on $[0,1]$ by (3.85)–(3.86) using cubic splines for the spatial discretization on a uniform mesh and taking $\beta(x) = 1 - 0.3\exp(-1000(x-0.5)^2)$, $h = 0.02$, $k = 0.01$, $u_h(0) = 0$, $d_h(0) = \mathrm{P}\,\beta$. Table 3.3 shows the error $d_h(1) - d_h(0)$ (where $d_h(1) = d_h\big|_{T=1}$) in the L^2 and L^∞ norms when the analytical formula of β or $\beta_h = \mathrm{P}\,\beta$ is taken in the source term, and a 3- or a 5-point Gauss rule is used. It is evident that when $b_h = \mathrm{P}\,\beta$ and a 5-point Gauss quadrature rule is used, the

$d_h(0)$	β in source term	s (-point Gauss rule)	$\|d_h(1) - d_h(0)\|$	$\|d_h(1) - d_h(0)\|_\infty$
$\mathrm{P}\,\beta$	analytical formula	3	1.8191e-4	8.3845e-4
$\mathrm{P}\,\beta$	$\beta_h = \mathrm{P}\,\beta$	3	1.2204e-6	4.7085e-6
$\mathrm{P}\,\beta$	$\beta_h = \mathrm{P}\,\beta$	5	3.7458e-15	1.0214e-14

Table 3.3: Treatment of source terms and effect of quadrature in (3.85)–(3.86), cubic splines and RK4, $h = 0.02$, $k = 0.01$, $T = 1$.

scheme is well balanced to roundoff and there is no influence of the time-stepping error. It should be noted that similar results were found when $d_h(0)$ and β_h were taken as the cubic spline interpolant of β at the nodes, and when piecewise smooth bottom profiles, e.g. like a parabolic perturbation of $\beta = 1$ supported in the interval of $[0,1]$, were considered.

Chapter 4

Discontinuous Galerkin Finite Element methods for the numerical solution of the Shallow Water equations over variable bottom

4.1 Introduction

In this Chapter we will consider high-order discontinuous Galerkin Finite Elements methods coupled with Runge-Kutta time-stepping (RKDG) for the solution of (SW) in conservation-law form. In the introductory section, 4.1, we will give a brief overview of RKDG methods for a general system of hyperbolic conservation laws. In section 4.2 we will focus in the system of the Shallow Water equations in balance-law form, that is, a system in conservation-law form with a source term. We will discuss some specific issues of implementation, such as the well-balancing (4.2.1), the positivity preservation (4.2.2), and the application of a slope limiter to achieve TVB discretization (4.2.3). Finally in the last section, 4.3, we will show the results of numerical experiments with our code intended to test the convergence rates and to simulate some standard test problems from the literature.

4.1.1 Overview of RKDG methods for a system of conservation laws

For this introduction we refer to [CS89], [CLS89] and to Cockburn's lecture notes [Coc99].
We consider the hyperbolic system

$$\begin{aligned} &\boldsymbol{u}_t + \boldsymbol{f}(\boldsymbol{u})_x = 0, \quad x \in [0,1], \;\; t \in [0,T], \\ &\boldsymbol{u}(x,0) = \boldsymbol{u}_0(x), \quad x \in [0,1], \end{aligned} \tag{4.1}$$

where $\boldsymbol{u} = (u_1, \ldots, u_m)^{\mathsf{T}} \in \mathbb{R}^m$, under periodic boundary conditions. Here $\boldsymbol{f} : \mathbb{R}^m \to \mathbb{R}^m$ is a sufficiently smooth map. We assume that (4.1) has a local in time smooth solution, which may possibly develop discontinuities eventually. Let $\{x_{j-1/2}\}_{j=1}^{N+1}$ be a partition of $[0,1]$ with $x_{1/2} = 0$, $x_{N+1/2} = 1$, and $I_j = \left[x_{j-1/2}, x_{j+1/2}\right]$, $1 \le j \le N$, with length h_j. Due to the periodic conditions we shall identify $x^+_{N+1/2}$ with $x^+_{1/2}$ and $x^-_{N+1/2}$ with $x^-_{1/2}$. We will accordingly identify the values of 1-periodic functions at $x^+_{N+1/2}$ with their values at $x^+_{1/2}$ and similarly at $x^-_{N+1/2}$, $x^-_{1/2}$. We seek an approximation $\boldsymbol{u}_h = (u_{1,h}, \ldots, u_{m,h})^{\mathsf{T}}$ to $\boldsymbol{u}$ such that $u_{i,h}$ belongs to the finite element space

$$V_h = \left\{ v \in L^2(0,1) : \ v\big|_{I_j} \in \mathbb{P}_r(I_j), \ j = 1, \ldots, N \right\}.$$

and such that

$$\begin{aligned}
&\int_0^1 \boldsymbol{u}_{ht} \boldsymbol{v} \, \mathrm{d}x - \sum_{j=1}^{N} \int_{I_j} \boldsymbol{f}(\boldsymbol{u}_h) \boldsymbol{v}_x \, \mathrm{d}x \\
&\quad + \sum_{j=1}^{N} \left(\boldsymbol{f}(\boldsymbol{u}_h(x^-_{j+1/2}, t)) \boldsymbol{v}(x^-_{j+1/2}) - \boldsymbol{f}(\boldsymbol{u}_h(x^+_{j-1/2}, t)) \boldsymbol{v}(x^+_{j-1/2}) \right) = 0, \\
&\int_{I_j} \boldsymbol{u}_h(x,0) \boldsymbol{v}(x) \, \mathrm{d}x = \int_{I_j} \boldsymbol{u}_0(x) \boldsymbol{v}(x) \, \mathrm{d}x, \qquad \forall \boldsymbol{v} \in \boldsymbol{V}_h, \ \ 1 \le j \le N.
\end{aligned} \tag{4.2}$$

Since we have not made any assumptions about the continuity of $\boldsymbol{u}_h$ at the points $x_{j+1/2}$, it is necessary to replace the flux $\boldsymbol{f}(\boldsymbol{u}_h(x^{\pm}_{j+1/2}, t))$ by a *numerical flux* $\widehat{\boldsymbol{f}}(\boldsymbol{u}_{j+1/2}, t)$ that depends on the two values of $\boldsymbol{u}_h$ at the point $x_{j+1/2}$. The numerical fluxes that we will mainly consider, see also [CLS89], (suppressing t) are:
(i) the *local Lax-Friedrichs flux* (LLF)

$$\widehat{\boldsymbol{f}}^{\mathrm{LLF}}(\boldsymbol{u}^-_{j+1/2}, \boldsymbol{u}^+_{j+1/2}) = \tfrac{1}{2}[\boldsymbol{f}(\boldsymbol{u}^-_{j+1/2}) + \boldsymbol{f}(\boldsymbol{u}^-_{j+1/2}) - \alpha_{j+1/2}(\boldsymbol{u}^+_{j+1/2} - \boldsymbol{u}^-_{j+1/2})] \tag{4.3}$$

where $\alpha_{j+1/2} = \max_{1 \le p \le m}(|\lambda^{(p)+}_{j+1/2}|, |\lambda^{(p)-}_{j+1/2}|)$, and $\lambda^{(p)\pm}_{j+1/2}$, $p = 1, \ldots, m$ are the m (real) eigenvalues of the Jacobian $\partial \boldsymbol{f} / \partial \boldsymbol{u}\big|_{\boldsymbol{u} = \boldsymbol{u}^{\pm}_{j+1/2}}$,

(ii) the *Lax-Friedrichs flux*, $\widehat{\boldsymbol{f}}^{\mathrm{LF}}$, which is the same as (4.3) but with $\alpha_{j+1/2} = \alpha = \max_{1 \le p \le m, 1 \le j \le N}(|\lambda^{(p)+}_{j+1/2}|, |\lambda^{(p)-}_{j+1/2}|)$

We will choose the basis functions of V_h as follows. Let P_ℓ be the shifted Legendre polynomials of degree ℓ in $[0,1]$ where $0 \le \ell \le r$. We note that the Legendre polynomials are orthogonal in $L^2(0,1)$; specifically, with respect to the $L^2(0,1)$ norm we have $(P_\ell, P_{\ell'}) = \left(\frac{1}{2\ell+1}\right) \delta_{\ell\ell'}$, leading to a diagonal mass matrix. Furthermore it holds that $P_\ell(0) = (-1)^\ell$, $P_\ell(1) = 1$. If we express the solution $\boldsymbol{u}_h$ of (4.2) as

$$\boldsymbol{u}_h(x,t) = \sum_{\substack{\ell = 0 \ldots r \\ j = 1 \ldots N}} \boldsymbol{u}^\ell_j(t) \phi^\ell_j(x),$$

where ϕ_j^ℓ are the local basis functions, $\phi_j^\ell(x) = P_\ell\left(\frac{x-x_{j-1/2}}{h_j}\right)$, from (4.2) we arrive to the semidiscrete form

$$\begin{aligned}\partial_t \boldsymbol{u}_j^\ell(t) - \frac{2\ell+1}{h_j}\int_{I_j} \boldsymbol{f}(\boldsymbol{u}_h(x,t))\partial_x\phi_j^\ell(x)\,\mathrm{d}x \\ + \frac{2\ell+1}{h_j}\left\{\widehat{\boldsymbol{f}}(\boldsymbol{u}(x_{j+1/2},t)) - (-1)^\ell\widehat{\boldsymbol{f}}(\boldsymbol{u}(x_{j-1/2},t))\right\} = 0, \quad (4.4)\\ \boldsymbol{u}_j^\ell(0) = \frac{2\ell+1}{h_j}\int_{I_j}\boldsymbol{u}_0(x)\phi_j^\ell(x)\,\mathrm{d}x,\end{aligned}$$

for $j = 1\ldots N,\ \ell = 0\ldots r$.

Concerning the temporal discretization, notice that the semidiscretization may be be written as

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u}_j^\ell(t) = L_{j,\ell}(\boldsymbol{u}_h(x,t)), \quad j = 1\ldots N,\ \ell = 0\ldots r. \tag{4.5}$$

Unless stated otherwise, we will use for time stepping the 3rd order Shu-Osher Runge-Kutta method, [SO88], given by the Butcher tableau

$$\begin{array}{ccc|c} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{2} \\ \hline \frac{1}{6} & \frac{1}{6} & \frac{2}{3} & \end{array}$$

The scheme has been used in many computations with fully discrete DG methods for conservation laws due to its relatively high order and TVD property, cf. [SO88]. For the actual implementation we follow [SO88], i.e. for an explicit RK scheme with k steps we compute the intermediate steps $\boldsymbol{u}_j^{\ell,i}$ by

$$\boldsymbol{u}_j^{\ell,i} = \sum_{q=0}^{i-1}\alpha_{iq}\boldsymbol{u}_j^{\ell,q} + \beta_q\Delta t L(\boldsymbol{u}_j^{\ell,q}), \quad i = 0\ldots k+1,$$

and $\boldsymbol{u}_j^\ell$ at the next time step is given by $\boldsymbol{u}_j^{\ell,k+1}$. Using this algorithm the aforementioned RK scheme can be written as a two-stage method with coefficients

$$\begin{array}{ccc|c} \multicolumn{3}{c|}{\alpha_{iq}} & \beta_q \\ \hline 1 & & & 1 \\ \frac{2}{4} & \frac{1}{4} & & \frac{1}{4} \\ \frac{1}{3} & 0 & \frac{2}{3} & \frac{2}{3} \end{array}$$

Another feature of the RKDG methods is the *slope limiter*. Since no smoothness is imposed at the boundary of each element, the numerical solution might achieve arbitrarily large values within a cell containing a discontinuity, which will result in

oscillations that will inevitably propagate to neighboring cells. To handle such issues, a *slope-limiting* procedure based on the values of the solution in neighbouring cells is applied. Limiting the solution while maintaining high-order accuracy is not trivial, but many such slope-limiters exist, cf. [CS89], [BDF94] among other.

We will use the simple $\Lambda\Pi_h^k$ limiter (sometimes referred to as "minmod limiter") as described in [CS89]. The procedure for piecewise linear polynomials, (limiter $\Lambda\Pi_h^1$), is as follows. Since $\phi_j^\ell(x_{j+1/2}^-) = 1$ and $\phi_j^\ell(x_{j-1/2}^+) = (-1)^\ell$ we can easily write the solution at these nodes as

$$\boldsymbol{u}_{j+1/2}^- = \boldsymbol{u}_j^{(0)} + \tilde{\boldsymbol{u}}_j, \quad \boldsymbol{u}_{j-1/2}^+ = \boldsymbol{u}_j^{(0)} - \tilde{\tilde{\boldsymbol{u}}}_j, \tag{4.6}$$

where $\boldsymbol{u}_j^{(0)}$ are the coefficients of the constant terms. To limit the solution, all we have to do is to modify $\tilde{\boldsymbol{u}}_j$, $\tilde{\tilde{\boldsymbol{u}}}_j$ by

$$\tilde{\boldsymbol{u}}_j^{(\text{mod})} = \boldsymbol{m}(\tilde{\boldsymbol{u}}_j, \boldsymbol{u}_{j+1}^{(0)} - \boldsymbol{u}_j^{(0)}, \boldsymbol{u}_j^{(0)} - \boldsymbol{u}_{j-1}^{(0)}),\ \tilde{\tilde{\boldsymbol{u}}}_j^{(\text{mod})} = \boldsymbol{m}(\tilde{\tilde{\boldsymbol{u}}}_j, \boldsymbol{u}_{j+1}^{(0)} - \boldsymbol{u}_j^{(0)}, \boldsymbol{u}_j^{(0)} - \boldsymbol{u}_{j-1}^{(0)})$$

where $\boldsymbol{m}$ is the component-wise applied *modified minmod function*, that is

$$\widetilde{\boldsymbol{m}}(\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3) = \begin{pmatrix} \widetilde{m}((a_1)_1, (a_2)_1, (a_3)_1) \\ \vdots \\ \widetilde{m}((a_1)_m, (a_2)_m, (a_3)_m) \end{pmatrix}$$

where

$$\widetilde{m}(a_1, a_2, a_3) = \begin{cases} a_1, & \text{if } |a_1| \le Mh^2, \\ s \cdot \min(|a_1|, |a_2|, |a_3|), & \text{if } |a_1| > Mh^2, \text{ and} \\ & \text{sign}(a_1) = \text{sign}(a_2) = \text{sign}(a_3) = s, \\ 0, & \text{otherwise,} \end{cases}$$

where M is an estimate of the value of the second derivatives of the solution near smooth critical points of $\boldsymbol{u}_0$. Finally we reconstruct the linear terms, $\boldsymbol{u}_j^{(1)}$, from either of the two relations in (4.6). In the case of higher order polynomials, (limiter $\Lambda\Pi_h^k$), we repeat the procedure described above, and if either of $\tilde{\boldsymbol{u}}_j$ or $\tilde{\tilde{\boldsymbol{u}}}_j$ is actually modified, we restrict the solution to linear and apply the $\Lambda\Pi_h^1$ limiter. Note that the limiter has to be applied in each stage of the Runge Kutta method.

Observe that this limiter is applied not only in cells that may contain a discontinuity, but near local extrema as well. When applied it reduces the solution to linear, which is not really an issue for cells containing a discontinuity. However near smooth extrema it tends to flatten the solution.

Actually the slope limiter, as well as the numerical flux, have to be applied in the local characteristic variables that we will define in the next section in the case of the Shallow Water equations.

4.2 RKDG methods for Shallow Water equations over variable bottom

The system in which we are particularly interested in this Chapter is the Shallow Water equations in balance law form, see also subsection 3.3.2. In this section we mainly follow, with additions and modifications the paper [XZS10] by Zing, Zhang and Shu. We rewrite the system here for the convenience of the reader:

$$\begin{aligned} &d_t + (du)_x = 0, \\ &(du)_t + \left(du^2 + \tfrac{1}{2}d^2\right)_x = -\beta'(x)\, d. \end{aligned} \tag{4.7}$$

Here u denotes the fluid velocity, β represents the bottom topography, and d is the water height, assumed to be non-negative. Handling of dry areas, where $d = 0$, introduces an important difficulty in numerical methods. As we saw in Chapter 3 this system has still-water steady-state solutions, wherein the flux gradients are nonzero and are exactly balanced by the source terms. Another issue that arises in numerical schemes for (4.7) is the conservation of mass, especially in the presence of dry areas.

If we set $\boldsymbol{u} = (d, du)^\mathsf{T}$, $\boldsymbol{f} = (du, \frac{(du)^2}{d} + \frac{d^2}{2})^\mathsf{T}$, and $\boldsymbol{s} = (0, -\beta' d)^\mathsf{T}$, then the system may be written in short as

$$\boldsymbol{u}_t + \boldsymbol{f}(\boldsymbol{u})_x = \boldsymbol{s}(d, \beta),$$

and the semidiscretization, following Section 4.1.1, is for $\boldsymbol{v} \in \boldsymbol{V}_h$, $t > 0$

$$\begin{aligned} &\int_{I_j} \boldsymbol{u}_{h,t}\boldsymbol{v}\,\mathrm{d}x - \int_{I_j} \boldsymbol{f}(\boldsymbol{u}_h)\partial_x\boldsymbol{v}\,\mathrm{d}x \\ &\quad + \widehat{\boldsymbol{f}}(\boldsymbol{u}^{\pm}_{h,j+1/2})\boldsymbol{v}(x^-_{j+1/2}) - \widehat{\boldsymbol{f}}(\boldsymbol{u}^{\pm}_{h,j-1/2})\boldsymbol{v}(x^+_{j-1/2}) = \int_{I_j} \boldsymbol{s}(d_h, \beta_h)\boldsymbol{v}\,\mathrm{d}x, \end{aligned} \tag{4.8}$$

where $\widehat{\boldsymbol{f}}$ is the numerical flux. It is important to note that in the fully discrete form we will use the L^2-projection of the bottom topography function, β_h, onto V_h rather than the original function β.

4.2.1 Well-balancing

A method is said to be *well balanced* if it preserves exactly the still-water steady-state solution, that is

$$d + \beta = \text{const.} \quad \text{and} \quad u = 0. \tag{4.9}$$

As we saw in section 3.3.2 continuous Galerkin methods are inherently well balanced; however this is not true in the discontinuous case. By substituting $d_h =$

$-\beta_h$, $u_h = 0$ in (4.8), we see that the first equation vanishes, while the second equation, since $\beta_h \in V_h$, becomes

$$\int_{I_j} \tfrac{1}{2} d_h^2 v_x \,\mathrm{d}x + \widehat{f}^{(2)}(\boldsymbol{u}^{\pm}_{h,j+1/2}) v(x^-_{j+1/2}) - \widehat{f}^{(2)}(\boldsymbol{u}^{\pm}_{h,j-1/2}) v(x^+_{j-1/2}) = \int_{I_j} (\tfrac{1}{2} d_h^2)_x v \,\mathrm{d}x. \tag{4.10}$$

Given a suitable quadrature rule both integrals in (4.10) are computed exactly, however, since the bottom β_h is in general discontinuous, the numerical flux does not reduce to the system's flux and the integration by parts is not exact. (Notice that if β_h were to be continuous then d_h is continuous too, since $d_h = -\beta_h + \text{const.}$.)

An idea that is due to [Aud+04], see also [XS06, §3], is to modify slightly (that is by terms of $\mathcal{O}(h^{r+1})$) the numerical flux so that in the case of the still-water solution it matches the system's flux.

If we set $\widehat{\boldsymbol{f}}_{j+1/2} := \widehat{\boldsymbol{f}}(\boldsymbol{u}^{\pm}_{h,j+1/2})$, the method can be written as

$$\int_{I_j} \partial_t \boldsymbol{u}_h v \,\mathrm{d}x - \int_{I_j} \boldsymbol{f}(\boldsymbol{u}_h) \partial_x v \,\mathrm{d}x + \widehat{\boldsymbol{f}}_{j+\frac{1}{2}} v\big(x^-_{j+\frac{1}{2}}\big) - \widehat{\boldsymbol{f}}_{j-\frac{1}{2}} v\big(x^+_{j-\frac{1}{2}}\big) =$$
$$\int_{I_j} \boldsymbol{s}(d_h, \beta_h) v \,\mathrm{d}x + \left(\widehat{\boldsymbol{f}}_{j+\frac{1}{2}} - \widehat{\boldsymbol{f}}^{l}_{j+\frac{1}{2}}\right) v\big(x^-_{j+\frac{1}{2}}\big) - \left(\widehat{\boldsymbol{f}}_{j-\frac{1}{2}} - \widehat{\boldsymbol{f}}^{r}_{j-\frac{1}{2}}\right) v\big(x^+_{j-\frac{1}{2}}\big),$$

where $\widehat{\boldsymbol{f}}_{j+\frac{1}{2}} - \widehat{\boldsymbol{f}}^{l}_{j+\frac{1}{2}}$ and $\widehat{\boldsymbol{f}}_{j-\frac{1}{2}} - \widehat{\boldsymbol{f}}^{r}_{j-\frac{1}{2}}$ are correction terms of $\mathcal{O}(h^{r+1})$ and are calculated as follows. After computing the cell boundary values $\boldsymbol{u}^{\pm}_{h,j+\frac{1}{2}}$, we set

$$d^{*,\pm}_{h,j+\frac{1}{2}} = \max\left(0, d^{\pm}_{h,j+\frac{1}{2}} + \beta^{\pm}_{h,j+\frac{1}{2}} - \max(\beta^{+}_{h,j+\frac{1}{2}}, \beta^{-}_{h,j+\frac{1}{2}})\right),$$

and redefine the left and right values of $\boldsymbol{u}$ as

$$\boldsymbol{u}^{*,\pm}_{h,j+\frac{1}{2}} = \begin{pmatrix} d^{*,\pm}_{h,j+\frac{1}{2}} \\ d^{*,\pm}_{h,j+\frac{1}{2}} u^{\pm}_{h,j+\frac{1}{2}} \end{pmatrix}.$$

Then the left and right modified fluxes $\widehat{\boldsymbol{f}}^{l}_{j+\frac{1}{2}}$ and $\widehat{\boldsymbol{f}}^{r}_{j+\frac{1}{2}}$ are given by

$$\widehat{\boldsymbol{f}}^{l}_{j+\frac{1}{2}} = \widehat{\boldsymbol{f}}\left(\boldsymbol{u}^{*,\pm}_{h,j+\frac{1}{2}}\right) + \begin{pmatrix} 0 \\ \frac{1}{2}\left(d^{-}_{h,j+\frac{1}{2}}\right)^2 - \frac{1}{2}\left(d^{*,-}_{h,j+\frac{1}{2}}\right)^2 \end{pmatrix},$$
$$\widehat{\boldsymbol{f}}^{r}_{j-\frac{1}{2}} = \widehat{\boldsymbol{f}}\left(\boldsymbol{u}^{*,\pm}_{h,j-\frac{1}{2}}\right) + \begin{pmatrix} 0 \\ \frac{1}{2}\left(d^{+}_{h,j-\frac{1}{2}}\right)^2 - \frac{1}{2}\left(d^{*,+}_{h,j-\frac{1}{2}}\right)^2 \end{pmatrix}.$$

It is easy to see that since $\widehat{\boldsymbol{f}}$ if a monotone flux and assuming β is smooth enough, $\widehat{\boldsymbol{f}}^{l}_{j+\frac{1}{2}}$ and $\widehat{\boldsymbol{f}}^{r}_{j+\frac{1}{2}}$, are indeed $\mathcal{O}(h^{k+1})$. Also under the still-water stationary

state (4.9), $\boldsymbol{u}^{*,-}_{h,j+1/2} = \boldsymbol{u}^{*,+}_{h,j+1/2}$ and the modified numerical flux is equal to the original flux,

$$\widehat{\boldsymbol{f}}^{l}_{j+\frac{1}{2}} = \begin{pmatrix} 0 \\ \frac{1}{2}\left(d^{*,-}_{h,j+\frac{1}{2}}\right)^2 \end{pmatrix} + \begin{pmatrix} 0 \\ \frac{1}{2}\left(d^{-}_{h,j+\frac{1}{2}}\right)^2 - \frac{1}{2}\left(d^{*,-}_{h,j+\frac{1}{2}}\right)^2 \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{1}{2}\left(d^{-}_{h,j+\frac{1}{2}}\right)^2 \end{pmatrix} = \widehat{\boldsymbol{f}}(\boldsymbol{u}^{-}_{h,j+1/2}),$$

and similarly

$$\widehat{\boldsymbol{f}}^{r}_{j-\frac{1}{2}} = \widehat{\boldsymbol{f}}(\boldsymbol{u}^{+}_{h,j-1/2}).$$

In order to verify the well-balanced property we perform a simulation. We let the bottom topography to be $\beta(x) = 0.5\exp(-100(x-4.5)^2)$, the water height $\zeta_0 = 1$, the velocity $u_0 = 0$, and the computational domain $x \in [0, 8]$, $N = 1000$. (Note that the bottom does not reach the free surface.) The exact solution is $d(x,t) = 1 - \beta(x)$, $u(x,t) = 0$. We will use Gauss-Legendre quadrature with enough nodes so that all integrals will be computed exactly.

The errors for various polynomial orders r of the standard RKDG scheme (without the well-balanced modification) at $T = 1$ can be see in Table 4.1. Note that all norms are computed exactly and the quadrature rule is accurate enough. We observe that the errors are $\mathcal{O}(10^{-4})$ and diminish as r grows. (The errors seem to be independent of the grid, and depend only on the smoothness of β, in this example $\max_x |\beta'| \simeq 4.3$ and $\max_x |\beta''| = 100$.)

r	quadrature	$\|d_h + \beta_h - 1\|_\infty$	$\|d_h + \beta_h - 1\|$	$\|d_h + \beta_h - 1\|_{L^1}$
1	G-Leg-3	4.2657e-04	1.6003e-04	1.5597e-04
2	G-Leg-3	1.0650e-04	3.9976e-05	3.8980e-05
3	G-Leg-5	1.3000e-07	4.8095e-08	4.5363e-08

Table 4.1: Still-water steady-state solution errors of the standard RKDG scheme for $\beta(x) = .5\exp(-100(x-4.5)^2)$, and $\zeta(x) = 1$. Simulation run up to $T = 1$ (assume d positive).

Errors of the well-balanced scheme can be seen in Table 4.2 and are of the order of machine precision. (The errors depend on h in this case. As N gets larger, since more computations are made, the errors increase slowly due to roundoff.)

4.2.1.1 Non-negative water height

Special attention has to be given to the case where the bottom rises above the free surface. We will define the still-water steady-state solution for such bottoms, $\boldsymbol{u}_{\text{still}}$,

r	quadrature	$\|d_h + \beta_h - 1\|_\infty$	$\|d_h + \beta_h - 1\|$	$\|d_h + \beta_h - 1\|_{L^1}$
1	G-Leg-3	1.3434e-14	1.5145e-14	2.4021e-14
2	G-Leg-3	2.5202e-14	3.2691e-14	5.4322e-14
3	G-Leg-5	4.3077e-14	7.9845e-14	1.7335e-13

Table 4.2: Still-water steady-state solution errors of the well-balanced RKDG scheme for $\beta(x) = .5\exp(-100(x-4.5)^2)$, and $\zeta(x) = 1$. Simulation run up to $T = 1$ (assume d positive).

by

$$u_{\text{still}} = 0 \quad \text{and} \quad \begin{cases} d_{\text{still}} + \beta = \text{const.} =: c, & \text{if } \beta \le c \\ d_{\text{still}} = 0, & \text{if } \beta > c. \end{cases}$$

For simplicity we assume that the bottom is strictly increasing and let the wet-dry interface be located at x_*, where $\beta(x_*) = c$. We can easily see that if x_* lies in the interior of some cell, d_{still} no longer belongs to V_h, even if $\beta = \beta_h \in V_h$ (due to the added assumption that $d \ge 0$ for $x > x_*$), for any polynomial of order $r \ge 1$; see for example Figure 4.1. Furthermore, since we require that $d_h(x) > 0$ for all x,

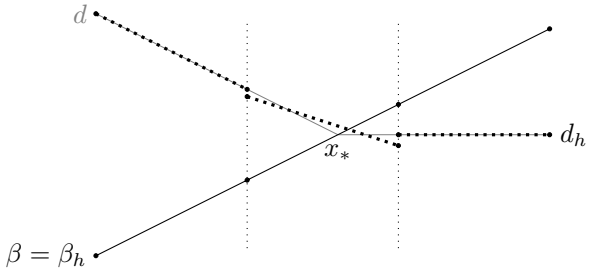


Figure 4.1: L^2-projection of d onto V_h when $x_* \notin \{x_i\}_{i=1}^{N+1}$, $c = 0$, $r = 1$. Solid black line: bottom $\beta(x)$, solid gray line: d, dotted line: d_h.

a positivity-preserving limiter will be applied (see section 4.3) making the solution d_h even larger for $x > x_*$, cf. Figure 4.2. Observe now that d_h is far from the

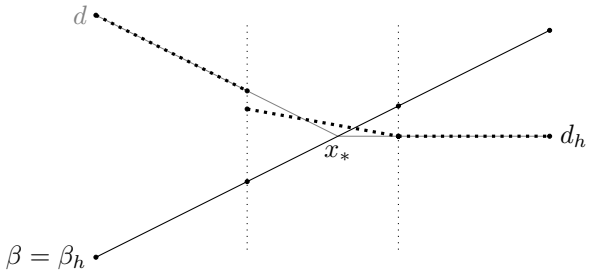


Figure 4.2: d_h after positivity-preserving limiter is applied, $c = 0$, $r = 1$. Solid black line: bottom $\beta(x)$, solid gray line: d, dotted line: d_h.

steady-state solution and as time evolves, an artificial reflection (initially of the form of a left-traveling pulse) will be generated.

To handle this issue without additional computational cost, we may include the wet-dry interface point in the spatial grid. A simple way of achieving this is to include x_* itself in the computational grid. But doing so will introduce additional (coding) complexity and might result to arbitrarily small cells in the case of an adaptive grid. A second way is to modify the projection of the bottom β_h, so that the wet-dry interface moves to the boundary of the corresponding cell. Since the bottom topography is given, if we assume that it is smooth and varies slowly, we see that this modification will not introduce large perturbations. (In the case of adaptive grids we can always introduce an additional penalty term in the estimator near the wet-dry interface, ensuring that the cell length is satisfactory small.) The procedure of bottom modification is similar to that of the positivity-preserving limiter, see section 4.2.3, i.e.

a) Project β and ζ_0 into V_h (notice that the projection of $\zeta_0 = d_0 - \beta$ is trivial since $\zeta_0 = c$).

b) In the cell I_* containing x_* modify β_h as follows: If $\overline{\beta}_h$ is the cell average of β_h in this particular cell, $\overline{\beta}_h = \frac{1}{h}\int_{I_*} \beta_h$, let

$$m = \begin{cases} \min_{x\in I_*} \beta_h(x), & \text{if } \overline{\beta}_h > c, \\ \max_{x\in I_*} \beta_h(x), & \text{if } \overline{\beta}_h \leq c, \end{cases}$$

and set $\beta_h^{\text{mod}} = \left(\frac{\overline{\beta}_h - c}{\overline{\beta}_h - m}\right)\left(\beta_h - \overline{\beta}_h\right) + \overline{\beta}_h$. Observe now that I_* will be either completely dry, if $\overline{\beta}_h > c$, or completely wet, if $\overline{\beta}_h \leq c$.

c) Set $d_{h,0} = \zeta_{h,0} + \beta_h^{\text{mod}}$ for all cells where $\overline{\beta}_h < c$.

The resulting bottom and water height can be seen in Figure 4.3. Note that this

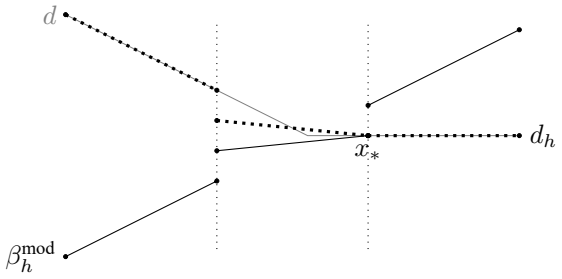


Figure 4.3: β_h, d_h after the bottom modification, $c = 0$, $r = 1$. Solid black line: bottom $\beta(x)$, solid gray line: d, dotted line: d_h.

procedure can be applied to polynomials of arbitrary order $r \geq 1$.

In order to verify the well-balanced property for the case where the bottom rises above the free surface, as previously, we perform a simulation. We now let

the bottom topography to be given by $\beta(x) = 1.1\exp(-100(x-4.5)^2)$, and as before we take the water height $\zeta_0 = 1$, the velocity $u_0 = 0$, and the computational domain $x \in [0, 8]$. Notice now that the bottom rises above the free surface. The exact solution will be $u(x,t) = 0$ and $d(x,t) = 1 - \beta(x)$ if $\beta(x) < 1$, $d(x,t) = 0$, otherwise. The errors for various polynomial orders r at $T = 1$ can be seen in Table 4.3 and are of the order of machine precision as expected.

r	quadrature	$\|d_h + \beta_h^{\text{mod}} - 1\|_\infty$	$\|d_h + \beta_h^{\text{mod}} - 1\|$	$\|d_h + \beta_h^{\text{mod}} - 1\|_{L^1}$
1	G-Leg-3	1.5432e-14	1.5501e-14	2.4681e-14
2	G-Leg-3	3.0753e-14	3.2321e-14	5.3523e-14
3	G-Leg-5	5.6399e-14	8.0099e-14	1.7330e-13

Table 4.3: Still-water steady-state solution errors for $\beta(x) = 1.1\exp(-100(x-4.5)^2)$, and $\zeta(x) = 1$. Simulation run up to $T = 1$.

4.2.2 Slope limiting

An important procedure, in the presence of discontinuities, is *slope limiting*. We will use the usual minmod limiter, see section 4.1.1. As described in [CLS89], in order to get qualitatively better results (i.e. avoid "wriggle" formation near discontinuities) at the cost of more complicated computations, the slope limiting (and the numerical flux calculation in general) has to be done in the local characteristic variables.

We briefly describe the procedure; see also [CLS89, §2]. Let $A_{j+1/2} = \partial \boldsymbol{f}/\partial \boldsymbol{u}\big|_{\boldsymbol{u}=\boldsymbol{u}_{j+1/2}}$ be some "average" Jacobian, where $\boldsymbol{u}_{j+1/2} = (\boldsymbol{u}_j^{(0)} + \boldsymbol{u}_{j+1}^{(0)})/2$, and let $\lambda_{j+1/2}^{(p)}$, $\boldsymbol{l}_{j+1/2}^{(p)}$, $\boldsymbol{r}_{j+1/2}^{(p)}$, $p = 1, \dots, m$, be the eigenvalues and left and right normalized eigenvectors of $A_{j+1/2}$ respectively. To apply the slope limiter, we project all the required quantities onto the left eigenspace of $A_{j+1/2}$ using

$$a^{(p)} = \boldsymbol{l}_{j+1/2}^{(p)} \cdot \boldsymbol{a}, \tag{4.11}$$

for $\boldsymbol{a} = \tilde{\boldsymbol{u}}_j$, $\tilde{\tilde{\boldsymbol{u}}}_j$, $\boldsymbol{u}_j^{(0)}$, $(\boldsymbol{u}_j^{(0)} - \boldsymbol{u}_{j-1}^{(0)})$, $(\boldsymbol{u}_{j+1}^{(0)} - \boldsymbol{u}_j^{(0)})$, and apply the minmod slope limiter as described in subsection 4.1.1. After the application of the slope limiter in the two adjacent cells, I_j, I_{j+1}, in order to calculate the numerical flux, $\widehat{\boldsymbol{f}}(\boldsymbol{u}_{j+1/2}^{\pm})$, we need to return to the component space using the formula

$$\boldsymbol{u}_{j+1/2}^{\pm} = \sum_{p=1}^{m} u_{j+1/2}^{\pm(p)} \boldsymbol{r}_{j+1/2}^{(p)},$$

compute $\boldsymbol{f}(\boldsymbol{u}_{j+1/2}^{\pm}) := \boldsymbol{f}_{j+1/2}^{\pm}$, project the flux again onto the eigenspace using (4.11), $(f_{j+1/2}^{\pm})^{(p)}, p = 1, \dots, m$, and then compute the numerical flux as described in subsection 4.1.1 in each characteristic field $\widehat{f}_{j+1/2}^{(p)}$, where now in the case of the

local Lax-Friedrichs flux $\alpha_{j+1/2} = a^{(p)}_{j+1/2} = \max(|\lambda^{(p)}_j|, |\lambda^{(p)}_{j+1}|)$. Finally in order to find $\widehat{\boldsymbol{f}}_{j+1/2}$, we need to return to the component space by

$$\widehat{\boldsymbol{f}}_{j+1/2} = \sum_{p=1}^{m} \widehat{f}^{(p)}_{j+1/2} \boldsymbol{r}^{(p)}_{j+1/2}.$$

Similarly we return to component space for the actual solution $\boldsymbol{u}^{(i)}_j$, $i = 0, \ldots, r$, i.e. $\widehat{\boldsymbol{u}}^{(i)}_{j+1/2} = \sum_{p=1}^{m} u^{(i)(p)}_{j+1/2} \boldsymbol{r}^{(p)}_{j+1/2}$, $i = 0, \ldots, r$.

For the particular system of interest, the Shallow Water equations in balance law form, the Jacobian of $\boldsymbol{f}$ is

$$J_{\boldsymbol{f}} = \begin{bmatrix} 0 & 1 \\ d - u^2 & 2u \end{bmatrix},$$

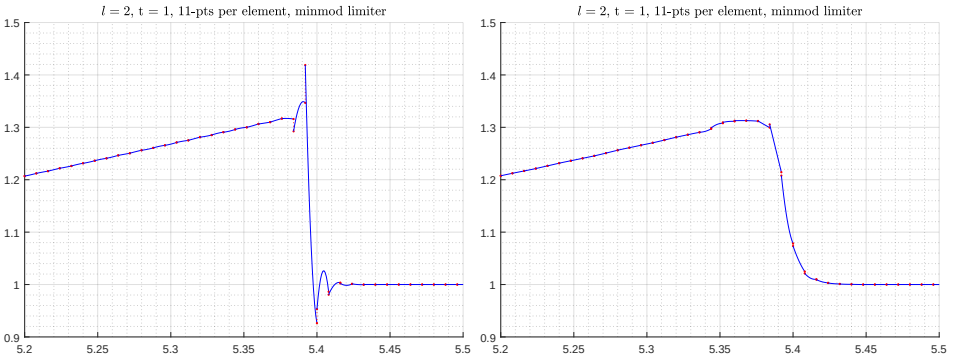
the eigenvalues are

$$\boldsymbol{\lambda} = \begin{bmatrix} \frac{du}{d} + \sqrt{d} \\ \frac{du}{d} - \sqrt{d} \end{bmatrix},$$

and the left and right normalized eigenvectors are given by

$$V_r = \begin{bmatrix} \frac{1}{2\sqrt{d}} & \frac{-1}{2\sqrt{d}} \\ \frac{du+d^{3/2}}{2d^{3/2}} & \frac{-(du-d^{3/2})}{2d^{3/2}} \end{bmatrix}, \quad V_l = \begin{bmatrix} \frac{-(du-d^{3/2})}{d} & 1 \\ \frac{-(du+d^{3/2})}{d} & 1 \end{bmatrix}.$$

The need for a slope limiter can be seen in the following experiment. We use a flat bottom $\beta(x) = 0$, and a heap of water centered at $x = 4$ with zero velocity as initial condition. Specifically we define $d_0(x) = 1 + \exp(-100(x-4)^2)$, $u_0(x) = 0$. Using quadratic polynomials, $r = 2$, and the aforementioned minmod slope limiter, we obtain the numerical solution at time $T = 1$ whose right-travelling component is shown in Figure 4.4. The initial pulse resolves itself into two pulses



(a) no slope limiter (b) minmod slope limiter

Figure 4.4: Demonstration of minmod slope limiter in the presence of shocks.

that travel in opposite direction and which quickly form a shock. Without the slope

limiting procedure, oscillations in the cells near the shock are formed. These oscillations disappear when the minmod limiter is applied. (Observe also that in the presence of the slope limiter the solution tends to become flat right before the shock, and decays smoothly just after, indicating perhaps that this limiter is not ideal for describing such solutions. There is however a large selection of slope limiters that are able to better describe the shape of the solution.)

In the case of variable bottom, applying the slope limiter in $(d, du)^\mathsf{T}$ variables might affect the preservation of the still-water steady-state solution. This can be prevented by applying the slope limiter to $(d+\beta, du)^\mathsf{T}$ instead, as can be verified in the following experiment. We set $\beta(x) = 0.5\exp(-100(x-4.5)^2)$, $\zeta(x) = 1$ and $u(x) = 0$. The exact solution is of course $d(x,t) = 1-\beta(x)$. The errors can be seen in Table 4.4: When no slope limiter is applied, as we have seen before, the errors are of the order of machine precision; when slope limiter is applied to $(d, du)^\mathsf{T}$, since d_h is not constant, the errors are significant. Finally, when the limiter is applied to $(d+\beta, du)^\mathsf{T}$, the errors are again negligible.

slope limiter	$\|d_h + \beta_h - 1\|_\infty$	$\|d_h + \beta_h - 1\|$	$\|d_h + \beta_h - 1\|_{L^1}$
none	2.6090e-14	3.4205e-14	5.6736e-14
$(d, du)^T$	2.4059e-03	2.1749e-04	4.9506e-05
$(d+\beta, du)^T$	2.5202e-14	3.2691e-14	5.4322e-14

Table 4.4: Still-water steady-state solution errors after the application of slope limiter. $\beta(x) = 0.5\exp(-100(x-4.5)^2)$, $\zeta(x) = 1$, and $r = 2$. Simulation run up to $T = 1$ (assume d positive).

4.2.3 Positivity-preserving limiter

Preservation of the non-negativity of d_h in the case of DG polynomials needs special attention. Recall from section 4.2.1 that the well-balanced RKDG semidiscretization is

$$\int_{I_j} \partial_t \boldsymbol{u}_h v \,\mathrm{d}x - \int_{I_j} \boldsymbol{f}(\boldsymbol{u}_h)\partial_x v \,\mathrm{d}x + \widehat{\boldsymbol{f}}^l_{j+\frac{1}{2}} v\big(x^-_{j+\frac{1}{2}}\big) - \widehat{\boldsymbol{f}}^r_{j-\frac{1}{2}} v\big(x^+_{j-\frac{1}{2}}\big) = \int_{I_j} \boldsymbol{s}(d_h, \beta_h) v \,\mathrm{d}x \quad (4.12)$$

Let $\{x^q_j\}_{q=1}^{n_q}$, $\{w_q\}_{q=1}^{n_q}$ be the nodes and weights of the n_q-point Gauss-Lobatto quadrature rule on $I_j = [x_{j-1/2}, x_{j+1/2}]$, and $\overline{d}^n_j$ be the average water height in cell I_j at time step n, $\overline{d}^n_h = \frac{1}{h}\int_{I_j} d^n_h \,\mathrm{d}x$. From Proposition 3.2 in combination with Remark 3.3 of [XZS10] we have

Proposition 4.1 ([XZS10, Proposition 3.2, Remark 3.3])**.** *Let* (4.12) *be satisfied by the cell averages of the water height. Let* $d^n_j(x)$ *be the DG polynomial for the*

water height in the cell I_j. If $d^-_{j-1/2}$, $d^+_{j+1/2}$, and $d^n_j(x^q_j)$, $q = 1, \ldots, n_q$ are all non-negative, then $\overline{d}^{n+1}_j$ is also non-negative under the CFL condition

$$\alpha \frac{k}{h} \leq w_1$$

where $\alpha = \max(|u| + \sqrt{d})$ and w_1 the first quadrature weight.

This proposition ensures the non-negativity of the average water height at the next time step given the non-negativity of the height for all $x \in I_j$ at the current time step, which is not guaranteed. To enforce this requirement we apply an additional *positivity-preserving* limiter to the solution $\boldsymbol{u}^n_j$, which is a linear scaling around its cell average. In particular we set

$$\begin{pmatrix} \widetilde{d}^n_j \\ (\widetilde{du})^n_j \end{pmatrix} =: \widetilde{\boldsymbol{u}}^n_j = \theta_j(\boldsymbol{u}^n_j - \overline{\boldsymbol{u}}^n_j) + \overline{\boldsymbol{u}}^n_j, \quad \theta_j = \min\left\{1, \frac{\overline{d}^n_j}{\overline{d}^n_j - m_j}\right\}, \quad (4.13)$$

with

$$m_j = \min_{x \in I_j} d^n_j(x) = \min_{q=1,\ldots,n_q} d^n_j(x^q_j).$$

It is easy to see that the limiter will modify the solution only if $d^n_j(x) < 0$ for some $x \in I_j$.

It is worth noting that this limiter preserves the conservation of height, since only higher than constant terms are modified for $\widetilde{d}_j$ (given that the basis functions are Legendre polynomials), and the conservation of momentum $\widetilde{d}_j u_j$ since $\widetilde{u}_j \equiv u_j$ is not modified. Also it maintains the well-balanced property, since in the case of still-water solutions $d^n_j(x) = \text{const.} - \beta_j(x) \geq 0$, so that the limiter will not be applied.

4.3 Numerical experiments

We will first mention some additional issues concerning the practical implementation of the method. As mentioned in [XZS10, §4] regarding the well-balanced property, there may be a conflict between the slope limiter, when applied to $(d+\beta, du)^\mathsf{T}$, and the positivity-preserving limiter (in the absence of the bottom modification described in section 4.2.1.1). In particular it is observed that the numerical step becomes smaller and smaller as time increases and the code eventually stops. This is issue is not present if we apply the slope limiter to $(d, du)^\mathsf{T}$, which is not actually well balanced as demonstrated in section 4.2.2. Although we will use the procedure described in 4.2.1.1, we will also implement the slope limiter as suggested in [XZS10], i.e. in two steps. For each cell we check first if limiting is actually required based on $(d + \beta, du)^\mathsf{T}$ if the cell is in the wet region ($\theta = 1$ in (4.13)), or based on $(d, du)^\mathsf{T}$ if it is in the dry or nearly dry region ($\theta < 1$ in (4.13)). Then, if

limiting is required, we preform the slope limiting on $(d, du)^\mathsf{T}$. Notice that this procedure will not destroy the well-balanced property since when we have a still-water solution, $d + \beta = \text{const.}$, the slope limiter is not actually applied.

Another issue is the estimation of the velocity, u, which is not a variable of the system but is given 'implicitly' by $u = (du)/d$. In the dry or nearly dry regions, where the water height is close to zero, a small error in d will result in a large error in u. Since u is required (for example) for the calculation of the eigenvalues of $J_{\boldsymbol{f}}$ used in the CFL condition, this will result in very small time steps. To combat this issue we will set $u = 0$ when $d \leq 1\mathsf{e}{-6}$.

Finally another modification that improves the stability of the method in some experiments is a more precise estimation of the CFL condition. For this purpose, after each intermediate Runge-Kutta step q, we estimate again the CFL number, $CFL^{n,q}$, based on the intermediate solution $\boldsymbol{u}_h^{n,q}$. If it is much smaller than the one for the current time step, CFL^n, (say $CFL^{n,q} < \frac{1}{10} CFL^n$), we reduce the time step and restart the RK method for the current time step.

The complete algorithm for the q-th intermediate Runge-Kutta step of the n-th time step is as follows. (We assume that the positivity preserving process and the slope limiter have already been applied at the first time step.) Also note that for a k-step RK method we assume that $\boldsymbol{u}_h^{n+1} = \boldsymbol{u}_h^{n,k+1}$).

- Calculate CFL^{n,q} and if it is much smaller than CFL^n restart the RK for the current time step using smaller k.
- Calculate $\boldsymbol{u}_h^{n,q+1}$ using (4.12).
- For each cell j, evaluate θ_j by (4.13).
- For each cell check whether a slope limiter is required based on $(d+\beta, du)^\mathsf{T}$ if the cell is in the wet region ($\theta_j = 1$), or based on $(d, du)^\mathsf{T}$ if it is in the dry or nearly dry region ($\theta_j < 1$). If a limiter is required, then apply it to $(d, du)^\mathsf{T}$.
- Apply the positivity-preserving limiter.

The numerical setup in the experiments to follow, unless otherwise indicated, will be the following. For the spatial discretization we will use a uniform grid with $N = 1000$ elements ($h = 8\mathsf{e}{-3}$) on the spatial interval $[0, 8]$ and use piecewise linear polynomials, $r = 1$. We will employ an accurate enough n_q-point Gauss-Legendre quardature so that all integrals are computed exactly. The nodes of a n_q-point Gauss-Lobatto rule will also be used for the positivity preserving limiter. For the temporal discretization we will use the third-order Shu-Osher Runge-Kutta method, as described in section 4.1.1, with CFL condition

$$\frac{k}{h} = \frac{0.9}{2r+1} \frac{1}{\max\limits_{\substack{1 \leq i \leq N \\ 1 \leq q \leq n_q}} |\lambda_{i,q}^n|} \tag{4.14}$$

where $\lambda_{i,q}^n$ is the maximum eigenvalue of the $J_{\boldsymbol{f}}$ and is evaluated using quadrature points in each element, and $0.9/(2r+1)$ is an empirical term. Finally we use the minmod slope limiter as described in section 4.1.1.

4.3.1 Convergence rates

To verify the accuracy of the scheme we use periodic b.c., a smooth variable bottom given by

$$\beta(x) = 0.1\sin(2\pi x),$$

for $x, t \in [0,1]$, and the smooth exact solution

$$\begin{aligned} d(x,t) &= 1 - \exp(-t)(\sin(2\pi x) + \sin(4\pi x))/4, \\ (du)(x,t) &= \exp(t^2)\cos(2\pi x)/10, \end{aligned}$$

for which we calculate the appropriate source terms. Notice that $\max_{x,t}|d(x,t)| = \max_x|d(x,0)| \simeq 46.605$ and $\max_{x,t}|(du)(x,t)| = \max_x|(du)(x,1)| \simeq 10.731$ and $d, du \in C^2_{\text{per}}([0,1])$. We we will use a $(n_q{+}2)$ Gauss-Legendre quadrature rule for the computation of the norms of the error. Since the minmod slope limiter, when applied, will degrade the accuracy of the scheme, it is interesting to distinguish two cases.

No slope limiter: Errors and rates for $r = 0, 1, 2$ can be seen in Tables 4.5–4.7 and the rates are equal to $r+1$ as expected.

Minmod slope limiter: As mentioned in section 4.1.1, M has to be proportional to the size of the second derivatives of the solution near the smooth extrema. For this example we choose $M = 2/3 \cdot 52$ and $r \geq 1$ of course. As expected, since the solution is smooth, the slope limiter is inactive and order of accuracy is maintained. The errors and rates of convergence are very close to those of Tables 4.6 and 4.7.

Another interesting test is to see what happens when the slope limiter is applied unconditionally. If we set $M = 2/3 \cdot M_2 = 0$ the resulting errors can be seen in Tables 4.8 and 4.9. The slope limiter is applied to 4 regions near the critical points of d, du and rates are degraded for $r = 2$.

Similar results hold for non-uniform grids too.

4.3.2 Riemann problems over a flat bottom

In this subsection we will consider two Riemann test problems that demonstrate the ability of the method to maintain the water height positive. The experiments, originally presented in [Bok05, §6.1], can be also found in [XZS10, §6.3]. To compare with the original simulations we now take gravity into account; therefore the second equation of (4.7) becomes

$$(du)_t + \left(du^2 + \tfrac{1}{2}gd^2\right)_x = -g\beta' d,$$

N	L^1	rate	L^2	rate	L_∞	rate
8	7.170e-2	-	7.648e-2	-	1.473e-1	-
16	3.571e-2	1.006	4.005e-2	0.933	8.230e-2	0.840
32	1.694e-2	1.076	1.964e-2	1.028	4.185e-2	0.976
64	7.759e-3	1.126	9.331e-3	1.074	1.942e-2	1.108
128	3.778e-3	1.038	4.617e-3	1.015	1.023e-2	0.925
256	1.945e-3	0.958	2.437e-3	0.922	5.859e-3	0.804
512	1.004e-3	0.954	1.290e-3	0.918	3.156e-3	0.893

(a) d

N	L^1	rate	L^2	rate	L_∞	rate
8	7.180e-2	-	8.387e-2	-	1.651e-1	-
16	3.607e-2	0.993	4.193e-2	1.000	9.574e-2	0.786
32	1.636e-2	1.140	2.000e-2	1.068	5.061e-2	0.920
64	6.533e-3	1.325	8.573e-3	1.222	2.310e-2	1.131
128	3.573e-3	0.871	4.247e-3	1.013	8.982e-3	1.363
256	2.264e-3	0.658	2.603e-3	0.706	5.756e-3	0.642
512	1.331e-3	0.766	1.557e-3	0.742	3.377e-3	0.770

(b) du

Table 4.5: Errors and convergence rates, no slope limiter, $r = 0$.

N	L^1	rate	L^2	rate	L_∞	rate
8	7.314e-3	-	8.636e-3	-	2.111e-2	-
16	1.639e-3	2.158	2.102e-3	2.039	6.482e-3	1.704
32	3.960e-4	2.049	5.080e-4	2.049	1.604e-3	2.015
64	9.785e-5	2.017	1.253e-4	2.019	3.849e-4	2.059
128	2.429e-5	2.010	3.117e-5	2.008	9.380e-5	2.037
256	6.051e-6	2.005	7.774e-6	2.003	2.317e-5	2.018

(a) d

N	L^1	rate	L^2	rate	L_∞	rate
8	7.221e-3	-	9.010e-3	-	2.334e-2	-
16	1.756e-3	2.040	2.200e-3	2.034	6.289e-3	1.892
32	4.140e-4	2.085	5.281e-4	2.059	1.601e-3	1.974
64	1.004e-4	2.044	1.287e-4	2.036	4.007e-4	1.998
128	2.473e-5	2.021	3.181e-5	2.017	9.948e-5	2.010
256	6.141e-6	2.010	7.907e-6	2.008	2.473e-5	2.008

(b) du

Table 4.6: Errors and convergence rates, no slope limiter, $r = 1$.

N	L^1	rate	L^2	rate	L_∞	rate
8	9.115e-4	-	1.035e-3	-	3.101e-3	-
16	1.113e-4	3.034	1.308e-4	2.984	3.686e-4	3.073
32	1.353e-5	3.040	1.631e-5	3.004	4.430e-5	3.057
64	1.698e-6	2.994	2.038e-6	3.000	5.373e-6	3.044
128	2.125e-7	2.999	2.546e-7	3.001	6.599e-7	3.025
256	2.656e-8	3.000	3.181e-8	3.001	8.179e-8	3.012

(a) d

N	L^1	rate	L^2	rate	L_∞	rate
8	9.709e-4	-	1.149e-3	-	2.757e-3	-
16	1.342e-4	2.855	1.658e-4	2.793	4.078e-4	2.757
32	1.732e-5	2.954	2.150e-5	2.947	5.460e-5	2.901
64	2.151e-6	3.010	2.696e-6	2.995	7.015e-6	2.960
128	2.670e-7	3.010	3.360e-7	3.005	8.865e-7	2.984
256	3.320e-8	3.007	4.187e-8	3.004	1.113e-7	2.993

(b) du

Table 4.7: Errors and convergence rates, no slope limiter, $r = 2$.

N	L^1	rate	L^2	rate	L_∞	rate
8	3.581e-2	-	4.850e-2	-	1.508e-1	-
16	9.025e-3	1.988	1.085e-2	2.160	2.809e-2	2.424
32	9.523e-4	3.245	1.208e-3	3.167	4.072e-3	2.786
64	2.061e-4	2.208	2.934e-4	2.042	1.897e-3	1.102
128	7.352e-5	1.487	9.349e-5	1.650	5.255e-4	1.852
256	1.464e-5	2.329	1.823e-5	2.358	7.667e-5	2.777

(a) d

N	L^1	rate	L^2	rate	L_∞	rate
8	2.528e-2	-	3.673e-2	-	1.223e-1	-
16	9.280e-3	1.446	1.119e-2	1.716	2.492e-2	2.294
32	1.200e-3	2.951	1.526e-3	2.874	5.876e-3	2.085
64	3.066e-4	1.969	3.843e-4	1.989	1.129e-3	2.380
128	6.795e-5	2.174	8.516e-5	2.174	4.098e-4	1.462
256	1.178e-5	2.528	1.553e-5	2.455	1.216e-4	1.753

(b) du

Table 4.8: Errors and convergence rates, minmod slope limiter, $M = 0$, $r = 1$.

N	L^1	rate	L^2	rate	L_∞	rate
8	2.562e-2	-	3.568e-2	-	1.181e-1	-
16	9.650e-3	1.409	1.458e-2	1.290	4.251e-2	1.475
32	2.362e-3	2.031	3.217e-3	2.181	1.152e-2	1.883
64	4.840e-4	2.287	6.569e-4	2.292	2.902e-3	1.989
128	1.036e-4	2.225	1.408e-4	2.222	9.505e-4	1.611
256	2.490e-5	2.056	3.289e-5	2.098	1.875e-4	2.342

(a) d

N	L^1	rate	L^2	rate	L_∞	rate
8	4.458e-2	-	5.881e-2	-	1.398e-1	-
16	8.366e-3	2.414	1.265e-2	2.217	4.392e-2	1.670
32	2.255e-3	1.891	3.076e-3	2.040	1.051e-2	2.062
64	4.528e-4	2.317	6.621e-4	2.216	3.253e-3	1.693
128	1.037e-4	2.127	1.392e-4	2.250	7.749e-4	2.070
256	2.351e-5	2.141	3.164e-5	2.137	2.028e-4	1.934

(b) du

Table 4.9: Errors and convergence rates, minmod limiter, $M = 0$, $r = 2$.

where we set $g = 9.812\,\mathrm{m/s^2}$. For both numerical simulations we will use the local Lax-Friedrichs flux.

Dam Break

For the dam break problem the initial condition will be

$$d(x,0) = \begin{cases} D_0, & \text{if } x \le 0, \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad du(x,0) = 0.$$

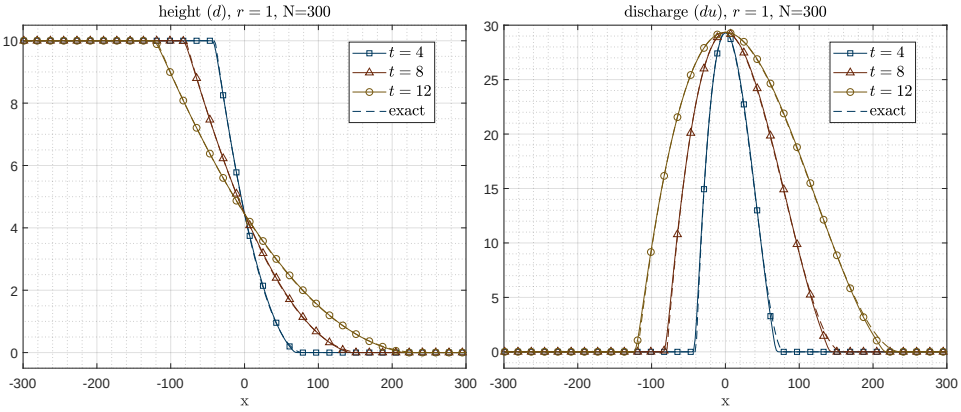
The water front, initially located at $x = 0$, will move to the right with speed $u + c$, where $c = 2a_0$ and $a_0 = \sqrt{gD_0}$. The exact solution, see [Bok05], is given by

$$d(x,t) = \begin{cases} \frac{1}{g}a_0^2, & \text{if} \quad x < -a_0 t, \\ \frac{1}{9g}\left(2a_0 - x/t\right)^2, & \text{if} \quad -a_0 t \le x < 2a_0 t, \\ 0, & \text{if} \quad x \ge 2a_0 t, \end{cases}$$

$$u(x,t) = \begin{cases} 0, & \text{if} \quad x < -a_0 t, \\ \frac{2}{3}\left(a_0 + x/t\right), & \text{if} \quad -a_0 t \le x < 2a_0 t, \\ 0, & \text{if} \quad x \ge 2a_0 t. \end{cases}$$

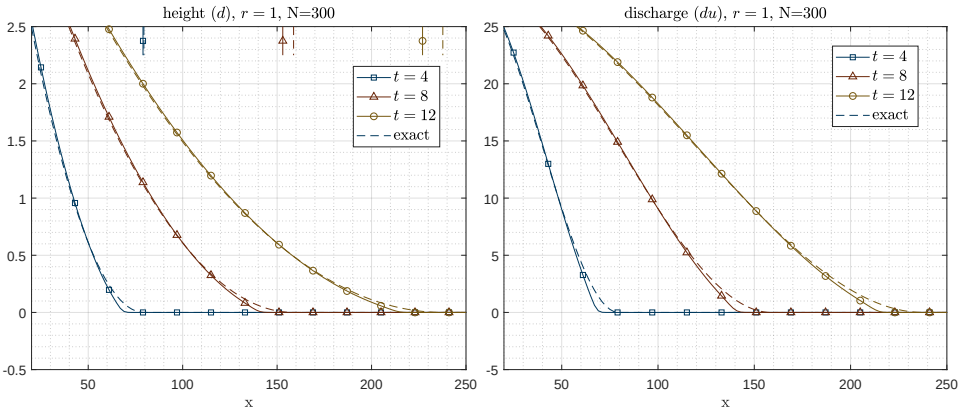
For the numerical experiment we set $D_0 = 10$, $x \in [-300, 300]$, $t \in [0, 12]$ and we use $N = 300$ elements. CFL condition is given by (4.14), and for this experiment is bounded below by $1.59\text{e}{-2}$. Since no water perturbation will reach the boundary we may simply use Dirichlet boundary conditions for both ends of the domain.

For the minmod slope limiter we set $M = 0$ (notice that the application of the slope limiter is mandatory). The solution at various time instances, as well as the location of the water front for piecewise linear polynomials, $r = 1$, can be seen in Figure 4.5. A magnification near the front, along with the numerical and exact locations of the front, can be seen in Figure 4.6. We observe that the numerical solution is satisfactorily close to the exact one.



(a) d (b) du

Figure 4.5: Numerical and exact solution for the dam break problem at times $t \simeq 4, 8, 12$, $N = 300$, $r = 1$. Solid lines with markers: numerical solution, dashed lines: exact solution.



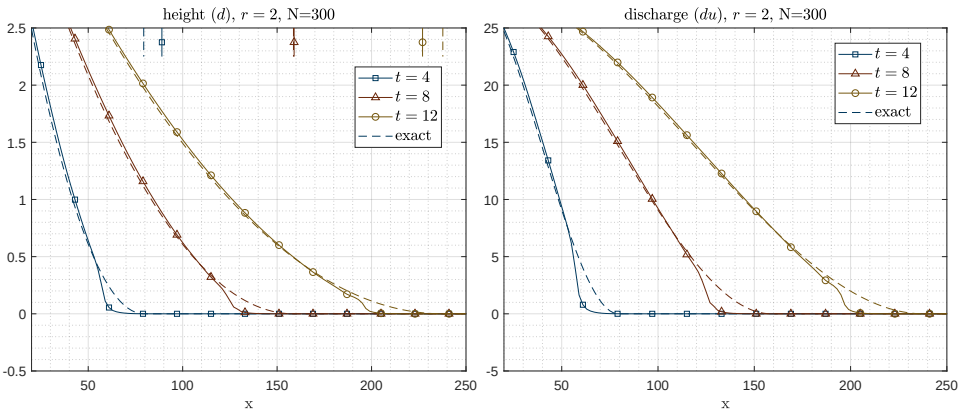
(a) d (b) du

Figure 4.6: Magnification of Figure 4.5 near the wet/dry front and front location. Solid lines with markers: numerical solution, dashed lines: exact solution, vertical lines: numerical and exact front location.

To assess the dispersion that is due to the slope limiter for higher order polynomials, we repeat the experiment using quadratic polynomials ($r = 2$). A magnification of the numerical and exact solution as well as the numerical and exact

front location can be seen in Figure 4.7. We see that the numerical solution is again close to the exact one, but now more dispersion is observed near the water front, effectively affecting the front location.

For both polynomial orders, the numerical solution converges to the exact as N grows, as expected.



(a) d (b) du

Figure 4.7: Magnification of numerical and exact solution for the dam break problem at time $t \simeq 4, 8, 12$, $N = 300$, $r = 2$. Solid lines with markers: numerical solution, dashed lines: exact solution, vertical lines: numerical and exact front location.

Drying

The initial conditions for the second Riemann test problem will be chosen such that a small dry patch is generated as the solution evolves. They are given by

$$d(x,0) = \begin{cases} d_l, & \text{if } x \leq 0, \\ d_r, & \text{otherwise} \end{cases} \quad \text{and} \quad u(x,0) = \begin{cases} u_l, & \text{if } x \leq 0, \\ u_r, & \text{otherwise} \end{cases}.$$

The solution is explicitly given by

$$h(x,t) = \begin{cases} d_l, & \text{if} \quad x \leq (u_l - a_l)t, \\ \frac{1}{9g}\left(u_l + c_l - x/t\right)^2, & \text{if} \quad (u_l - a_l)t < x \leq S_l t, \\ 0, & \text{if} \quad S_l t < x \leq S_r t, \\ \frac{1}{9g}\left(x/t - u_r + c_r\right)^2, & \text{if} \quad S_r t < x \leq (u_r + a_r)t, \\ d_r, & \text{if} \quad x > (u_r + a_r)t, \end{cases}$$

$$u(x,t) = \begin{cases} u_l, & \text{if} \quad x \leq (u_l - a_l)t, \\ \frac{1}{3}\left(u_l + c_l + 2x/t\right)^2, & \text{if} \quad (u_l - a_l)t < x \leq S_l t, \\ 0, & \text{if} \quad S_l t < x \leq S_r t, \\ \frac{1}{3}\left(2x/t + u_r - c_r\right)^2, & \text{if} \quad S_r t < x \leq (u_r + a_r)t, \\ u_r, & \text{if} \quad x > (u_r + a_r)t, \end{cases}$$

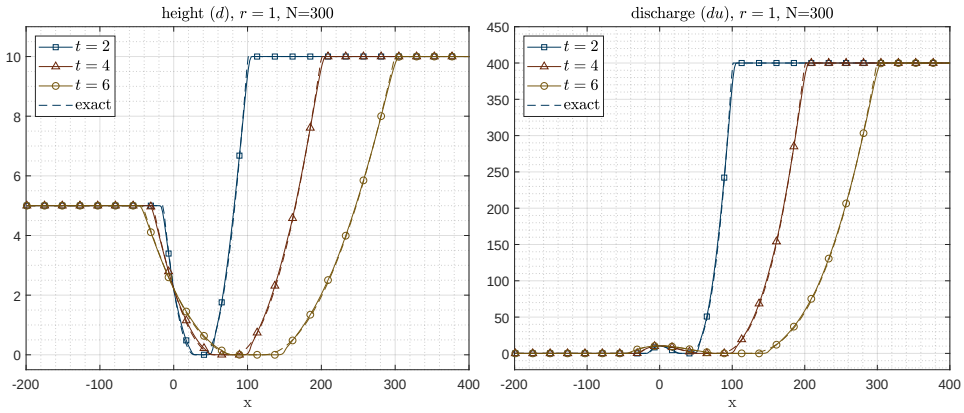
where $a_{l,r} = \sqrt{g\, d_{l,r}}$, $c_{l,r} = 2\, a_{l,r}$, $S_l = u_l + c_l$, and $S_r = u_r - c_r$. As time increases a dry region is (immediately) formed at $x = 0$ and two opposing travelling expansion waves are generated and travel away from and dry region. Drying occurs when $c_l + c_r - u_r + u_l < 0$.

For the numerical simulation we took $d_l = 5$, $d_r = 10$, $u_l = 0$, $u_r = 40$, $x \in [-200, 400]$, $t \in [0, 6]$ and $N = 300$ elements. As in the previous simulation, the solution remains constant near the boundary and we may use Dirichlet boundary conditions; we also set the minmod slope limiter threshold $M = 0$. CFL condition varies between $3\text{e}{-3}$ and $6\text{e}{-3}$ for almost all of the time steps. The solution at various time instances for linear polynomials, $r = 1$ can be seen in Figure 4.8, while a magnification near the dry region can be seen in Figure 4.9. We observe that the numerical solution is close to the exact, and that there is noticeable dispersion due to the minmod slope limiter especially near the dry-wet interface.

In Figure 4.10 we repeat the experiment using quadratic polynomials. At first, the numerical solution seems to be closer to the exact near the dry-wet interface. However, due to the dispersion introduced by the slope limiter, the actual height in the supposedly dry region is only $\mathcal{O}(10^{-2})$, which is worse than that for the linear polynomials (which was $\mathcal{O}(10^{-4})$).

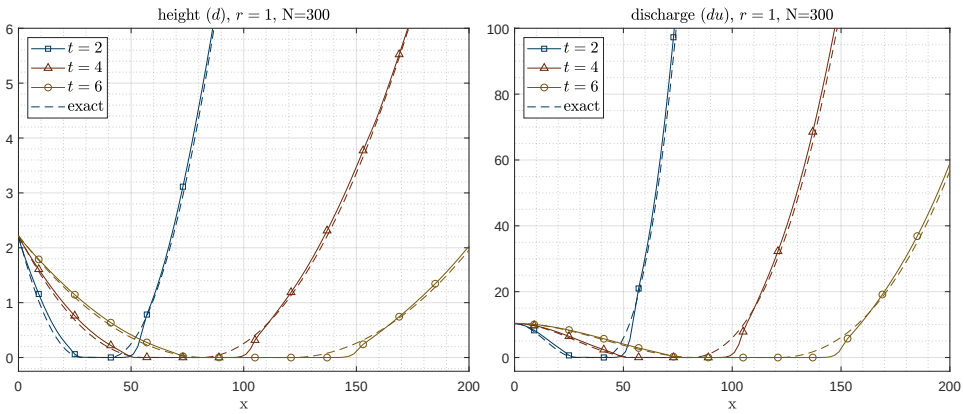
Finally we repeat the experiment using linear polynomials and $M = 2/3 \cdot 1$ in the minmod slope limiter. The solution can be seen in Figure 4.11. The numerical solution now actually becomes dry in the region of interest (that is $d < 10^{-6}$ which is the dry region threshold of our scheme). Also the dispersion of the numerical solution is smaller near the dry-wet front, since the slope limiter is applied fewer times. Notice however that in this case the CFL condition has to be chosen carefully, as described in the beginning of this section, for the scheme to be stable.

We have verified that the numerical solution converges to the exact as N grows for all the simulations printed in this section.



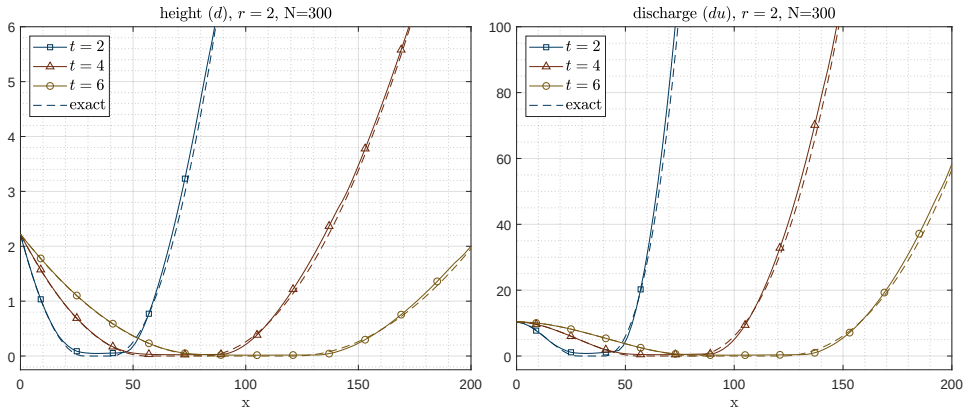
(a) d (b) du

Figure 4.8: Numerical and exact solution for the drying problem at times $t \simeq 2, 4, 6$, $N = 300$, $r = 1$. Solid lines with markers: numerical solution, dashed lines: exact solution.



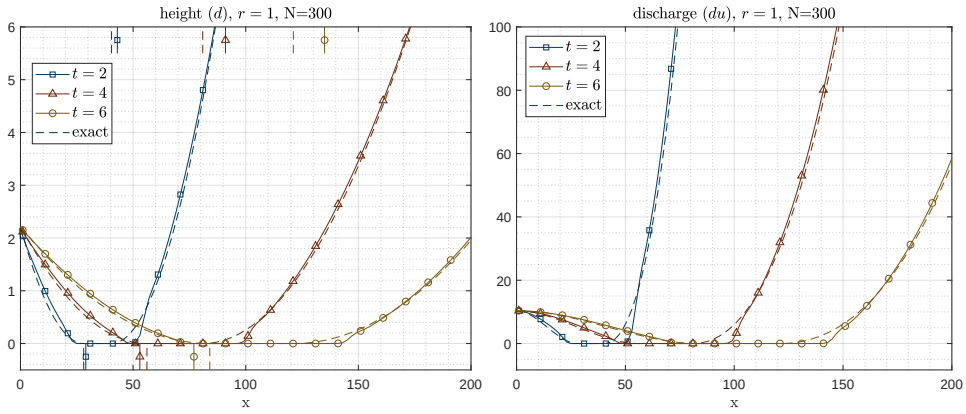
(a) d (b) du

Figure 4.9: Magnification of Firure 4.8 near the dry region



(a) d (b) du

Figure 4.10: Numerical and exact solution for the drying problem using quadratic polynomials. $t \simeq 2, 4, 6$, $N = 300$.



(a) d (b) du

Figure 4.11: Numerical and exact solution for the drying problem using linear polynomials and minmod slope limiter threshold $M = 2/3 \cdot 1$. $t \simeq 2, 4, 6$, $N = 300$, $r = 1$. Solid lines with markers: numerical solution, dashed lines: exact solution, bottom vertical lines: location of the wet-dry front, top vertical lines: location of the dry-wet front.

4.3.3 Parabolic bowl

In this section we consider an experiment with a periodic in time solution which has a moving wet/dry front. We consider a parabolic bottom of the form

$$\beta(x) = h_0(x/\alpha)^2,$$

where h_0 and α are constants. Analytical solutions, assuming frictional bottom, have been derived by Sampson et. al., [SES06]. Sampson assumed that the velocity u is a function of time only. By dropping the friction term we have

$$\begin{aligned} d(x,t) + \beta(x) &= h_0 - \frac{B^2}{4g}\cos(2\omega t) - \frac{B^2}{4g} - \frac{B}{2\alpha}\sqrt{\frac{8h_0}{g}}\cos(\omega t)x, \quad x_{\mathrm{l}} \le x \le x_{\mathrm{r}}, \\ u(x,t) &= B\sin(\omega t), \end{aligned} \qquad t > 0, \tag{4.15}$$

where $\omega = \sqrt{2gh_0}/\alpha$ and B is a given constant. Initial conditions are given by the formulas (4.15) for $t = 0$ and can be seen in Figure 4.12. The exact location of the wet/dry front is given by

$$x_{\mathrm{l,r}} = -\frac{B\omega\alpha^2}{2gh_0}\cos(\omega t) \mp \alpha.$$

Observe that the water height $d(x,t) + \beta(x)$ is a linear function of x for every t. Also, due to the absence of friction, d and u are periodic in time with period $2\pi/\omega$.

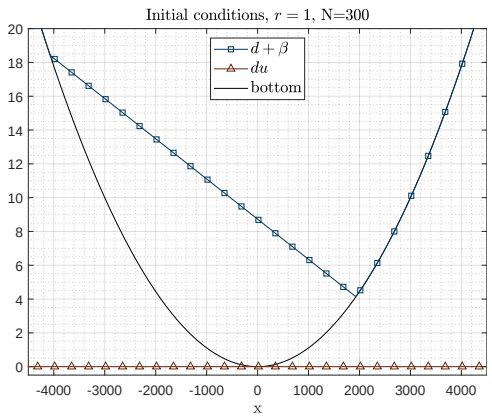
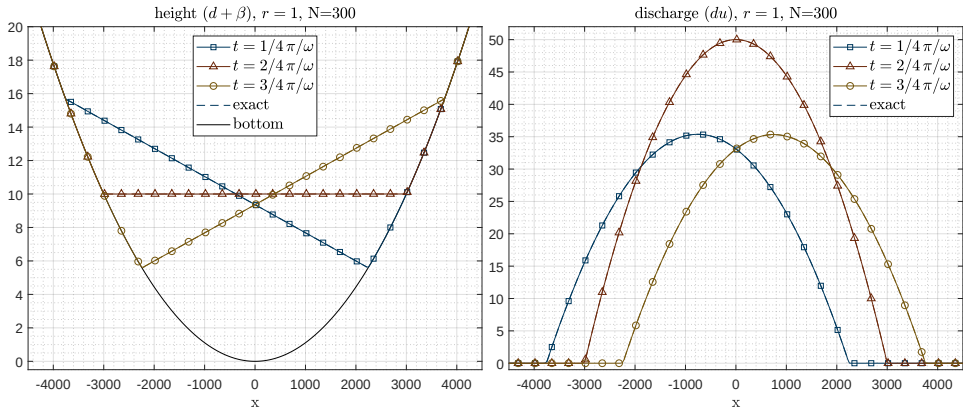


Figure 4.12: Initial conditions and bottom for the parabolic bowl problem

We perform the following numerical simulation, originally found in [LM09, §4.3] for frictional bottom, and later in [XZS10, §6.5] for the frictionless case. We let $x \in [-5000, 5000]$ and set $h_0 = 10$, $\alpha = 3000$, $B = 5$. Since the water does not reach the boundary we may again use Dirichlet boundary conditions. The simulation run up to $t = 6000$ using $N = 300$ cells and can be seen for various time instances in Figures 4.13 and 4.14. We observe that the numerical solution is close to the exact.



(a) $d+\beta$ (b) du

Figure 4.13: Numerical and exact solution for the parabolic bowl problem using linear polynomials and minmod slope limiter. $t \simeq 1/4\,\pi/\omega,\ 2/4\,\pi/\omega,\ 3/4\,\pi/\omega$, $N = 300$. Solid lines with markers: numerical solution, dashed lines: exact solution (indistinguishable).

4.3.4 An experiment with complex bottom topography

In the final section we will perform a novel experiment on a complex bottom topography that presents several computational difficulties. The bottom, seen in Figure 4.15(a), consist of two hills of different height (H_1, H_2) that are separated by a valley (V), it is smooth (C^∞) and is given by

$$\beta(x) = 1.33\exp(-50x^2) - 0.28\exp(-400(x-0.005)^2).$$

The initial conditions will be chosen such that no slope limiting will be required and consist of a Gaussian pulse centered at $x = 0$ for d and zero for u:

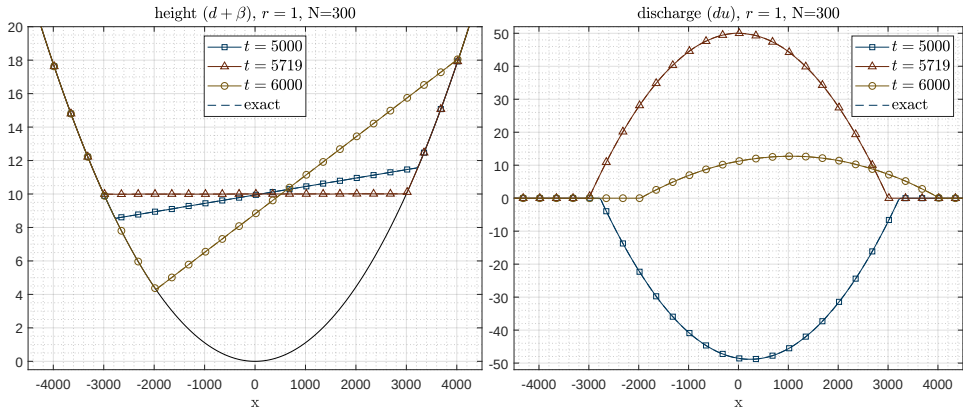
$$d(x,0) = 0.1\exp(-100x^2), \quad (du)(x,0) = 0.$$

For the numerical simulation we let $x \in [-4,4]$, $T = 3$ and use $N = 4000$ elements. The water is not disturbed near the boundary of the domain in our time of interest, so we simply use Dirichlet boundary conditions. In order to test the stability of the method near the dry region we will not utilize slope limiting. Notice that, despite the bottom being continuous, after its projection onto V_h and the procedure described in the well-balancing section, 4.2.1.1, the bottom becomes discontinuous at the cells adjacent to the wet-dry interface at H_1. (The same happens to H_2, though the discontinuity is too small to be noticeable in the figures.)

The exact solution is not known for this experiment. Some significant instances of the simulation can be seen in Figure 4.16, in particular:

4.16(a) At $t = 0.400$ the main pulse reaches and starts climbing H_1.

4.16(b) At $t = 0.480$ the main pulse goes over H_1.



(a) $d+\beta$ (b) du

Figure 4.14: Numerical and exact solution for the parabolic bowl problem using linear polynomials and minmod slope limiter. $t \simeq 5000,\ 5719\,(8.5\,\pi/\omega),\ 6000$, $N = 300$. Solid lines with markers: numerical solution, dashed lines: exact solution (indistinguishable).

4.16(c) At $t = 0.625$ the main pulse goes over H_2.

4.16(d) At $t = 0.800$ the bottom is completely flooded. A part of the water continues to travel to the right, while the other part is reflected (from H_2) to the left.

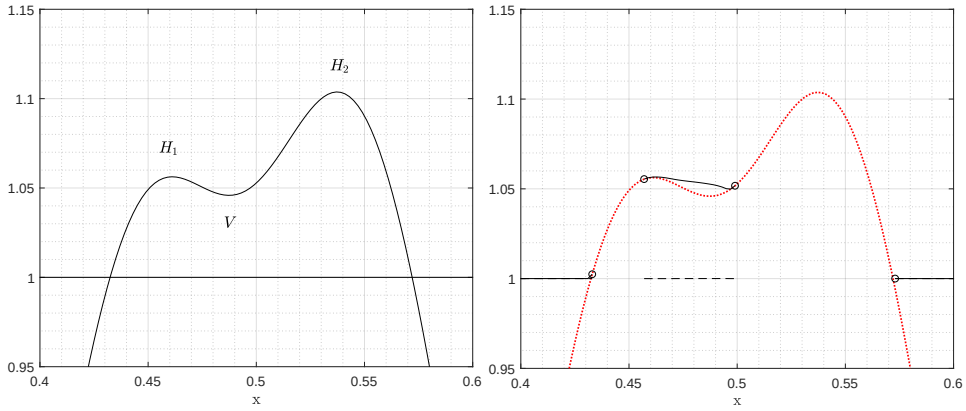
4.16(e) At $t = 1.135$ the reflected (left-traveling) pulse from H_2 overtakes H_1 (since H_2 is taller than H_1, the reflected mass is large enough for this purpose). Part of this pulse will continue traveling to the left, while the other part will be reflected to the right.

4.16(f) At $t = 1.675$ the right-traveling pulse reflected from H_1 (4.16(e)) reaches the maximum height on H_2 (runup), and is reflected to the left. The top and the right-hand parts of H_2 are now dry.

4.16(g) At $t = 2.115$ the left-hand part of H_1 is dry. The left-traveling pulse reflected from H_2 (4.16(f)) overtakes H_1 (since the reflected discharge from the instance of 4.16(e) happened to be large enough), and part of it will continue traveling to the left, while the other part will be reflected to the right. This procedure (of figures 4.16(e)–4.16(g)) will be repeated until the water height at the instance of figure 4.16(g) at the top of H_1 becomes less than 10^{-6}. Then the top of H_1 will be considered as dry and we will transition to an "oscillating lake" case.

4.16(h) At $t = 2.285$ the left-hand part of H_1 is wet again. Part of the water can be seen traveling leftwards.

The numerical solution at the final time, $t = 3$, can be seen in figure 4.15(b). Many properties of the numerical method can be tested in this experiment. First we

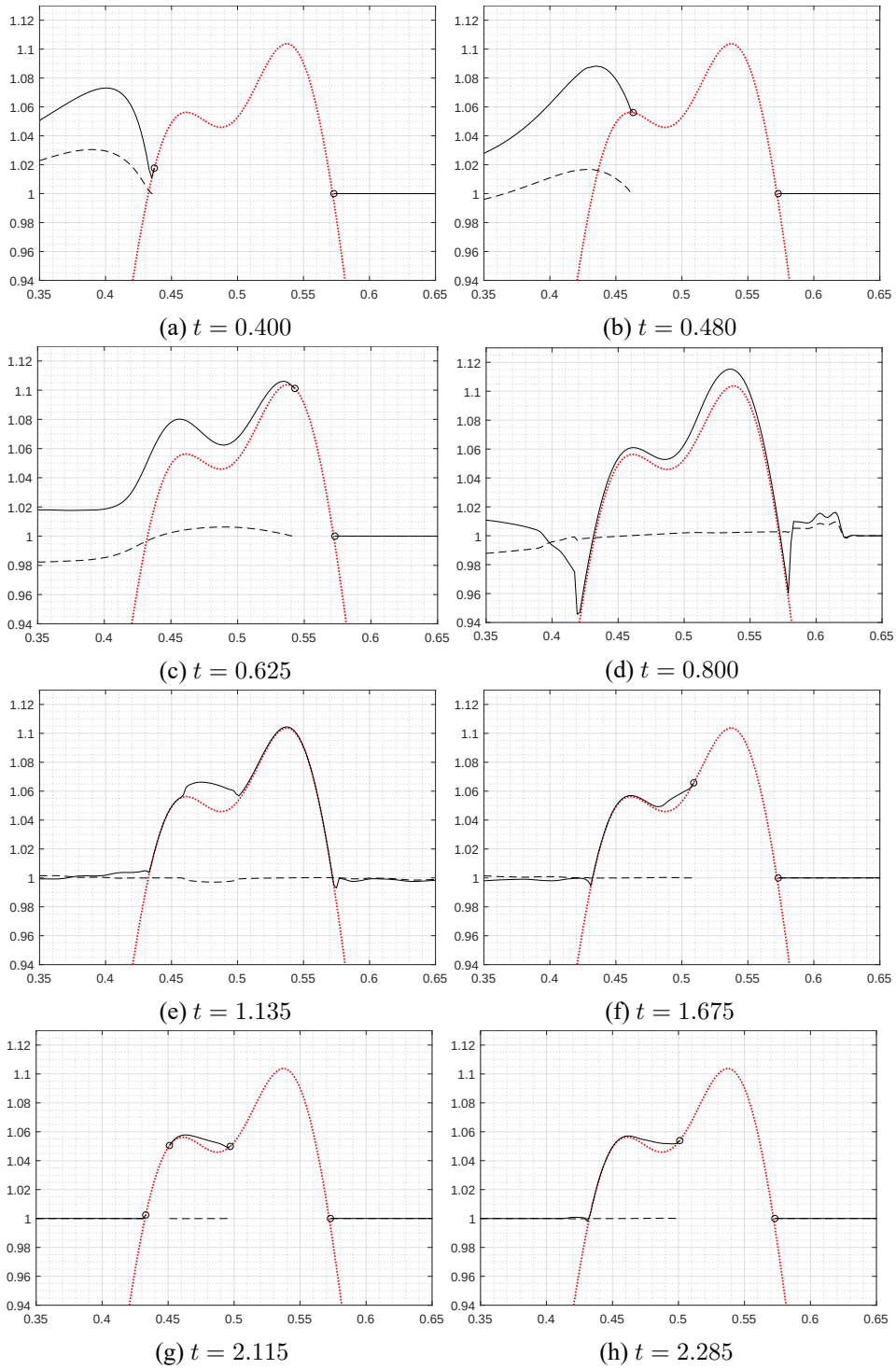


(a) Bottom (b) Numerical solution at time $t = 3$

Figure 4.15: Bottom and solution at final time step

observe that the water height, d, is horizontal at the wet-dry interface left of H_1 and right of H_2, indicating that the scheme is indeed well balanced. Many regions of the domain change between dry and wet during the simulation, testing the ability of the method to maintain the water height positive.

Finally, despite the initial conditions being chosen such that no shocks will be formed if the water propagated over a flat bottom (within our spatial domain of interest), we observe various type of discontinuities that are generated during the runup/rundown process. In particular: (a) After the reflection of the main pulse on H_2 at $t = 1$ a left-traveling shock is formed (the water is very shallow in this region). (b) During the draining of V two (almost) static shocks are formed near the feet of H_1 and H_2, where the water draining from V has a significant discharge when it enters the region of zero or negative velocity.



(a) $t = 0.400$ (b) $t = 0.480$

(c) $t = 0.625$ (d) $t = 0.800$

(e) $t = 1.135$ (f) $t = 1.675$

(g) $t = 2.115$ (h) $t = 2.285$

Figure 4.16: Flow over a complex bottom. Numerical solution ($d + \beta$: solid line, $du+1$: dashed lines, β: dotted lines, shore location: small circles) at various times.

Bibliography

[Ada11] K. Adamy, "Existence of solutions for a Boussinesq system on the half line and on a finite interval", *Discrete & Continuous Dynamical Systems - A* 29 (2011), 25–49, DOI: 10.3934/dcds.2011.29.25.

[Ami84] C. J. Amick, "Regularity and uniqueness of solutions to the Boussinesq system of equations", *Journal of Differential Equations* 54 (1984), 231–247, DOI: 10.1016/0022-0396(84)90160-8.

[AD12] D. C. Antonopoulos and V. A. Dougalis, "Numerical solution of the 'classical' Boussinesq system", *Mathematics and Computers in Simulation* 82 (2012), Nonlinear Waves: Computation and Theory-IX, WAVES 2009, 984–1007, DOI: 10.1016/j.matcom.2011.09.006.

[AD13] D. C. Antonopoulos and V. A. Dougalis, "Error estimates for Galerkin approximations of the "classical" Boussinesq system", *Mathematics of Computation* 82 (2013), 689–717, DOI: 10.1090/S0025-5718-2012-02663-9.

[AD16] D. C. Antonopoulos and V. A. Dougalis, "Error estimates for the standard Galerkin-finite element method for the shallow water equations", *Mathematics of Computation* 85 (2016), 1143–1182, DOI: 10.1090/mcom3040.

[AD17] D. C. Antonopoulos and V. A. Dougalis, "Galerkin-finite element methods for the shallow water equations with characteristic boundary conditions", *IMA Journal of Numerical Analysis* 37 (2017), 266–295, DOI: 10.1093/imanum/drw017.

[ADK19] D. C. Antonopoulos, V. A. Dougalis, and G. Kounadis, "On the standard Galerkin method with explicit RK4 time stepping for the Shallow Water equations", *IMA Journal of Numerical Analysis* (2019), DOI: 10.1093/imanum/drz033.

[ADM10] D. C. Antonopoulos, V. A. Dougalis, and D. E. Mitsotakis, "Galerkin approximations of periodic solutions of Boussinesq systems", *Bull. Greek Math. Soc* 57 (2010), 13–30.

[ADM17] D. C. Antonopoulos, V. A. Dougalis, and D. E. Mitsotakis, "Error estimates for Galerkin approximations of the Serre equations", *SIAM Journal on Numerical Analysis* 55 (2017), 841–868, DOI: 10.1137/16M1078355.

[ADM] D. C. Antonopoulos, V. A. Dougalis, and D. E. Mitsotakis, "Numerical methods for the Serre-Green-Naghdi equations over variable bottom", *(to appear)*.

[Aud+04] E. Audusse, F. Bouchut, M.-O. Bristeau, R. Klein, and B. t. Perthame, "A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows", *SIAM Journal on Scientific Computing* 25 (2004), 2050–2065, DOI: 10.1137/S1064827503431090.

[Bar04] E. Barthélemy, "Nonlinear shallow water theories for coastal waves", *Surveys in Geophysics* 25 (2004), 315–337, DOI: 10.1007/s10712-003-1281-7.

[BV94] A. Bermudez and M. E. Vázquez, "Upwind methods for hyperbolic conservation laws with source terms", *Computers & Fluids* 23 (1994), 1049–1071, DOI: 10.1016/0045-7930(94)90004-3.

[BDF94] R. Biswas, K. D. Devine, and J. E. Flaherty, "Parallel, adaptive finite element methods for conservation laws", *Applied Numerical Mathematics* 14 (1994), 255–283, DOI: 10.1016/0168-9274(94)90029-9.

[Bok05] O. Bokhove, "Flooding and Drying in Discontinuous Galerkin Finite-Element Discretizations of Shallow-Water Equations. Part 1: One Dimension", *Journal of Scientific Computing* 22 (2005), 47–82, DOI: 10.1007/s10915-004-4136-6.

[BCS04] J. L. Bona, M. Chen, and J.-C. Saut, "Boussinesq equations and other systems for small-amplitude long waves in nonlinear dispersive media: II. The nonlinear theory", *Nonlinearity* 17 (2004), 925–952, DOI: 10.1088/0951-7715/17/3/010.

[Bon+11] P. Bonneton, F. Chazel, D. Lannes, F. Marche, and M. Tissier, "A splitting approach for the fully nonlinear and weakly dispersive Green–Naghdi model", *Journal of Computational Physics* 230 (2011), 1479–1498, DOI: 10.1016/j.jcp.2010.11.015.

[Cha07] F. Chazel, "Influence of bottom topography on long water waves", *ESAIM: Mathematical Modelling and Numerical Analysis* 41 (2007), 771–799, DOI: 10.1051/m2an:2007041.

[Cia78] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, (Reprinted SIAM, 2002), North Holland, 1978.

[CBB07] R. Cienfuegos, E. Barthélemy, and P. Bonneton, "A fourth-order compact finite volume scheme for fully nonlinear and weakly dispersive Boussinesq-type equations. Part II: boundary conditions and validation", *International Journal for Numerical Methods in Fluids* 53 (2007), 1423–1455, DOI: 10.1002/fld.1359.

[Coc99] B. Cockburn, *Discontinuous Galerkin methods for convection-dominated problems*, 1999, URL: http://www-users.math.umn.edu/~cockburn/lecture_notes/DG-2.pdf (visited on 11/01/2019).

[CLS89] B. Cockburn, S.-Y. Lin, and C.-W. Shu, "TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws III: one-dimensional systems", *Journal of Computational Physics* 84 (1989), 90–113, DOI: 10.1016/0021-9991(89)90183-6.

[CS89] B. Cockburn and C.-W. Shu, "TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. II. General framework", *Mathematics of Computation* 52 (1989), 411–435, DOI: 10.1090/S0025-5718-1989-0983311-4.

[DM10] F. Dias and P. Milewski, "On the fully-nonlinear shallow-water generalized Serre equations", *Physics Letters A* 374 (2010), 1049–1053, DOI: 10.1016/j.physleta.2009.12.043.

[Dod98] N. Dodd, "Numerical model of wave run-up, overtopping, and regeneration", *Journal of Waterway, Port, Coastal, and Ocean Engineering* 124 (1998), 73–81, DOI: 10.1061/(ASCE)0733-950X(1998)124:2(73).

[Dou14] V. A. Dougalis, *Surface Water Waves: Mathematical Models, Solitary Waves*, Notes for a graduate course, Hellenic Open University, Patras, 2014 (in Greek).

[DDW75] J. Douglas, T. Dupont, and L. Wahlbin, "Optimal L_∞ error estimates for Galerkin approximations to solutions of two-point boundary value problems", *Mathematics of Computation* 29 (1975), 475–483, DOI: 10.1090/S0025-5718-1975-0371077-0.

[Dup73] T. Dupont, "Galerkin methods for first order hyperbolics: an example", *SIAM Journal on Numerical Analysis* 10 (1973), 890–899, DOI: 10.1137/0710074.

[GN76] A. E. Green and P. M. Naghdi, "A derivation of equations for wave propagation in water of variable depth", *Journal of Fluid Mechanics* 78 (1976), 237–246, DOI: 10.1017/s0022112076002425.

[GLN74] A. E. Green, N. Laws, and P. M. Naghdi, "On the theory of water waves", *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences* 338 (1974), 43–55, DOI: 10.1098/rspa.1974.0072.

[Gri+94] S. Grilli, R. Subramanya, I. Svendsen, and J. Veeramony, "Shoaling of solitary waves on plane beaches", *Journal of Waterway, Port, Coastal, and Ocean Engineering* 120 (1994), 609–628, DOI: 10.1061/(ASCE)0733-950X(1994)120:6(609).

[HK68] D. D. Houghton and A. Kasahara, "Nonlinear shallow fluid flow over an isolated ridge", *Communications on Pure and Applied Mathematics* 21 (1968), 1–23, DOI: 10.1002/cpa.3160210103.

[HPT11] A. Huang, M. Petcu, and R. Temam, "The one-dimensional supercritical shallow-water equations with topography", *Annals of the University of Bucharest (Mathematical Series)* 2 (LX) (2011), 63–82.

[Isr11] S. Israwi, "Large time existence for 1D Green-Naghdi equations", *Nonlinear Analysis: Theory, Methods & Applications* 74 (2011), 81–93, DOI: 10.1016/j.na.2010.08.019.

[KD19] G. Kounadis and V. A. Dougalis, *Galerkin finite element methods for the Shallow Water equations over variable bottom*, *(to appear in J. Comput. Appl. Math)*, 2019, arXiv: 1901.04230.

[Lan13] D. Lannes, *The Water Waves Problem: Mathematical Analysis and Asymptotics*, vol. 188, American Mathematical Society, Providence, RI, 2013, DOI: 10.1090/surv/188.

[LB09] D. Lannes and P. Bonneton, "Derivation of asymptotic two-dimensional time-dependent equations for surface water wave propagation", *Physics of Fluids* 21 (2009), 016601, DOI: 10.1063/1.3053183.

[LM09] Q. Liang and F. Marche, "Numerical resolution of well-balanced shallow water equations with complex source terms", *Advances in Water Resources* 32 (2009), 873–884, DOI: 10.1016/j.advwatres.2009.02.010.

[MM69] O. S. Madsen and C. C. Mei, "The transformation of a solitary wave over an uneven bottom", *Journal of Fluid Mechanics* 39 (1969), 781–791, DOI: 10.1017/S0022112069002461.

[MAT18] MATLAB, *version 9.4.0 (R2018a)*, The MathWorks Inc., Natick, Massachusetts, 2018.

[ML66] C. C. Mei and B. Le Méhauté, "Note on the equations of long waves over an uneven bottom", *Journal of Geophysical Research* 71 (1966), 393–400, DOI: 10.1029/JZ071i002p00393.

[MSM17] D. E. Mitsotakis, C. Synolakis, and M. McGuinness, "A modified Galerkin/finite element method for the numerical solution of the Serre-Green-Naghdi system", *International Journal for Numerical Methods in Fluids* 83 (2017), 755–778, DOI: 10.1002/fld.4293.

[NHF08] J. Nycander, A. M. Hogg, and L. M. Frankcombe, "Open boundary conditions for nonlinear channel flow", *Ocean Modelling* 24 (2008), 108–121, DOI: 10.1016/j.ocemod.2008.06.003.

[Per67] D. H. Peregrine, "Long waves on a beach", *Journal of Fluid Mechanics* 27 (1967), 815–827, DOI: 10.1017/S0022112067002605.

[Per72] D. H. Peregrine, "Equations for water waves and the approximations behind them", in *Waves on Beaches and Resulting Sediment Transport*, ed. by R. E. Meyer, New York: Academic Press, 1972, 95–121, DOI: 10.1016/B978-0-12-493250-0.50007-2.

[PT11] M. Petcu and R. Temam, "The one dimensional Shallow Water equations with Dirichlet boundary conditions on the velocity", *Discrete & Continuous Dynamical Systems - Series S* 4 (2011), 209–222, DOI: 10.3934/dcdss.2011.4.209.

[PT13] M. Petcu and R. Temam, "The one-dimensional shallow water equations with transparent boundary conditions", *Mathematical Methods in the Applied Sciences* 36 (2013), 1979–1994, DOI: 10.1002/mma.1482.

[QZ16] J. Qiu and Q. Zhang, "Stability, error estimate and limiters of discontinuous Galerkin methods", in *Handbook of Numerical Methods for Hyperbolic Problems*, ed. by R. Abgrall and C.-W. Shu, vol. 17, Handbook of Numerical Analysis, Elsevier, 2016, 147–171, DOI: 10.1016/bs.hna.2016.06.001.

[SES06] J. Sampson, A. Easton, and M. Singh, "Moving boundary shallow water flow above parabolic bottom topography", *Proceedings of the 7th Biennial Engineering Mathematics and Applications Conference, EMAC-2005*, ed. by A. Stacey, B. Blyth, J. Shepherd, and A. J. Roberts, vol. 47, 2006, C373–C387, DOI: 10.21914/anziamj.v47i0.1050, (visited on 10/16/2006).

[Sch81] M. E. Schonbek, "Existence of solutions for the Boussinesq system of equations", *Journal of Differential Equations* 42 (1981), 325–352, DOI: 10.1016/0022-0396(81)90108-X.

[Ser53a] F. Serre, "Contribution à l'étude des écoulements permanents et variables dans les canaux", *La Houille Blanche* (1953), 374–388, DOI: 10.1051/lhb/1953034.

[Ser53b] F. Serre, "Contribution à l'étude des écoulements permanents et variables dans les canaux", *La Houille Blanche* (1953), 830–872, DOI: 10.1051/lhb/1953058.

[Shi+11] M.-C. Shiue, J. Laminie, R. Temam, and J. Tribbia, "Boundary value problems for the shallow water equations with topography", *Journal of Geophysical Research: Oceans* 116 (2011), 1–22, DOI: 10.1029/2010JC006315.

[SO88] C.-W. Shu and S. Osher, "Efficient implementation of essentially non-oscillatory shock-capturing schemes", *Journal of computational physics* 77 (1988), 439–471, DOI: 10.1016/0021-9991(88)90177-5.

[SG69] C. H. Su and C. S. Gardner, "Korteweg-de Vries equation and generalizations. III. Derivation of the Korteweg-de Vries equation and Burgers equation", *Journal of Mathematical Physics* 10 (1969), 536–539, DOI: 10.1063/1.1664873.

[WB99] M. Walkley and M. Berzins, "A finite element method for the one-dimensional extended Boussinesq equations", *International Journal for Numerical Methods in Fluids* 29 (1999), 143–157, DOI: 10.1002/(SICI)1097-0363(19990130)29:2<143::AID-FLD779>3.0.CO;2-5.

[Whi74] G. B. Whitham, *Linear and Nonlinear Waves*, Wiley, New York, 1974.

[Xin17] Y. Xing, "Numerical Methods for the Nonlinear Shallow Water Equations", in *Handbook of Numerical Methods for Hyperbolic Problems*, ed. by R. Abgrall and C.-W. Shu, vol. 18, Handbook of Numerical Analysis, Elsevier, 2017, 361–384, DOI: 10.1016/bs.hna.2016.09.003.

[XS06] Y. Xing and C.-W. Shu, "High order well-balanced finite volume WENO schemes and discontinuous Galerkin methods for a class of hyperbolic systems with source terms", *Journal of Computational Physics* 214 (2006), 567–598, DOI: 10.1016/j.jcp.2005.10.005.

[XZS10] Y. Xing, X. Zhang, and C.-W. Shu, "Positivity-preserving high order well-balanced discontinuous Galerkin methods for the shallow water equations", *Advances in Water Resources* 33 (2010), 1476–1493, DOI: 10.1016/j.advwatres.2010.08.005.