JACOBS
UNIVERSITY

A numerical investigation of optimal balance for rotating shallow water flow

by

Gökce Tuba Masur

a Thesis submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy
in Mathematics

Approved Dissertation Committee:

Prof. Dr. Marcel Oliver (Chair)
Jacobs University Bremen, Germany.

Prof. Dr. Ulrich Achatz
Goethe University Frankfurt, Germany.

Prof. Dr. Sergey Danilov
Jacobs University Bremen &
Alfred Wegener Institute, Germany.

Date of Defense: October 7, 2021

Department of Mathematics & Logistics

JACOBS UNIVERSITY

Statutory Declaration

Family name, Given/First name	Masur, Gökce Tuba
Matriculation number	20329758
What kind of thesis are you submitting: Bachelor-, Master- or PhD-Thesis	PhD-Thesis

English: Declaration of Authorship

I hereby declare that the thesis submitted was created and written solely by myself without any external support. Any sources, direct or indirect, are marked as such. I am aware of the fact that the contents of the thesis in digital form may be revised with regard to usage of unauthorized aid as well as whether the whole or parts of it may be identified as plagiarism. I do agree my work to be entered into a database for it to be compared with existing sources, where it will remain in order to enable further comparisons with future theses. This does not grant any rights of reproduction and usage, however.

The Thesis has been written independently and has not been submitted at any other university for the conferral of a PhD degree; neither has the thesis been previously published in full.

German: Erklärung der Autorenschaft (Urheberschaft)

Ich erkläre hiermit, dass die vorliegende Arbeit ohne fremde Hilfe ausschließlich von mir erstellt und geschrieben worden ist. Jedwede verwendeten Quellen, direkter oder indirekter Art, sind als solche kenntlich gemacht worden. Mir ist die Tatsache bewusst, dass der Inhalt der Thesis in digitaler Form geprüft werden kann im Hinblick darauf, ob es sich ganz oder in Teilen um ein Plagiat handelt. Ich bin damit einverstanden, dass meine Arbeit in einer Datenbank eingegeben werden kann, um mit bereits bestehenden Quellen verglichen zu werden und dort auch verbleibt, um mit zukünftigen Arbeiten verglichen werden zu können. Dies berechtigt jedoch nicht zur Verwendung oder Vervielfältigung.

Diese Arbeit wurde in der vorliegenden Form weder einer anderen Prüfungsbehörde vorgelegt noch wurde das Gesamtdokument bisher veröffentlicht.

August 23, 2021

Date Signature

To my family...

Acknowledgements

Throughout my research and writing this thesis, many people have contributed in several different ways, but first, I want to express my deepest appreciation to my supervisor Prof. Marcel Oliver for giving me an opportunity to join his research group and sharing his broad scientific knowledge. He was also very kind to share his experience and thoughts in various matters not directly related to science. While I was having hardships in my project and personal life, he supported me by giving enough time to deal with them. I would also like to extend my thank to him for introducing me this thesis topic and providing PDE-based approach formulations in Section 4.3.2. Without his invaluable suggestions, this thesis would not have the shape that it is in now.

Many thanks also go to everyone in our research group, I benefited the academic atmosphere and the friendship. I am especially grateful to Dr. Stefan Juricke for his generous time to answer all my numerical-based questions and for keeping his office door open whenever I needed. The special thanks also go to Dr. Anton Kutsenko for introducing me a new scientific problem each time that we saw each other, and to Dr. Haidar Mohamad for his valuable comments and discussion on Chapter 3.

I greatly appreciate my thesis committee members, Prof. Ulrich Achatz, Dr. Gualtiero Badin (old member), Prof. Sergey Danilov for their time to join my committee meetings, to read this thesis and helpful comments. I especially acknowledge Dr. Gualtiero Badin for always being encouraging and supportive, and Prof. Ulrich Achatz for giving a Postdoc position in his research group, which makes me enthusiastic.

I am thankful to the funding provided by Deutsche Forschungsgemeinschaft through *TRR* 181 *Energy Transfers in Atmosphere and Ocean* to help me continue this project, and to Jacobs University Bremen for providing facilities.

At last but not least, there is no word to express my thanks to my husband Dr. Qaisar Latif and my little son Arif Latif. During my research, my husband as a mathematician listened my ideas and contributed with insightful suggestions. His time to discuss the theoretical part in Chapter 3 and to give feedback the first draft of this thesis was extremely helpful. I am grateful to my son for making our life colourful.

Contents

List of Figures

List of Tables

Abstract

Balancing geophysical flows is a procedure to decompose the system state into a balanced flow and unbalanced gravity-wave components. These gravity waves are generated spontaneously, orographically or via forcing, and presumed to contribute significantly to dissipation of energy in the ocean. This energy route is understood to a limited extent, and our understanding needs to be improved to develop more accurate climate models. For this purpose, accurate gravity-wave diagnostics are crucial. In this thesis, we focus on the decomposition method called *optimal balance*.

Optimal balance was introduced by Viúdez and Dritschel (2004) in a special numerical setting where potential vorticity is advected by a semi-Lagrangian scheme, so it was specifically named *optimal potential vorticity balance*. Optimal balance works through adiabatically deforming the nonlinear model into its linear form where mode decomposition is exact. This leads to a boundary value problem in time, which is solved by iterative backward-forward nudging scheme. At the linear-end boundary, gravity waves are removed. At the nonlinear-end boundary, a base-point coordinate is restored. This process is repeated until convergence to a balanced state that is characterised by the given base-point coordinate. Global geophysical ocean models use traditional variables (velocity and tracer fields) rather than potential vorticity and complementary variables, so that optimal balance should be performed in these variables. In this thesis, we therefore apply optimal balance to the f-plane shallow water model governed by the primitive velocity-height variables as a first step toward optimal balance for full ocean models.

Our model implementation is based on a pseudo-spectral scheme in velocity-height variables on a spatially periodic domain; nevertheless, kinematic potential vorticity inversion formulas appear if potential vorticity is set as base point. The core of this thesis focuses on a systematic investigation of design parameters included in our numerical setting: integration time scale, ramp function, linear projector at the linear end, and base-point coordinate at the nonlinear end. We also explored numerically the applicability of viscosity term, the sensitivity of convergence tolerance, diagnostic tools, the initial-condition structure. Out of all parameters, the linear projector and the base point are found as being the most critical. The use of the linear oblique projector and the base point potential vorticity is the best combination. The oblique projector spectrally maps the flow onto its Rossby-wave mode along its gravity-wave mode, and it can be reformulated as a PDE-based projector, which corresponds to preservation of linear potential vorticity. As base point, the potential vorticity is advantageous with its fast convergence. With the choice of this combination, we support the use of the potential vorticity-preserving end-point conditions in the balancing procedure. On the other hand, the height field as base point gives good quality balance with some restrictions, so that it is also a prominent candidate as base point, which is convenient for models formulated in primitive variables.

The numerical investigation in the thesis points to convergence issues of the backward-

forward nudging scheme. Though the method always gives well-balanced flows, the sequence of iterates produced by the nudging scheme converges to the base point only up to a small residual that does not vanish as the number of iterations is increased. Working on a lower-dimensional model, we split the overall error into a balance error and a termination residual. We prove that the termination residual is of the same small order as the balance error. As a result, the nudging iterates "quasi-converge," i.e., the termination residual decreases at algebraic order in the time-scale separation parameter.

We conclude that optimal balance an accurate balance-imbalance decomposition in primitive variables. Its simplicity comes from its applicability as a diagnostic tool over existing numerical codes without the need of any asymptotic analysis. We expect that the behaviour of the different linear and nonlinear end boundary condition will carry over when implementing optimal balance for complicated systems such as the models on the sphere and global ocean models. In practice, the PV-based conditions will be of particular importance as they can be implemented by solving an elliptic PDE on general domains.

Chapter 1

Introduction

In large-scale ocean dynamics, geophysical flows at mid-latitudes are dominated by a balance between the pressure gradient and the Coriolis force in the horizontal direction, and by the gravity and buoyancy forces in the vertical direction, which are respectively called geostrophic balance and hydrostatic balance. These flows evolve over long-time scales and are described as slow balanced flow. The large-scale dynamics of the ocean is predominantly balanced; however, on small scales there are also gravity waves that evolve over short-time scales and are described as fast unbalanced waves. In the literature, the balanced flow is found under terms such as Rossby waves, vortex dynamics, vortical flow, slow-geostrophic mode and so on; terms used to represent the unbalanced flow are gravity wave, gravity-wave mode, fast-ageostrophic mode and so on.

The balanced flow and unbalanced wave evolving on different time scales can be separated from each other. For a linear flow, this separation is exact, where the flow is divided into balanced Rossby-wave and unbalanced gravity-wave components. For a nonlinear flow, explicit separation is not possible, as gravity waves are excited spontaneously by the nonlinear coupling between the different modes. If linear-mode decomposition is used to describe a nonlinear state, the nonlinear balanced (geostrophic) component contains not only linear Rossby waves, but also a contribution from the linear gravity-wave modes that is "slaved" to the Rossby-wave modes. To clarify the terminology, we call the nonlinear unbalanced component as unbalanced gravity waves throughout the thesis.

The strength of gravity-wave excitation is characterised by the time-scale separation parameter ε, which is, in our work, Rossby number

$$\varepsilon = \frac{U}{fL}$$

with horizontal velocity scale U, horizontal length scale L, and Coriolis frequency f. This number ε indicates the effect of Earth's rotation on the fluid, and in terms of time scales, it characterises the magnitude of the slow dynamics at frequency $1/T$ relative to the Coriolis frequency f. When ε takes relatively large values, the slow-fast dynamics are strongly coupled resulting in a substantial amount of gravity-wave emission. This occurs, e.g., in the equatorial ocean and in regions close to strong fronts. In this regime, the flows with frequency equal to or larger than f are excited and classified under "unbalanced flow". When ε takes smaller values, $\varepsilon \ll 1$, the rotation effects become dominant and near-geostrophic balance persists over long-time scales. In this regime, the coupling between motion on the two-time scales is weak and gravity-wave emission is small with respect to ε. The flows excited with frequency much smaller than f are, then, identified as "balanced

flow". In our work, we mainly focus on the small-Rossby-number regime, where the two time scales can be distinguished due to the large frequency gap.

The gravity-wave-permitting nature of geophysical fluid models makes the flow-wave decomposition to characterise balance flows a long-standing issue, which we also deal with in this thesis. The decomposition is important for several reasons:

i) Flows completely void of (geostrophic) gravity waves fail to capture accurate dynamics for ageostrophic phenomena such as intense mesoscale fronts. An improved understanding of gravity-wave emission and activity is crucial to improve our understanding of such processes (Vanneste and Yavneh, 2004).

ii) Available numerical weather prediction or climate models require to be initialised by near-balanced flows. Starting with flows disturbed by spurious gravity waves leads to erroneous short-range forecasts (Lynch, 2006).

iii) The effect of gravity waves can be seen on larger scales despite their small scale structure, so that unresolved gravity waves contributes significantly to the uncertainty of weather forecast (Kim et al., 2003).

iv) The bulk of energy in the ocean is assumed to be dissipated by gravity waves and this energy transfer is not well understood. The energy transfer can be included in the global climate models after the identification of its route to provide accurate climate forecasts (von Storch et al., 2019).

This thesis contributes to the development of applicable diagnostics that accurately quantify gravity waves in geophysical flows.

The characterisation of balanced flows is usually done by asymptotic analysis. We, here, work with the purely numerical method "optimal balance". The numerical study of optimal balance is achieved, earlier than the theoretical results, in the context of geophysical flow models by Viúdez and Dritschel (2004). The novelty of their work rests in constructing balanced flows implicitly while a specified potential vorticity is taken as a base point coordinate (or fixed reference field), so that their work is called optimal potential vorticity (PV) balance. The application of optimal balance involves solving a boundary value problem in time, which is done via an iterative backward-forward nudging scheme. The formulation of Viúdez and Dritschel (2004) requires a special numerical scheme formulated in geostrophic-ageostrophic variables, where the q fields is advected in the Lagrangian form; the other fields in the Eulerian form. Operational global models, however, do not employ these geostrophic-ageostrophic variables. The implementation of optimal PV balance with its scheme on the global models can also be computationally expensive despite very strong convergence of the method. This drawback will persist with optimal balance in primitive variables!

The theoretical study of optimal balance is achieved in a finite-dimensional model analogous to the evolution of single-column fluid in a specific regime, which is derived in Section 2.3.1. After optimal balance is realised as adiabatic invariance of slow manifolds (Cotter and Reich, 2006), the improvement in the later work (Gottwald et al., 2017) provided exponentially small residual of the method in the same model with rigorous settings. The theoretical results are discussed in details in Section 2.4.

In this thesis, we apply the theory of optimal balance in such a way that it can potentially provide diagnostics for global models without the requirement of a special numerical

treatment. To explore the concept, we choose to work in a simple setting: a single-layer rotating shallow water model moving on the f-plane in velocity-height variables. We implement the scheme over an existing pseudo-spectral shallow water code. The numerical implementation of optimal balance comes together with several design parameters, which are thoroughly examined. Among a wide range of parameters, our numerical results emphasised the significance in the selection of two parameters: the projector at the linear-end boundary and base point at the nonlinear-end boundary. We find that optimal balance works with PV-based projectors – which are the strongest choice – or with primitive variable-based projectors which are useful for general domains and global models, which all are reported and compared in Chapter 6.

While Viúdez and Dritschel (2004) report rapid convergence of the backward-forward nudging scheme, the proof of convergence to the solution of the boundary value problem is, yet, an open question. In this thesis, we provide initial theoretical results to this issue in the finite-dimensional setting that the other theoretical results are also built upon. According to our results, convergence does not occur in the strict mathematical sense: The nudging scheme gives sufficiently close consecutive iterates, but as the number of iterates gets larger, a small residual does not disappear. However, we can show that this termination residual will be small alongside with the balance error, and is of algebraic order in ε.

1.1 Thesis Outline

Chapter 2 introduces the general concept of balance in a generic slow-fast model. We discuss an infinite-dimensional model, the rotating shallow water model on the f-plane, and finite-dimensional models, an analogue evolutionary model of single-column fluid and the spring-mas pendulum. Optimal balance is introduced in the framework of the generic model with precise explanation of each component. The theoretical work related on the finite-dimensional model is recalled, and as an initiative implementation, the optimal PV balance is described.

Chapter 3 focuses on the quasi-convergence of the backward-forward nudging scheme in the context of the finite-dimensional system.

Chapter 4 deals with the application of optimal balance to the rotating shallow water model on the f-plane, in primitive velocity-height variables. After analysing the time scales of the model, the optimal balance boundary value problem is constructed. We suggest different choices for the formulation of the boundary conditions to be tested numerically.

Chapter 5 describes the experimental set-up. The numerical implementation of the problem over an available code, the choice of ramp function and the choice of initial conditions to start up test cases are introduced. To diagnose the performance of optimal balance, a special diagnostics, named "diagnosed imbalance", is used. Gathering all components, we describe the complete numerical set-up.

Chapter 6 reports the systematic investigation of the behaviour of optimal balance depending different combinations of the design parameters. We identify the best combination which yields high-quality balance with reasonable convergence.

Chapter 7 concludes this thesis with the highlights of the theoretical and numerical aspects and the discussion of possible future works.

Appendix A includes additional simulation results, mostly in the quasi-geostrophic scaling limit, which are not presented in Chapter 6.

Chapter 2

Model and method description

In this chapter, we describe several mathematical models and the method of optimal balance. We initially formulate an abstract slow-fast system to review the general concept of balance. Then, two concrete examples, the shallow water equations and a finite-dimensional model are introduced. We then formulate optimal balance on the abstract model. As last, relevant theoretical and numerical studies on these models are revised, as a starting point for our original work.

2.1 A class of abstract slow-fast systems

The geophysical fluid models can be expressed in a general compact form. If the compact form is linear, the time-scale separation of fluid dynamics is exact, and we denote fluid dynamics with slow geostrophic component s and fast ageostrophic components $\boldsymbol{f}$. We keep the same s-$\boldsymbol{f}$ notation for the nonlinear form, but the components are, here, separated only approximately, which will be explained clearly later on. The compact two-time-scale form is, then, written as

$$\partial_t s = \mathcal{N}_s(s, \boldsymbol{f}) \,, \tag{2.1.1a}$$

$$\partial_t \boldsymbol{f} = \frac{1}{\varepsilon}\mathcal{L}\boldsymbol{f} + \mathcal{N}_{\boldsymbol{f}}(s, \boldsymbol{f}) \,, \tag{2.1.1b}$$

where ε is a small-scale separation parameter, $\mathcal{N}_s$ and $\mathcal{N}_{\boldsymbol{f}}$ are nonlinear operators, and $\mathcal{L}$ is a linear operator with purely imaginary eigenvalues. The nonlinear operator represents, for example, the advection term in fluid dynamical equations while the linear operator causes oscillatory behaviour of fast variables. The rotating shallow-water model and the Euler-Boussinesq equations can be written in this compact form, see respectively Dritschel et al. (2017) and Franzke et al. (2019).

2.1.1 Balanced flow components

In geophysical flows, we separate the dynamical fields into a geostrophic, a balanced ageostrophic and unbalanced components, as it is expressed in Danioux et al. (2012). Regarding the compact form (2.1.1), the general flow is written as

$$s + \boldsymbol{f}_{\text{bag}} + \boldsymbol{f}_{\text{unb}} \,.$$

The balanced ageostrophic component $\boldsymbol{f}_{\text{bag}}$ is given by higher-order corrections to the leading-order geostrophic component s. This component is obtained systematically in

powers of time-scale separation parameter ε. In this context, the parameter is the Rossby number. The remaining part of $\boldsymbol{f}$ forms the unbalanced ageostrophic component $\boldsymbol{f}_{\mathrm{unb}}$, which we call the unbalanced flow or gravity waves. We will explain how to derive $\boldsymbol{f}_{\mathrm{bag}}$ as a power series in ε and how to determine the amplitude of $\boldsymbol{f}_{\mathrm{unb}}$ in the upcoming sections. The balanced flow is, then, identified by the geostrophic component s for linear flows, where the mode separation is exact, and by $s + \boldsymbol{f}_{\mathrm{bag}}$ for nonlinear flows, where the nonlinear-mode interaction is present. The main point of the thesis is to characterize balance for nonlinear flows.

2.1.2 Balance relation and balanced model

The balanced flow $s + \boldsymbol{f}_{\mathrm{bag}}$ is described by a relation between the slow component s and the fast component $\boldsymbol{f}$ at an instantaneous time. By this relation, the dynamics of the flow is reduced to expressed by fewer degrees of freedom, generally by a single dependent variable, the slow variable s. The geostrophic balance is the least accurate member of these relations at leading order, which excludes the fast variable $\boldsymbol{f}$. For higher-order relations, the nonlinear interactions $\mathcal{N}_s(s, \boldsymbol{f}_{\mathrm{bag}})$ and $\mathcal{N}_{\boldsymbol{f}}(s, \boldsymbol{f}_{\mathrm{bag}})$ are crucial to describe minimal ageostrophic features, $\boldsymbol{f}_{\mathrm{bag}}$. The *balance relation* or *filtering condition* is defined as a functional map $G_n : s \to \boldsymbol{f}_{\mathrm{bag}}$ such that

$$\boldsymbol{f}_{\mathrm{bag}} = G_n(s)\,, \tag{2.1.2}$$

where n is the order of the relation. As G_n parameterizes a slow manifold by s, we call s *base-point coordinate* or *base variable*. In other words, the balance relation G_n slaves the component $\boldsymbol{f}_{\mathrm{bag}}$ to the slow variable s, so that $\boldsymbol{f}_{\mathrm{bag}}$ is also referred to as *slaved variable*.

The relation G_n has the properties of being kinematic or diagnostic. Due to being kinematic, the relation does not involve any explicit time dependency: It is free of any expression of time, time derivatives and time integrals. At the same time, it is diagnostic in the sence that it can reconstruct the complete phase space coordinates when the slow variable s is known at some certain time t, but any time-dependent information of s stays unknown (McIntyre, 2015). The balance relation, thus, does not provide the evolutionary state of the flow.

A balance model, on the other hand, approximates the evolutionary process of the whole dynamics providing only slow-time evolution. It consists of a single prognostic equation, and the simplest model is obtained by inserting $G_n(s)$ into the slow equation (2.1.1a), that is,

$$\partial_t s = \mathcal{N}_s(s, G_n(s))\,. \tag{2.1.3}$$

The single variable s is enough to initialise the balanced model and provide the dynamics in the extended phase space. Balance model hierarchies are, then, constructed employing increasingly accurate relations G_n.

To derive higher-order balance relations, we combine the equations in the model (2.1.1) using G_n instead of f, then we obtain the superbalance relation,

$$\varepsilon\mathcal{N}_s(s, G_n) \cdot \partial_s G_n + \mathcal{L}G_n - \varepsilon\mathcal{N}_{\boldsymbol{f}}(s, G_n) = 0\,. \tag{2.1.4}$$

This superbalance relation is approximated in increasing order using different approaches, which lead to next-order accurate G_n. The one of the two approaches that we will introduce

is to expand $\boldsymbol{f}_{\text{bag}}$ in power series of ε, i.e.,

$$\boldsymbol{f}_{\text{bag}} = \varepsilon \boldsymbol{f}_1 + \varepsilon^2 \boldsymbol{f}_2 + \cdots + \varepsilon^n \boldsymbol{f}_n = \sum_{i=1}^{n} \varepsilon^i \boldsymbol{f}_i(s)\,. \tag{2.1.5}$$

The variable s is updated at each step according to the obtained $\boldsymbol{f}_{\text{bag}}$. As $\boldsymbol{f}_{\text{bag}}$ has zero leading order, the leading-order model is

$$\boldsymbol{f}_{\text{bag}} = 0, \qquad \partial_t s = \mathcal{N}_s(s, 0)\,.$$

The model at $O(\varepsilon)$ becomes

$$\boldsymbol{f}_{\text{bag}} = \varepsilon \mathcal{L}^{-1} \mathcal{N}_{\boldsymbol{f}}(s, 0), \qquad \partial_t s = \mathcal{N}_s(s, \boldsymbol{f}_{\text{bag}})\,.$$

The pattern of $\boldsymbol{f}_{\text{bag}}$ becomes clear at $O(\varepsilon^2)$, so the model at nth ε-order yields

$$\boldsymbol{f}_{\text{bag}} = \varepsilon \mathcal{L}^{-1}\Big(\mathcal{N}_{\boldsymbol{f}}(s, \boldsymbol{f}_n) - \mathcal{N}_s(s, \boldsymbol{f}_n) \cdot \partial_s \boldsymbol{f}_n\Big), \qquad \partial_t s = \mathcal{N}_s(s, \boldsymbol{f}_{\text{bag}})\,.$$

As the second approach, the superbalance equation (2.1.4) is approximated iteratively, and the $(n+1)$th iterate gives

$$G_{n+1} = \varepsilon \mathcal{L}^{-1}\Big(\mathcal{N}_{\boldsymbol{f}}(s, G_n) - \mathcal{N}_s(s, G_n) \cdot \partial_s G_n\Big)\,,$$

where G_n is known from the previous iteration. In the same manner, a model hierarchy follows at each iteration step. We refer Warn et al. (1995) for these approaches and Vanneste (2013) for a review in other type of derivations.

2.1.3 Slow manifold

Slow manifold is a subset of phase space, not a manifold in topological sense, more precisely, it is an invariant manifold tangent to slow subspace of equilibrium points for nonlinear dynamical systems see Verhulst (1990). The notion of the slow manifold is established by the balance relation G_n and the corresponding balanced model is understood as a projection of the whole system dynamics on this slow manifold. The whole concept of balance is, then, reduced to find a flow state placed on the slow manifold.

The system dynamics on the slow manifold has contribution of only slow balanced motion but of no fast oscillations. The slow manifold and these fast oscillations cannot, thus, theoretically coexist. The manifold is unique and invariant for dissipative systems; however, there is no slow object with these properties for nondissipative systems. The uniqueness and invariance of slow manifold are relaxed due to nonexistence of explicit scale separation, while the manifold properties can be maintained rigorously, see MacKay (2004). These nearly invariant slow manifolds are named as balanced manifolds or approximate slow manifolds. There are also terminologies like fuzzy manifold and quasi-manifold used to address the vagueness in invariance. In the thesis, we are interest in nondissipative systems, and so the term "slow manifold" always refer a nearly invariant one.

For nondissipative systems, the slow manifold is constructed as an asymptotic series of expansion or iterative procedures (2.1.5). The accuracy of balance can be improved to a desired order of accuracy, $O(\varepsilon^n)$, obtained by the series expansion of the first n terms or by the iteration at nth step. While the improvement is asymptotically possible in the

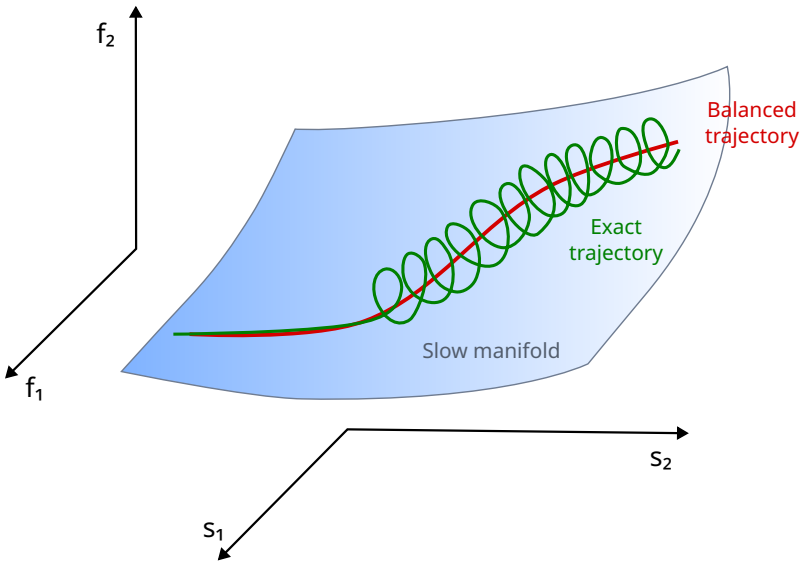

Figure 2.1: Exponentially small fast oscillations are spontaneously excited after some time. To provide an illuminating interpretation, we consider a dynamical system with an extended phase space, which consists of two slow (s_1, s_2) and two fast (f_1, f_2) variables, as being different than the phase space of our general model (2.1.1). The exact trajectory (green path) initialised on the nearly invariant slow manifold remains in close distance to it for long slow time. After some time, the model dynamics excites abrupt oscillations, which are exponentially small, and the exact trajectory drifts away from the balanced trajectory (red path) obtained from the balanced model. Drawn after Vanneste (2013).

limit of $\varepsilon \to 0$ for fixed n, the convergence to a slow manifold is limited by the divergent nature of the series expansion as $n \to \infty$. This analytical divergence corresponds to the nonexistence of an exactly invariant slow manifold.

The best balance approximation results after the truncation of the series expansion at optimal term, $n_{\text{opt}} \sim 1/\varepsilon$, which gives the smallest coefficient in the last term, and the accuracy turns out to be an exponential order, $O(\exp(-c/\varepsilon))$ for some $c > 0$. After the term n_{opt}, the remainder of the series becomes significant, and the balance error increases, see Vanneste (2013) and references therein. As the slow manifold is exponentially accurate, the dynamical systems with optimally truncated initialisations start exciting fast unbalanced motion with exponentially small amplitude, see Figure 2.1. This "spontaneous excitation" or "spontaneous generation" justifies the nonexistence of exactly invariant slow manifold.

The exponentially small deviation of fast dynamics from the slow manifold is achieved rigorously in different type of systems, but not all results provided the exponent 1 of ε, that is $O(\exp(-c/\varepsilon^{\alpha}))$ for different α. For a generic finite-dimensional system as in (2.1.1), the exponential decay of imbalance is obtained with $\alpha = 1/2$ and $\alpha = 1$ in Wirosoetisno (2004) and Vanneste (2008), respectively. When a specific Hamiltonian system is studied, to be introduced in Section 2.3.1, the fast dynamics remain close to the slow manifold with the residual of $O(\exp(-c/\varepsilon))$ over long times of $O(\exp(c/\varepsilon))$ provided that the fast dynamics are vanished at the initial time, see Cotter (2013). In the same system, the residual is found to decay slower with the power $\alpha = 1/3$, where it is not strict but 1 is not reached, in Gottwald et al. (2017). As an infinite-dimensional model, the hydrostatic primitive equations are studied, and the bound with $\alpha = 1/3$ is found, see Temam and Wirosoetisno (2007, 2010, 2011).

2.2 Infinite-dimensional model

In this part, we introduce the rotating shallow water equations as an infinite-dimensional model, which can be reformulated in the form of (2.1.1). Our main numerical work is also built on this model. Besides the description of the model, we separate linear shallow-water flow exactly into its balanced Rossby-wave and imbalanced gravity-wave components by the normal-mode decomposition. The separation can be conducted by the application of the spectral projectors on the flows.

2.2.1 The rotating shallow water equations

The rotating shallow-water equations describe the evolution of fluid columns in shallow fluid layer with mean height H_0. The fluid dynamics is described by free surface displacement $h = h(\boldsymbol{x}, t)$ and velocity field $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x}, t)$ on the 2π-doubly periodic domain in the following equations:

$$\partial_t \boldsymbol{u} + \boldsymbol{u} \cdot \boldsymbol{\nabla} \boldsymbol{u} + f\, \boldsymbol{u}^\perp + g\, \boldsymbol{\nabla} h = 0\,, \tag{2.2.1a}$$

$$\partial_t h + \boldsymbol{\nabla} \cdot (h\boldsymbol{u}) + H_0\, \boldsymbol{\nabla} \cdot \boldsymbol{u} = 0\,, \tag{2.2.1b}$$

where $\boldsymbol{u}^\perp = (-v, u)$, the fluid domain has the Coriolis parameter f and the gravity acceleration g.

The nondimensionalisation of the equations (2.2.1) characterises physical regimes in the fluid domain depending on scaling limits. We introduce a horizontal velocity scale U, horizontal length scale L, fluid depth scale H, and time scale T, and describe each quantity in (2.2.1) as multiplication of its scale with its nondimensional form, which we denote, for simplicity, by the same letter. The equations, then, become

$$\frac{U}{T}\partial_t \boldsymbol{u} + \frac{U^2}{L}\boldsymbol{u} \cdot \boldsymbol{\nabla} \boldsymbol{u} + fU\boldsymbol{u}^\perp + \frac{gH}{L}\boldsymbol{\nabla} h = 0\,, \tag{2.2.2a}$$

$$\frac{H}{T}\partial_t h + \frac{HU}{L}\boldsymbol{\nabla} \cdot (h\boldsymbol{u}) + \frac{H_0 U}{L}\boldsymbol{\nabla} \cdot \boldsymbol{u} = 0\,. \tag{2.2.2b}$$

Assuming an *advective* time scale, $T = L/U$, we obtain three nondimensional parameters: The Rossby number is denoted by

$$\varepsilon = \frac{U}{fL}, \tag{2.2.3}$$

and it can be written as $\varepsilon = 1/(fT)$ in terms of time scales demonstrating the magnitude of the dynamics frequency $1/T$ in the domain relative to the local Coriolis frequency f. The second parameter is called the Burger number, and it is given by

$$\mathrm{Bu} = \frac{L_d^2}{L^2} = \frac{gH_0}{f^2 L^2}\,, \tag{2.2.4}$$

where the Rossby deformation radius $L_d = c/f$ with characteristic wave speed $c = \sqrt{gH_0}$ indicates the length scale that a gravity wave can travel until having a significant change under the effect of the Coriolis force. The third parameter is the nondimensional layer depth

$$h_0 = \frac{H_0}{H}\,. \tag{2.2.5}$$

The nondimensional shallow-water equations (2.2.2), then, become

$$\varepsilon\left(\partial_t \boldsymbol{u} + \boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{u}\right) + \boldsymbol{u}^{\perp} + \frac{1}{h_0}\frac{\mathrm{Bu}}{\varepsilon}\boldsymbol{\nabla}h = 0\,, \tag{2.2.6a}$$

$$\partial_t h + \boldsymbol{\nabla}\cdot(h\boldsymbol{u}) + h_0\boldsymbol{\nabla}\cdot\boldsymbol{u} = 0\,. \tag{2.2.6b}$$

Focusing on the small-Rossby-number regime, $\varepsilon \ll 1$, the Coriolis term $\boldsymbol{u}^{\perp}$ dominates the material advection terms $(\partial_t\boldsymbol{u} + \boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{u})$. In this case, the Coriolis term is balanced by the pressure gradient $\boldsymbol{\nabla}h$ to a leading order in the momentum equation (2.2.6a), when $\mathrm{Bu} = \varepsilon h_0$. This leading-order geostrophic balance is

$$\boldsymbol{u}_{\mathrm{g}} = \boldsymbol{\nabla}^{\perp}h\,. \tag{2.2.7}$$

Two main distinguished scaling limits are considered:

i) The quasi-geostrophic limit arises from Burger number of order one, $\mathrm{Bu} = O(1)$. The nondimensional mean height, therefore, becomes $h_0 = O(\varepsilon^{-1})$, and small amplitude of height variations $h = O(\varepsilon)$ follow.

ii) The semi-geostrophic limit arises from Burger number of order ε, $\mathrm{Bu} = 0(\varepsilon)$. The heights h_0 and h, therefore, take the same order $O(1)$.

We, thus, admit $h_0 = \varepsilon^{-1}$ for the quasi-geostrophic scaling; $h_0 = 1$ for the semi-geostrophic scaling. In this thesis, we work on both scaling regimes.

2.2.2 Normal-mode decomposition

For the linear rotating shallow-water equations,

$$\varepsilon\partial_t\boldsymbol{u} + \boldsymbol{u}^{\perp} + \boldsymbol{\nabla}h = 0\,, \tag{2.2.8a}$$

$$\partial_t h + h_0\boldsymbol{\nabla}\cdot\boldsymbol{u} = 0\,, \tag{2.2.8b}$$

described on the periodic domain, the $\boldsymbol{u}$-h variables are spectrally represented in Fourier space: When $\boldsymbol{k} = (k,l)$ is the wave-number vector with k and l, respectively, in x and y directions, for $\boldsymbol{u}$ we write

$$\boldsymbol{u}(\boldsymbol{x},t) = \sum_{\boldsymbol{k}\in\mathbb{Z}} \boldsymbol{u}_{\boldsymbol{k}}(t)\,\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}\,. \tag{2.2.9}$$

After multiplication with $e^{-i\boldsymbol{k}\cdot\boldsymbol{x}}$ and integration over domain, the linear equations (2.2.8) turn into the system of $\boldsymbol{z}_{\boldsymbol{k}} = (\boldsymbol{u}_{\boldsymbol{k}}, h_{\boldsymbol{k}})$,

$$\partial_t\boldsymbol{z}_{\boldsymbol{k}} = \mathrm{i}A_{\boldsymbol{k}}\boldsymbol{z}_{\boldsymbol{k}} \quad \text{with} \quad A_{\boldsymbol{k}} = \begin{pmatrix} 0 & -\mathrm{i}/\varepsilon & -k/\varepsilon \\ \mathrm{i}/\varepsilon & 0 & -l/\varepsilon \\ -h_0\,k & -h_0\,l & 0 \end{pmatrix}. \tag{2.2.10}$$

For each $\boldsymbol{k}$, the matrix $A_{\boldsymbol{k}}$ holds eigenvalues, so-called dispersion relations, $\lambda^{\mathrm{RW}}_{\boldsymbol{k}} = 0$ and $\lambda^{\mathrm{GW}}_{\boldsymbol{k}} = \pm(\varepsilon h_0|\boldsymbol{k}|+1)/\varepsilon^2$, and their corresponding eigenvectors $\boldsymbol{v}_{\boldsymbol{k},0}$ and $\boldsymbol{v}_{\boldsymbol{k},1},\ \boldsymbol{v}_{\boldsymbol{k},2}$ take the form

$$\boldsymbol{v}_{\boldsymbol{k},0} = \begin{pmatrix} -il \\ ik \\ 1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{v}_{\boldsymbol{k},1},\ \boldsymbol{v}_{\boldsymbol{k},2} = \begin{pmatrix} 0 \\ i/\varepsilon \\ -h_0k \end{pmatrix},\ \begin{pmatrix} -i/\varepsilon \\ 0 \\ -h_0l \end{pmatrix}. \tag{2.2.11}$$

The Rossby-wave mode $\boldsymbol{v}_{\boldsymbol{k},0}$ and the gravity-wave modes ($\boldsymbol{v}_{\boldsymbol{k},1},\ \boldsymbol{v}_{\boldsymbol{k},2}$) are not orthogonal to each other.

The Rossby-wave component is along $v_{\boldsymbol{k},0}$, which defines the kernel of $A_{\boldsymbol{k}}$ (denoted $\mathrm{Ker}\, A_{\boldsymbol{k}}$), and the gravity-wave component is the projection onto the span of $\boldsymbol{v}_{\boldsymbol{k},1}$ and $\boldsymbol{v}_{\boldsymbol{k},2}$, which define the range of $A_{\boldsymbol{k}}$ (denoted $\mathrm{Range}\, A_{\boldsymbol{k}}$). These components are, then, found by the Rossby-wave projector $\mathbb{P}^{\mathrm{RW}}$ and its complement, gravity-wave projector, $\mathbb{P}^{\mathrm{GW}} = I - \mathbb{P}^{\mathrm{RW}}$, providing the following decomposition

$$z_{\boldsymbol{k}} = (\mathbb{P}^{\mathrm{RW}} + \mathbb{P}^{\mathrm{GW}})\, z_{\boldsymbol{k}}. \tag{2.2.12}$$

To test in numerical experiments, we construct the projector $\mathbb{P}^{\mathrm{RW}}$ in two different ways: First, we follow the non-orthogonal nature of subspaces, which are then oblique, and for each wave number $\boldsymbol{k}$, we construct the *oblique projector* $\mathbb{P}^{\mathrm{RW,obliq}}_{\boldsymbol{k}}$ onto $\mathrm{Ker}\, A$ along $\mathrm{Range}\, A$. This Rossby-wave projector intrinsically retains existing modes on $\mathrm{Range}\, A$, while vanishes modes on $\mathrm{Ker}\, A$, i.e.,

$$\mathbb{P}^{\mathrm{RW,obliq}}_{\boldsymbol{k}} \begin{pmatrix} \boldsymbol{v}_{\boldsymbol{k},0} & \boldsymbol{v}_{\boldsymbol{k},1} & \boldsymbol{v}_{\boldsymbol{k},2} \end{pmatrix} = \begin{pmatrix} \boldsymbol{v}_{\boldsymbol{k},0} & \boldsymbol{0} \end{pmatrix}, \tag{2.2.13}$$

where the mode matrix is invertible, and as a result, we obtain the Rossby-wave projection matrix,

$$\mathbb{P}^{\mathrm{RW,obliq}}_{\boldsymbol{k}} = \frac{1}{h_0\,|\boldsymbol{k}|^2 + \varepsilon^{-1}} \begin{pmatrix} h_0\, l^2 & -h_0\, kl & -\mathrm{i}l/\varepsilon \\ -h_0\, kl & h_0\, k^2 & \mathrm{i}k/\varepsilon \\ \mathrm{i}h_0\, l & -\mathrm{i}h_0\, k & 1/\varepsilon \end{pmatrix}. \tag{2.2.14}$$

Second, we assume orthogonality between the subspaces, and the *orthogonal projector* is constructed directly onto $\mathrm{Ker}\, A_{\boldsymbol{k}}$, by the orthogonal projection formula,

$$\mathbb{P}^{\mathrm{RW,orth}}_{\boldsymbol{k}} = \frac{\boldsymbol{v}_{\boldsymbol{k},0}\, \boldsymbol{v}^*_{\boldsymbol{k},0}}{\|\boldsymbol{v}_{\boldsymbol{k},0}\|^2} = \frac{1}{|\boldsymbol{k}|^2 + 1} \begin{pmatrix} l^2 & -kl & -\mathrm{i}l \\ -kl & k^2 & \mathrm{i}k \\ \mathrm{i}l & -\mathrm{i}k & 1 \end{pmatrix}, \tag{2.2.15}$$

where $\boldsymbol{v}^*_{\boldsymbol{k},0}$ denotes the Hermitian conjugate. The complementary projectors $\mathbb{P}^{\mathrm{GW}}$, for the two cases, immediately follow. These projectors are used in our application of the diagnostic flow-wave separation method, and we leave the analysis of these projectors in Section 4.3.

2.3 Finite-dimensional models

The concept of balance and spontaneous wave generation are studied through simple finite-dimensional models which structurally resemble geophysical flow models. These simple models are advantageous for numerical computation and theoretical study, which are important to understand the fundamental features of the original ones. In this section, we cover two Hamiltonian systems: a system derived variationally in the semi-geostrophic scaling limit and a spring-mass pendulum derived by truncating the shallow-water model. The former system is important to build the background of our further studies, and the spring-mass pendulum is visited as an analogue of balance and spontaneous wave generation.

2.3.1 Approximating shallow-water flows

The first toy model describes the dynamics for the position vector $\boldsymbol{r} : [0, T] \to \mathbb{R}^{2d}$ and the corresponding momenta $\boldsymbol{p}$ in a smooth potential $V \in C^{n+1}$. Denoting time derivatives by overdot, e.g, $d\boldsymbol{r}/dt = \dot{\boldsymbol{r}}$, we write the system in the form of

$$\begin{aligned} \dot{\boldsymbol{r}} &= \boldsymbol{p}\,, \\ \varepsilon\dot{\boldsymbol{p}} &= J\boldsymbol{p} - \boldsymbol{\nabla} V(\boldsymbol{r})\,, \end{aligned} \tag{2.3.1}$$

where ε is again a small time-scale separation parameter, and J is the canonical symplectic matrix in $2d$ dimensions,

$$J = \begin{pmatrix} 0 & -I_d \\ I_d & 0_d \end{pmatrix}.$$

This model is analysed in several different studies (Cotter and Reich, 2006; Oliver, 2006; Cotter, 2013; Gottwald and Oliver, 2014) and we recall the transformation of this model in the form of the abstract formulation (2.1.1) in Section 2.1.

The model (2.3.1) is interpreted as the evolution of a single fluid column, when $d = 1$. The model reduction is explained by the rotating shallow-water model (2.2.6) in the semi-geostrophic scaling limit. For a single fluid column initiated at Lagrangian label coordinate $\boldsymbol{a}$, we define a fluid map $\boldsymbol{\eta}_t(\boldsymbol{a})$ in time t through where the fluid particle advects. The particle takes the Eulerian position $\boldsymbol{r}(t) = \boldsymbol{\eta}_t(\boldsymbol{a})$, and the Lagrangian velocity is given by $\partial_t \boldsymbol{\eta}_t(\boldsymbol{a}) = \boldsymbol{u}(\boldsymbol{\eta}_t(\boldsymbol{a}), t)$. The momentum conservation equation (2.2.6a), then, gives

$$\varepsilon\ddot{\boldsymbol{r}} - J\dot{\boldsymbol{r}} + \boldsymbol{\nabla} h(\boldsymbol{r}) = 0\,, \tag{2.3.2}$$

where $J\dot{\boldsymbol{r}}$ represents the Coriolis term $\boldsymbol{u}^{\perp}$, the limit $\varepsilon \to 0$ indicates rapid rotation, and h takes the role of the potential for a fixed layer depth, $V(\boldsymbol{r}) := h(\boldsymbol{r})$. After introducing the momenta $\boldsymbol{p} = \dot{\boldsymbol{r}}$, the equation (2.3.2) is written as a system of first-order differential equations (2.3.1).

An alternative understanding of the model underlies the evolution of a single charged particle $\boldsymbol{r}$ with a potential V on a plane in the presence of an external magnetic field, that is normal to the plane (Gottwald and Oliver, 2014; Gottwald et al., 2007). In this case, the term $J\boldsymbol{r}$ in the equation (2.3.2) is the Lorentz force and the limit $\varepsilon \to 0$ corresponds to the zero mass limit of the particle while its charge is unchanged.

Two different time scaled motions are identified in the toy model. When the potential is neglected, the linear system

$$\varepsilon\ddot{\boldsymbol{r}} = J\dot{\boldsymbol{r}} \tag{2.3.3}$$

describes fast harmonic oscillations, and when $\varepsilon = 0$, the slow geostrophically balanced motion is represented by

$$\dot{\boldsymbol{r}} = -J\boldsymbol{\nabla} V(\boldsymbol{r})\,. \tag{2.3.4}$$

The slow dynamics evolves on the time scale $O(1)$ and the fast dynamics evolves on the time scale $O(\varepsilon^{-1})$. As introduced earlier, the accuracy of balanced motion can be increased order by order. The asymptotic series with regard to the parameter ε are practically convenient, but their explicit construction are a tedious work for most of the dynamical systems due to complicated recursive terms. As a simple example, we restate the construction for the model (2.3.1) in the following theorem. The theorem shows only the algebraic-order accuracy, and the exponential accuracy is achieved, when the remainder term is estimated carefully. The core of our work, however, builds the slow manifold implicitly via numerical methods.

Theorem 2.3.1 (Gottwald and Oliver (2014)). *For $n \in \mathbb{N}$, suppose $V \in C^{n+2}$ and set*

$$G_n(\boldsymbol{r}) = \sum_{i=0}^{n} \varepsilon^i g_i(\boldsymbol{r}) \tag{2.3.5}$$

with coefficient functions g_i recursively defined via

$$\begin{aligned} g_0(\boldsymbol{r}) &= -J\boldsymbol{\nabla} V(\boldsymbol{r})\,, \\ g_k(\boldsymbol{r}) &= -J \sum_{i+j=k-1} Dg_i(\boldsymbol{r}) g_j(\boldsymbol{r})\,. \end{aligned} \tag{2.3.6}$$

For fixed initial positions $\boldsymbol{r}_0 \in \mathbb{R}^{2d}$ and $a > 0$, let $\boldsymbol{r}(t)$ denote a solution to

$$\dot{\boldsymbol{r}} = G_n(\boldsymbol{r}) \tag{2.3.7}$$

with $\boldsymbol{r}(0) = \boldsymbol{r}_0$. Let $\boldsymbol{r}_\varepsilon(t)$ solve the full parent dynamics (2.3.1) consistently initialised via $\boldsymbol{r}_\varepsilon(0) = \boldsymbol{r}_0$ and $\boldsymbol{p}_\varepsilon(0) = G_n(\boldsymbol{r}_0)$. Then, there exist $\varepsilon_0 > 0$ and $c = c(\boldsymbol{r}_0, a, V)$ such that

$$\sup_{t\in[0,T]} \|\boldsymbol{r}_\varepsilon(t) - \boldsymbol{r}(t)\| \leq c\ \varepsilon^{n+1} \tag{2.3.8}$$

for all $0 < \varepsilon < \varepsilon_0$.

2.3.2 The spring-mass pendulum

The use of finite-dimensional models is initiated when Lorenz (1980) suggested a model with five degrees of freedom. After the introduction of forced-dissipative terms (Lorenz and Krishnamurthy, 1987), it is also called the Lorenz-Krishnamurthy model. The Lorenz's model is reduced to a two-component Hamiltonian system that corresponds closely to a simple mechanical system, the spring-mass pendulum, or the swinging pendulum, by Camassa (1995) and Bokhove and Shepherd (1996).

The spring-mass pendulum is one of the well-known example to explain the essence of balanced and unbalanced motions. The pendulum is composed of a stiff elastic spring, whose mass is negligible, being attached to a fixed central point at one end and carrying a mass at the other end, see in Figure 2.2. When the pendulum departs at a certain angle θ from the vertical equilibrium position, it starts swinging with slow frequency in the angular direction, and it accelerates under the effect of the gravity. Over the time evolution, the mechanical system also generates small amount of spring compression with fast frequency, which changes the length of the spring ℓ, consequently the restored energy. The swinging mode corresponds to slow balanced flow, and the compressional mode corresponds to fast gravity waves. The small time scale-separation parameter ε is, then, given by the ratio of the pendulum frequency to the spring frequency. Adding initial spring stretch is also possible at the start of the motion, but then the amplitude of fast vibrations might not be small any more.

The stiffness of the spring determines the excitation of fast vibrations. The compressible spring extends close to the equilibrium position, where the mass moves faster; it contracts while traversing away from the inertial position, where the mass moves slower. If it is sufficiently stiff, then the compressional mode can be neglected, and the swinging mode gives the same analogy of geostrophic balance or the leading-order balance. If it is stiff to include minimal compression, the analogy of the quasi-geostrophic theory is obtained. More accurate approximations follows in hierarchy by changing the stiffness of the spring.

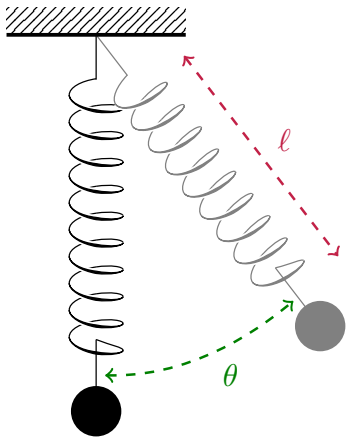

Figure 2.2: The spring-mass pendulum is a paradigm of balance. This system consists of a spring attached to a fixed point with a mass at the other end. The pendulum oscillations with angle θ are considered analogous to slow balanced dynamics, and the spring compression, so that the spring length ℓ, corresponds to fast gravity waves for geophysical fluids. The ratio of pendulum frequency to spring frequency gives the time scale-separation parameter.

2.4 Diagnostic flow-wave separation

The separation of balanced flow and gravity waves is performed by many different diagnostic approaches, but we study the theory of optimal balance. The foremost study of the method was a numerical implementation in the context of semi-Lagrangian schemes for rapidly-rotating fluid flows provided under the name "optimal PV balance" by Viúdez and Dritschel (2004), see Section 2.4.3.

The theoretical understanding of optimal balance is derived by Cotter (2013) for a system of single Lagrangian particle (2.3.1) in fast-time scale. He described optimal balance as an approach to constrain a phase space on the slow manifold, which is used to obtain the assimilated initial data from noisy observations. In this approach, the model is integrated symplectically, where the Hamiltonian characteristics are preserved, over very long time scaled with ε. The excitation of fast dynamics is, then, restricted to be exponentially small depending on his earlier theoretical work (Cotter and Reich, 2006), and by this, he put forward the optimal balance procedure as adiabatic invariance of slow manifolds.

The exponential residual between the fast dynamics and the slow manifold was achieved under the assumption of an analytic ramp function and an analytic potential (Cotter, 2013). The derivatives of the ramp function converged asymptotically to zero as the ramp time diverges to $\pm\infty$. For finitely-long ramp time, however, the issue of undefined derivatives at temporal boundaries emerged, which was not discovered due to the special choice of the ramp function. Gottwald et al. (2017) observed the need to define derivatives and proved theoretically the exponential estimate with smaller power of ε, when the derivatives of ramp functions vanish at the boundaries, see Section 2.4.2.

The remaining part of the section first describes general characteristics of optimal balance on the abstract model (2.1.1) with approaches to solve an optimal balance boundary value problem (BVP). Next, we recall the algebraic and exponential estimate of the balance error for (2.3.1) in Gottwald et al. (2017). As last, optimal PV balance is presented in details, which is greatly beneficial for our own implementation.

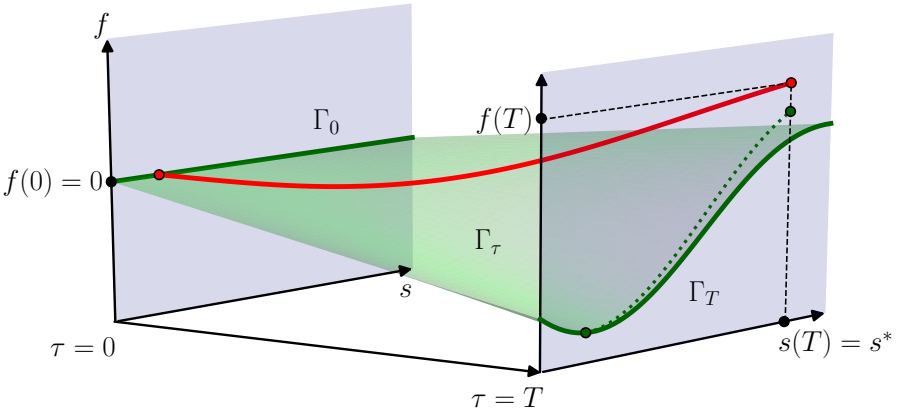

Figure 2.3: The schematic drawing presents the geometry of optimal balance in two-dimensional phase space. In the course of the artificial-time τ evolution, the slow manifold Γ_0 determined by $f(0) = 0$ is deformed, and a surface of approximate slow manifolds Γ_τ is generated by the homotopy (green-shaded area). A particle initially located on Γ_0 moves through a path (red line), which remains in close neighbourhood of Γ_τ for long-time τ horizon. When the deformation is inactive, a path on Γ_T deviates away due to the excitation of exponentially small oscillations (green-dotted line). Adapted from Gottwald et al. (2017). Adapted with permission.

2.4.1 Theory of optimal balance

The principal idea of optimal balance is to adiabatically deform the dynamical equations resulting in a smooth homotopy between the linear and the fully nonlinear equations, the simplified geometry of optimal balance is displayed in Figure 2.3. This deformation is performed over an artificial time τ. It gradually turns on the nonlinear interactions starting from the linear system on a linear slow manifold Γ_0 at $\tau = 0$, computed by $\boldsymbol{f}(0) = 0$. The system ramps up to the fully nonlinear form on an approximate slow manifold Γ_T at $\tau = T$. On Γ_T, it is desired to meet with the specified fixed coordinate $s(T) = s^*$. Over the time range $\tau \in [0, T]$, the nonlinear interactions change the slow manifold in a controlled way, and hence, time-dependent approximate slow manifolds Γ_τ are created. The manifolds are shown in the green-shaded area in Figure 2.3. As the deformation over long τ is slow, a path of a particle initiated on Γ_0 stay close to Γ_τ, which is drawn by the red line. In other words, the deviation of the path from Γ_τ between the temporal-end points goes to zero under the limit $\tau \to \infty$, and this asymptotic behaviour is called adiabatic invariance. When the deformation parameter is frozen, Γ_T is approximately invariant. A path on Γ_T, therefore, deviates away with exponential residual in terms of the time scale separation parameter ε, this deviation is illustrated by the green-dotted line.

Two different sources of imbalances are present in the optimal balance procedure. The first source is emerged by the excitation of the exponentially small oscillations without an active deformation through the physical-time dynamics of the nonlinear system on. The second one arises from the slow deformation of Γ_0 over artificial time τ. For the system (2.3.1), the error estimate of the first source has been stated in Theorem 2.3.1, and the error estimates the second source are in the following section (Gottwald et al., 2017).

The application of optimal balance performs an explicit slow deformation to the nonlinear terms by a smooth monotonic ramp function $\rho : [0,1] \to [0,1]$ with $\rho(0) = 0$ and $\rho(1) = 1$, which returns an optimal balance BVP in time. For the abstract system (2.1.1), the BVP is formed as

$$\begin{aligned} \partial_\tau s &= \rho(\tau/T)\mathcal{N}_s(s, \boldsymbol{f}) \,, \\ \partial_\tau \boldsymbol{f} &= \frac{1}{\varepsilon}\mathcal{L}\boldsymbol{f} + \rho(\tau/T)\mathcal{N}_{\boldsymbol{f}}(s, \boldsymbol{f}) \,, \end{aligned} \tag{2.4.1}$$

and the boundary conditions are imposed at the temporal-end points,

$$\boldsymbol{f}(0) = 0 \quad \text{and} \quad s(T) = s^* \,, \tag{2.4.2}$$

which are called respectively the linear-end and the nonlinear-end boundary condition as regards to the type of the dynamical system, where the conditions are applied upon. The linear-end condition alone is enough to provide a balanced state without the role of the nonlinear-end condition, but the latter condition ensures balance to build for a fixed quantity s^*. Besides this explicit deformation (Gottwald et al., 2017), the deformation can be also carried out implicitly without direct effect to the nonlinear terms (Viúdez and Dritschel, 2004).

The precise definition of balance is determined by the choice of projector which maps on an approximate slow manifold and the choice of base point s coordinate. The optimal balance method is our projector and the coordinate is chosen among fields which predominantly structure the geostrophic component of the flow. The use of height field h is a common choice in balance relations, especially in the variationally derived balanced models initiated by Salmon (1983). The balance relation for $\boldsymbol{u}$ is formulated as

$$\boldsymbol{u} = \boldsymbol{u}_{\mathrm{g}} + \boldsymbol{u}_{\mathrm{bag}} = \boldsymbol{u}_{\mathrm{g}} + O(\varepsilon) \tag{2.4.3}$$

where each term is kinematically dependent on h, see the geostrophic balance $\boldsymbol{u}_{\mathrm{g}}$ in (2.2.7). Alternatively, the use of the PV q is proposed by Viúdez and Dritschel (2004). The complementary fields can be, then, written as velocity divergence δ and ageostrophic vorticity γ, which are precisely defined for the rotating shallow-water fluid in Chapter 4. The q-δ-γ variables capture the leading-order separation, and a general flow becomes

$$q + (\delta, \gamma)_{\mathrm{bag}} + (\delta, \gamma)_{\mathrm{unb}} \,,$$

where the first two terms denote the balanced flow. Any balance relations in the form of (2.4.3) come with another kinematic relation between h and q, and so both points can be inverted to another one (Dritschel et al., 2017; Calik et al., 2013). We, therefore, employ h and q as base points with no theoretical obstacle, and our work proves their numerical applicability with some conditions in Chapter 6.

To solve the optimal balance BVP (2.4.1), the simple shooting method is implemented on a lower-dimensional system (2.3.1) by Gottwald et al. (2017). It integrates the system of equations as an initial value problem with the initial value $(s(0), \boldsymbol{f}(0))$ to match with the given boundary value $s(T) = s^*$. In higher dimensions, however, the shooting method needs to compute (approximate) Jacobian implicitly, which comes with expensive computational cost. Any other advanced boundary solvers can possibly fail in a similar way.

The backward-forward nudging scheme is suggested by Viúdez and Dritschel (2004) in the context of geophysical fluid dynamics, which is discovered as a stable and robust

solver. The scheme consists of repeated backward and forward integrations while nudging the desired boundary condition after each one-sided integration: At $\tau = 0$, $\boldsymbol{f}(0)$ is removed by the linear-end boundary, and the other component $s(0)$ is preserved. At $\tau = T$, $s(T)$ is restored to s^* by the nonlinear-end boundary, and $\boldsymbol{f}(T)$ is maintained. For the nudging scheme, we refer the work of Auroux and Nodet (2012), though the practical usage is different. In this work, the observational data to be converged by the state variable is added as a nudging term in the equations of a dynamical system, and the back-and-forth integrations are successively proceeded with the initial conditions provided by the final state of the previous integration. The convergence of the backward-forward nudging scheme is, yet, not understood theoretically, although its rapid convergence is reported numerically. For a lower-dimensional setting (2.3.1), we study its convergence in Chapter 3.

The balanced state $(s^*, \boldsymbol{f}(T))$ is obtained as an approximate solution to the BVP (2.4.1) with its boundaries (2.4.2). At each step in the τ-integration, optimal balance constructs a slow vector field on Γ_τ by projecting $s(\tau)$ onto the full phase space, i.e.,

$$Q_n(s(\tau), \tau) = \sum_{i=1}^{n} \varepsilon^i q_i(s(\tau), \tau)$$

with functions q_i to be determined. As $\tau \to T$, the solution of the BVP on Γ_τ becomes closer to the one obtained by the optimal truncation of the series describing Γ_T, in other words,

$$\boldsymbol{f}(\tau) = Q_n(s(\tau), \tau) \to G_n(s^*) \quad \text{as} \quad \tau \to T\,.$$

The error of optimal balance is, then, computed by the following residual

$$\|\boldsymbol{f}(T) - G_n(s^*)\|\,,$$

which is exponentially small. The rigorous establishment of Q_n and G_n is mostly computationally inconvenient, so that optimal balance becomes practically important to construct a balanced state. Gottwald et al. (2017) defined this method as "the best practically available characterisation of slow manifold" and named it as *optimal balance*.

The approximate slow manifold Γ_T is restricted to preserve its approximately invariant nature over long time disregarding the choice of base point (MacKay, 2004). This requirement is naturally satisfied, at least for small ε values, by the selection of the boundary conditions: The flow becomes free of fast dynamics Γ_0 at the linear end, and it is projected on Γ_T by a base point evolving mainly the geostrophic component of the flow at the nonlinear end. The slow manifold Γ_T is, thus, well-defined by the optimal balance BVP.

2.4.2 Optimal balance

The finite-dimensional Hamiltonian model (2.3.1) is considered to theoretically understand optimal balance, and the optimal balance BVP becomes

$$\partial_\tau \boldsymbol{r} = \boldsymbol{p}\,, \tag{2.4.4a}$$

$$\varepsilon \partial_\tau \boldsymbol{p} = J\boldsymbol{p} - \rho(\tau/T)\boldsymbol{\nabla} V(\boldsymbol{r})\,, \tag{2.4.4b}$$

with boundary conditions

$$\boldsymbol{p}(0) = 0 \quad \text{and} \quad \boldsymbol{r}(T) = \boldsymbol{r}^*\,. \tag{2.4.5}$$

The slow-fast dynamics are entangled in this system: $\boldsymbol{r}$ represents the slow dynamics, while $\boldsymbol{p}$ indicates both type of dynamics. At $\tau = 0$, however, slow and fast dynamics

are explicitly separated in the linear system. At $\tau = T$, the leading-order slow balanced motion in the nonlinear system is given by (2.3.4).

The asymptotic behaviour of optimal balance is, first, associated with the smoothness of the potential and ramp functions by Cotter (2013). Later on, Gottwald et al. (2017) indicate a crucial feature of ρ to obtain the desired order of accuracy. They achieved the algebraic accuracy $O(\varepsilon^{k+1})$ provided that the first finite n derivatives of ρ vanish at the temporal-end points of the problem, which can be seen in the following theorem. Without the vanishing derivatives, additional terms at $O(\varepsilon^k)$ stay present and limit the accuracy of optimal balance.

Theorem 2.4.1 (Gottwald et al. (2017))**.** *For $n \in \mathbb{N}$, suppose $\rho \in C^{n+1}$ with $\rho(0) = 0$ and $\rho(1) = 1$ satisfying the algebraic order condition*

$$\rho^{(i)}(0) = \rho^{(i)}(1) = 0 \tag{2.4.6}$$

for $i = 1, \ldots, n$. Suppose further that $V \in C^{n+2}$. Fix $a > 0$ and consider a sequence of ramp times $T = a$ and a sequence of solutions $(\boldsymbol{r}, \boldsymbol{p})$, implicitly parametrised by ε, to the boundary value problem (2.4.4)*. Then there exists a constant $c = c(\rho, a, n, V)$ such that*

$$\|\boldsymbol{p}(T) - G_n(\boldsymbol{r}^*)\| \leq c\ \varepsilon^{n+1}\,.$$

To obtain beyond-all-order accuracy, the ramp function ρ must have all derivatives zero at the end points as a consequence of Theorem 2.4.1. The ramp function ρ, nevertheless, cannot satisfy the analyticity and vanishing derivatives at the same time. In this case, ρ can be chosen from a special Gevrey class 2, $G^2(0,1)$, which is a class of analytic functions with bounded derivatives, see Gottwald et al. (2017) for the definition in exact terms. This class of ρ functions with analytic potential V lead an exponential separation of fast dynamics in the cube root of ε, $O(\exp(-c/\varepsilon^{1/3}))$, which is the optimal convergence.

Theorem 2.4.2 (Gottwald et al. (2017))**.** *Suppose $\rho \in G^2(0,1)$ with $\rho(0) = 0$ and $\rho(1) = 1$ satisfying the exponential order condition*

$$\rho^{(i)}(0) = \rho^{(i)}(1) = 0 \tag{2.4.7}$$

for all $i \in \mathbb{N}^*$. Fix $a > 0$, and consider a sequence of ramp times $T = a$ and a sequence of solutions $(\boldsymbol{r}, \boldsymbol{p})$, implicitly parametrised by $\varepsilon \leq 1$, to the boundary value problem (2.4.4)*. Now suppose there exists a compact subset of phase space $\mathcal{K} \subset \mathbb{R}^{2d}$ containing this sequence of solution trajectories and that there exist $R > 0$ and $z_0 \in \mathbb{R}^{2d}$ with $\mathcal{K} \subset B_{R/2}(z_o)$ such that V is analytic on $B_R(z_0)$. Then there exist $n = n(\rho, a, V, \varepsilon) \in \mathbb{N}$ and positive constants $c = c(\rho, a, V)$ and $d = s(\rho, a, V)$ such that*

$$\|\boldsymbol{p}(T) - G_n(\boldsymbol{r}^*)\| \leq d\ \exp\left(-\frac{c}{\varepsilon^{1/3}}\right)\,.$$

These theoretical results are significant to decide the order of optimal balance error in higher dimensions, as the system (2.3.1) is an approximation of the higher-dimensional rotating shallow-water model in the semi-geostrophic regime, see Section 2.3.1. We, thus, expect to partially meet with these results in our numerical implementation.

2.4.3 Optimal potential vorticity balance

Optimal PV balance is supported by the idea of PV being materially conserved and determining the bulk of the slow dynamics even in inertial-gravity-waves-permitting models, see Viúdez and Dritschel (2004). The method, thus, requires the geophysical flow models in the slow geostrophic and fast ageostrophic quantities. The required equations of motion obey the abstract form (2.1.1) for $s = q$ and $f = (\delta, \gamma)$. Due to the material conservation of q, the slow equation materially advects q, and the fast equation is maintained as the Eulerian evolution of ageostrophic fields (δ, γ), i.e.,

$$\partial_t q + \boldsymbol{u} \cdot \boldsymbol{\nabla} q = 0 \,, \tag{2.4.8a}$$

$$\partial_t \begin{pmatrix} \delta \\ \gamma \end{pmatrix} = \frac{1}{\varepsilon} \mathcal{L} \begin{pmatrix} \delta \\ \gamma \end{pmatrix} + \mathcal{N}_{\delta,\gamma}(q, \delta, \gamma) \,, \tag{2.4.8b}$$

where $\mathcal{L}$ and $\mathcal{N}_{\delta,\gamma}$ again indicate a skew linear and a nonlinear operators. As the velocity $\boldsymbol{u}$ is included in the modified formulation, the equations of motion are closed by a PV-inversion equation to obtain $\boldsymbol{u}$ from (q, δ, γ).

The intrinsic nature of their method rests in using Lagrangian advection to approximate the slow-q equation (2.4.8a) and to make this approximation dependent on time. The PV q as a Lagrangian quantity is advected through a fluid map $\boldsymbol{\eta}_t(\boldsymbol{a})$ in physical time t, the map is defined in Section 2.3.1. Given the initial PV field $q_0(\boldsymbol{a})$ at the coordinate a, the material conservation of q (2.4.8a) is written as

$$q(\boldsymbol{\eta}_t(\boldsymbol{a}), t) = q_0(\boldsymbol{a})$$

through the advection. The PV q has its reference field $q_r(\boldsymbol{r})$ satisfiying $\mathcal{N}_{\delta,\gamma}(q_r, \delta, \gamma) = 0$, and (q_r, δ, γ) gives the stationary solution of the system (2.4.8).

As next step, the Lagrangian approximation is evolved over artificial time $\tau \in [0, T]$ for a frozen physical time t. For the corresponding flow map $\boldsymbol{\eta}_\tau$ of the τ-evolution, the PV is ramped starting from the reference field q_r up to a given PV q^*, and the PV conservation yields

$$q(\boldsymbol{\eta}_\tau(\boldsymbol{r}), \tau) = q_r(\boldsymbol{\eta}_\tau(\boldsymbol{r})) + \rho(\tau/T)(q^*(\boldsymbol{r}) - q_r(\boldsymbol{r})) \,, \tag{2.4.9}$$

where the ramp function ρ is a cosine function,

$$\rho(\theta) = (1 - \cos(\pi\theta))/2 \,, \tag{2.4.10}$$

with $\rho(0) = 0$ and $\rho(T) = 1$. In the way q_r is defined, this ramping procedure implicitly provides smooth transition of the nonlinear system into the linear system. The flow map $\boldsymbol{\eta}_\tau$ is restricted to start from the position $\boldsymbol{r}$ of the actual fluid particles. At $\tau = T$, therefore, the particles return back their actual position, i.e., $\boldsymbol{\eta}_T(\boldsymbol{r}) = \boldsymbol{r}$, and the PV satisfies the desired condition $q(\boldsymbol{r}, T) = q^*(\boldsymbol{r})$.

The equations of motion, now, consist of the ramped PV equation (2.4.9) and the evolution of the δ-γ fields (2.4.8b) in τ. The conditions for q are automatically applied by the derivation of the former equation, and the ageostrophic δ-γ fields need the initial condition,

$$\delta(\boldsymbol{r}, 0) = \gamma(\boldsymbol{r}, 0) = 0 \,,$$

to evolve the later equation. The optimal PV method is tested for two systems: a three-dimensional baroclinic model on f-plane and a two-dimensional rotating shallow-water

model on the sphere. As a result, they reported rapid convergence to a balanced state $(q^*(\boldsymbol{r}), \delta(\boldsymbol{r}, T), \gamma(\boldsymbol{r}, T))$, only a few iterations in the backward-forward nudging scheme. The balanced state depends weakly on the ramp function ρ and the ramp-time length T. They implemented this model in a special contour advection code governed by the geostrophic-ageostrophic variables, which is explained below in details, but most numerical geophysical-fluid codes use a different set of variables.

The contour-advective semi-Lagrangian algorithm

For their purely numerical model, they modified their special contour-advective semi-Lagrangian (CASL) algorithm for the needs of the optimal PV balance (Dritschel and Ambaum, 1997). In this contour-based algorithm, the materially conserved PV q is described as a piecewise-uniform function by level sets, which are distinguished by contours, and these contours are advected in an entirely Lagrangian way. This special treatment to PV q makes their work innovative. On the other hand, fields which are not materially conserved are discretised on a coarse fixed grid and advected by a pseudo-spectral method in a standard Eulerian way.

To create these q-contours, the computational domain is divided into level sets sharing uniform q, and the difference between the level sets is given by the PV jump. The q-contours are, then, placed a set of nodes to separate these levels (Dritschel et al., 1999). Through the advection of the q-contours, the characteristics of q is resolved on very small scales finer than the grid scales, and steeper gradients of q are preserved by construction, which cannot be provided by the coarse grid resolution. As a result, when the computational grid is refined, the numerical solution is improved.

The advection of q-contours is carried out by the trajectory integration, which requires the velocity field $\boldsymbol{u}$. For the computation of $\boldsymbol{u}$, the PV q is interpolated on the grid by projecting in a finer grid and then averaging down to the actual computational-grid scale until obtaining a smooth q. Given h field, ageostrophic vorticity ζ is computed straight forward using the definition of q. The field $\boldsymbol{u}$ is, then, computed from the stream function ψ and velocity potential ϕ, these PV-inversion formulas will be also used in our work, see Section 4.4.2. As next, all nodes on the q-contours are evolved forward in time, which may follow by possible redistribution of nodes on each contour. Through the evolution, long narrow PV filaments emerge. Once they occur, contour surgery is applied to remove these filamentary structures and reconnect contours from the cut edges (Dritschel, 1988).

The implementation of the optimal PV balance method can be, now, explained in terms of the CASL algorithm. The ramp function is applied to the PV anomalies $q^* - q_r$ in (2.4.9). This ramping procedure is simply performed by multiplying the PV jumps within the level sets. In the course of the backward integration, the q-contours get similar structure with the ones of the reference field q_r. At the end, at $\tau = 0$, the q-contours are maintained to initialise the forward integration. After the forward integration, at $\tau = T$, the boundary condition is imposed by replacing the advected q-contours at the final time with the q-contours of the actual flow, and the other fields follow in a standard way. The iterative integrations are assumed to converge a balance state when sufficiently small flow states are produced at the end of each iteration. As a result, a high-quality balance is provided with affordable computational cost by the implicit definition of a balance relation.

Chapter 3

Quasi-convergence of the optimal balance nudging scheme

The theory of optimal balance was numerically studied by solving the associated BVP using backward-forward nudging by Viúdez and Dritschel (2004). There are, however, open questions that have not been answered up to now. We do not know whether the BVP is well-posed and if it is, whether the nudging scheme converges to the solution of the BVP. In this chapter, assumed that the BVP is well-posed, we analyse the convergence of the nudging scheme in the context of the finite-dimensional model (2.3.1).

3.1 Construction of slow manifold

The finite-dimensional model (2.3.1) and the application of optimal balance on this model are introduced in Section 2.3.1 and 2.4.2 respectively, yet we recall the essential parts with some details from Gottwald et al. (2017). The model evolves the position vector $\boldsymbol{r} : [0,T] \to \mathbb{R}^{2d}$ and the corresponding momenta $\boldsymbol{p}$ with the small time scale-separation parameter ε and the potential $V \in C^{n+1}$. Optimal balance is applied to this model,

$$\begin{aligned} \partial_t \boldsymbol{r} &= \boldsymbol{p}\,, \\ \varepsilon \partial_t \boldsymbol{p} &= J\boldsymbol{p} - \rho(t/T)\boldsymbol{\nabla} V(\boldsymbol{r}), \end{aligned} \tag{3.1.1}$$

with boundary conditions

$$\boldsymbol{p}(0) = 0 \quad \text{and} \quad \boldsymbol{r}(T) = \boldsymbol{r}^*\,. \tag{3.1.2}$$

using a smooth monotonic function $\rho : [0,1] \to [0,1]$ satisfying $\rho(0) = 0$ and $\rho(1) = 1$. We, here, assume that the ramp function holds the algebraic order condition (3.1), i.e.,

$$\rho^{(i)}(0) = \rho^{(i)}(1) = 0 \quad \text{for} \quad i = 1, \ldots, n\,.$$

A slow manifold Q_n of this model is constructed by a power series expansion,

$$Q_n(\boldsymbol{r}, t) = \sum_{i=0}^{n} q_i(\boldsymbol{r}, t)\varepsilon^i\,,$$

where the recursive functions q_i are defined by

$$\begin{aligned} q_0(\boldsymbol{r}, t) &= -\rho(t/T) J \boldsymbol{\nabla} V(\boldsymbol{r})\,, \\ q_k(\boldsymbol{r}, t) &= -J\partial_t q_{k-1} - J \sum_{i+j=k-1} Dq_i(\boldsymbol{x}, t) q_j(\boldsymbol{x}, t)\,, \end{aligned}$$

for $k = 1, \cdots, n$. As $\boldsymbol{p}$ denotes the fast and slow motions together, the fast motion is given by the remainder,

$$\boldsymbol{w}(t) = \boldsymbol{p}(t) - Q_{n+1}(\boldsymbol{r}, t)\,, \tag{3.1.3}$$

which depends on the order n of the slow manifold. The model (3.1.1) can be rewritten in the $\boldsymbol{r}$-$\boldsymbol{w}$ variables as

$$\begin{aligned}
\dot{\boldsymbol{r}} &= Q_{n+1} + \boldsymbol{w}\,,\\
\dot{\boldsymbol{w}} &= \left(\frac{1}{\varepsilon}J - DQ_{n+1}\right)\boldsymbol{w} + \frac{1}{\varepsilon}\left(JQ_{n+1} - \rho\boldsymbol{\nabla}V\right) - \partial_t Q_{n+1} - DQ_{n+1}Q_{n+1}\,.
\end{aligned}$$

After the cancellation of some terms up to the order of ε^{n+1}, we obtain

$$\dot{\boldsymbol{r}} = Q_{n+1} + \boldsymbol{w}\,, \tag{3.1.4a}$$

$$\dot{\boldsymbol{w}} = \left(\frac{1}{\varepsilon}J - DQ_{n+1}\right)\boldsymbol{w} + O(\varepsilon^{n+1})\,, \tag{3.1.4b}$$

and the boundary conditions (3.1.2) correspond to

$$\boldsymbol{w}(0) = 0 \quad \text{and} \quad \boldsymbol{r}(T) = \boldsymbol{r}^*. \tag{3.1.5}$$

Our analysis is built on this slow-fast splitted form (3.1.4) with the imposed boundary conditions (3.1.5) instead of the original system (3.1.1). In the analysis, we do not consider beyond-all-order estimates, so the exponential order condition (2.4.7) of the ramp function and rigorous construction of the remainder term are not required, see Masur et al. (2022) for further results.

3.2 Some primary estimates

To use in our main result, we first provide a general estimate to bound smoothly differentiable functions in the following lemma and the generalisation of Gronwall's inequality. In the following, we extend the lemma in Hunter and Nachtergaele (2001), which bounds the function $g : \mathbb{R}^k \to \mathbb{R}$.

Lemma 3.2.1. *Suppose that $g : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ is smoothly differentiable function with bounded partial derivatives. Then, there exist a ball of radius R, $B_R \subset \mathbb{R}^{2d}$, and for all $\boldsymbol{x},\ \boldsymbol{y} \in B_R$, we have*

$$\|g(\boldsymbol{x}) - g(\boldsymbol{y})\| \leq C(R)\|\boldsymbol{x} - \boldsymbol{y}\|\,, \tag{3.2.1}$$

and

$$C(R) = \sup_{\boldsymbol{z}\in B_R} \|Dg(\boldsymbol{z})\|\,. \tag{3.2.2}$$

Proof. Using the fundamental theorem of calculus, we get

$$g(\boldsymbol{x}) - g(\boldsymbol{y}) = \int_0^1 Dg(s\boldsymbol{x} + (1-s)\boldsymbol{y})(\boldsymbol{x} - \boldsymbol{y})ds\,,$$

and taking the norm of the equation implies

$$\begin{aligned}
\|g(\boldsymbol{x}) - g(\boldsymbol{y})\| &\leq \left\|\int_0^1 Dg(s\boldsymbol{x} + (1-s)\boldsymbol{y})ds(\boldsymbol{x} - \boldsymbol{y})\right\|\\
&\leq \left\|\int_0^1 Dg(s\boldsymbol{x} + (1-s)\boldsymbol{y})ds\right\|\ \|\boldsymbol{x} - \boldsymbol{y}\|\\
&\leq \int_0^1 \|Dg(s\boldsymbol{x} + (1-s)\boldsymbol{y})\|ds\|\boldsymbol{x} - \boldsymbol{y}\|
\end{aligned}$$

by the Jensen's inequality (Hunter and Nachtergaele (2001)). The upper bound, then, becomes

$$\|g(\boldsymbol{x}) - g(\boldsymbol{y})\| \leq \sup_{0\leq s\leq 1} \|Dg(s\boldsymbol{x} + (1-s)\boldsymbol{y})\| \, \|(\boldsymbol{x} - \boldsymbol{y})\| \,,$$

and for a ball of radius R such that $\|\boldsymbol{x} - \boldsymbol{y}\| < R$,

$$\|g(\boldsymbol{x}) - g(\boldsymbol{y})\| \leq \sup_{\boldsymbol{z}\in B_R} \|Dg(\boldsymbol{z})\| \, \|(\boldsymbol{x} - \boldsymbol{y})\| \,.$$

With the coefficient (3.2.2), hence, the estimate (3.2.1) follows. □

The generalisation of Gronwall's inequality could be obtained from Chandra and Davis (1976). This generalisation bounds two linear inequalities in two variables with exact solutions of the corresponding equalities. In particular, we seek a bound for $\boldsymbol{z} \in \mathbb{R}^2$ satisfying

$$\dot{\boldsymbol{z}}(t) \leq A\boldsymbol{z}(t) + k \,,$$

where k is a two-dimensional vector, and $A \in \mathbb{R}^{2\times 2}$ is a positive semi-definite matrix, i.e, $\boldsymbol{x}^T A\boldsymbol{x} \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^2$. The following proposition provides the bound for this problem.

Proposition 3.2.2. *Suppose $\boldsymbol{z} \in \mathbb{R}^2$, k is a two-dimensional vector, and A is a positive semi-definite matrix such that*

$$\dot{\boldsymbol{z}}(t) \leq A\boldsymbol{z}(t) + k \,, \tag{3.2.3}$$

for $0 \leq t \leq T$, then

$$\boldsymbol{z}(t) \leq \boldsymbol{z}(0) + kt + \int_0^t \mathrm{e}^{A(t-s)} A(\boldsymbol{z}(0) + ks)ds \,. \tag{3.2.4}$$

Proof. The linear inequalities (3.2.3) implies

$$\boldsymbol{z}(t) \leq \boldsymbol{z}(0) + kt + \int_0^t A\boldsymbol{z}(s)ds \,. \tag{3.2.5}$$

We have the integral operator, say $\mathcal{K}$, defined by

$$\mathcal{K}z(t) = \int_0^t A\boldsymbol{z}(s)ds.$$

To use the argument in Chandra and Davis (1976), we need to show that $\mathcal{K}$ is a monotone operator, which follows from, for $\boldsymbol{z} = (z_1, z_2)$,

$$\begin{aligned}
\langle \boldsymbol{z}, \mathcal{K}\boldsymbol{z}\rangle &= \int_0^T \int_0^t \boldsymbol{z}(t) A\boldsymbol{z}(s) dsdt \\
&= \int_0^T \int_0^t \big(z_1(t) + z_2(t)\big)\big(z_1(s) + z_2(s)\big) dsdt \\
&= \int_0^T \big(z_1(t) + z_2(t)\big) \int_0^t \big(z_1(s) + z_2(s)\big) dsdt \\
&= \frac{1}{2}\left(\int_0^T \big(z_1(t) + z_2(t)\big)dt\right)^2 \geq 0 \,.
\end{aligned}$$

Using the monotonicity of $\mathcal{K}$, we apply their theorem on (3.2.5) and we obtain (3.2.4).

□

3.3 Algebraic estimate of the nudging scheme

The backward-forward nudging scheme is applied to the BVP (3.1.4) with its boundary conditions (3.1.5). The backward scheme gives the solution $(\boldsymbol{r}^-, \boldsymbol{w}^-)$ starting with an arbitrary fiber coordinate $\boldsymbol{w}_0$ and a fixed base point $\boldsymbol{r}^*$ such that $(\boldsymbol{w}_0, \boldsymbol{r}^*)$ is a point on Q_{n+1}, which are written as

$$\boldsymbol{r}^-(T) = \boldsymbol{r}^* \quad \text{and} \quad \boldsymbol{w}^-(T) = \boldsymbol{w}_0 \,.$$

The system is evolved back to $t = 0$ (linear end) and the forward scheme with the solution $(\boldsymbol{r}^+, \boldsymbol{w}^+)$ is initialised by

$$\boldsymbol{r}^+(0) = \boldsymbol{r}^-(0) \quad \text{and} \quad \boldsymbol{w}^+(0) = 0 \,. \tag{3.3.1}$$

After solving the forward scheme up to $t = T$ (nonlinear end), the base point is set while the other variable is preserved for the next iteration,

$$\boldsymbol{r}^+(T) = \boldsymbol{r}^* \quad \text{and} \quad \boldsymbol{w}^+(T) = \boldsymbol{w}^-(T) \,. \tag{3.3.2}$$

We, hence, define an optimal balance map,

$$\Phi : \boldsymbol{w}_0 \mapsto \boldsymbol{w}^+(T) \,,$$

and by calling $\boldsymbol{w}^+(T) = \boldsymbol{w}_1$, the iterative application of the map Φ constructs a sequence $(\boldsymbol{w}_m)$, i.e.,

$$\boldsymbol{w}_m = \Phi(\boldsymbol{w}_{m-1}) = \Phi^{m-1}(\boldsymbol{w}_0) \,.$$

According to this, in the nudging scheme each cycle ends with its iterates $(\boldsymbol{r}_m, \boldsymbol{w}_m)$ where $(\boldsymbol{r}_m)$ is the sequence of complementary component.

The analysis of the nudging scheme is described on the approximate slow-fast variables $\boldsymbol{r}$-$\boldsymbol{w}$ for convenience, but the variable $\boldsymbol{w}$ is only theoretically exist. As the scheme itself is a numerical process, it is a right approach to analyse using the variable $\boldsymbol{p}$ and then switch to $\boldsymbol{w}$ through the definition in (3.1.3), see Masur et al. (2022) for improved results.

Theorem 3.3.1. *Solve the optimal balance BVP* (3.1.4) *with boundary conditions* (3.1.5) *over time horizon* $t \in [0, T]$ *using the backward-forward nudging scheme. Assume that the ramp function* ρ *satisfies the algebraic order condition* (3.1)*. For any* $0 < \varepsilon \leq T$*, there exists* $\sigma = \sigma(\rho, T, n, V, R)$ *and* $\beta = \beta(\rho, T, n, V, R)$ *such that the optimal balance map is bounded as follows*

$$\|\Phi(\boldsymbol{w}_m) - \Phi(\widetilde{\boldsymbol{w}}_m)\| \leq \sigma \|\boldsymbol{w}_m - \widetilde{\boldsymbol{w}}_m\| + \beta \varepsilon^{n+1} \,. \tag{3.3.3}$$

where $\boldsymbol{w}_m$ *and* $\widetilde{\boldsymbol{w}}_m$ *are constructed by the nudging scheme.*

Proof. First, we want to find the energy estimates for $\boldsymbol{w}$ and $\boldsymbol{r}$, and to do so, we use the two systems with the variables $(\boldsymbol{r}^\pm, \boldsymbol{w}^\pm)$ and $(\widetilde{\boldsymbol{r}}^\pm, \widetilde{\boldsymbol{w}}^\pm)$. Subtracting the forward systems from each other, we get the first equation as

$$\frac{d}{dt}(\boldsymbol{r}^+ - \widetilde{\boldsymbol{r}}^+) = Q_{n+1}(\boldsymbol{r}^+, t) - Q_{n+1}(\widetilde{\boldsymbol{r}}^+, t) + (\boldsymbol{w}^+ - \widetilde{\boldsymbol{w}}^+) \,.$$

Taking the dot product with $(\boldsymbol{r}^+ - \widetilde{\boldsymbol{r}}^+)$ implies

$$\|\boldsymbol{r}^+ - \widetilde{\boldsymbol{r}}^+\| \frac{d}{dt} \|\boldsymbol{r}^+ - \widetilde{\boldsymbol{r}}^+\| = (\boldsymbol{r}^+ - \widetilde{\boldsymbol{r}}^+) \cdot \left(Q_{n+1}(\boldsymbol{r}^+, t) - Q_{n+1}(\widetilde{\boldsymbol{r}}^+, t) + (\boldsymbol{w}^+ - \widetilde{\boldsymbol{w}}^+)\right) .$$

It, then, follows

$$\begin{aligned}\|\boldsymbol{r}^+ - \widetilde{\boldsymbol{r}}^+\|\frac{d}{dt}\|\boldsymbol{r}^+ - \widetilde{\boldsymbol{r}}^+\| &= \left|(\boldsymbol{r}^+ - \widetilde{\boldsymbol{r}}^+)\cdot\left(Q_{n+1}(\boldsymbol{r}^+,t) - Q_{n+1}(\widetilde{\boldsymbol{r}}^+,t) + (\boldsymbol{w}^+ - \widetilde{\boldsymbol{w}}^+)\right)\right| \\ &\leq \|\boldsymbol{r}^+ - \widetilde{\boldsymbol{r}}^+\|\left(\|Q_{n+1}(\boldsymbol{r}^+,t) - Q_{n+1}(\widetilde{\boldsymbol{r}}^+,t)\| + \|\boldsymbol{w}^+ - \widetilde{\boldsymbol{w}}^+\|\right),\end{aligned}$$

where first the triangle inequality and later the Cauchy-Schwarz inequality are used, and we obtain

$$\frac{d}{dt}\|\boldsymbol{r}^+ - \widetilde{\boldsymbol{r}}^+\| \leq \|Q_{n+1}(\boldsymbol{r}^+,t) - Q_{n+1}(\widetilde{\boldsymbol{r}}^+,t)\| + \|\boldsymbol{w}^+ - \widetilde{\boldsymbol{w}}^+\|. \tag{3.3.4}$$

Depending on the smooth differentiability of ρ and V, each term of Q_{n+1} satisfy the assumptions of Lemma 3.2.1, therefore

$$\|Q_{n+1}(\boldsymbol{r}^+,t) - Q_{n+1}(\widetilde{\boldsymbol{r}}^+,t)\| \leq \left(c_0 + \cdots + c_{n+1}\varepsilon^{n+1}\right)\|\boldsymbol{r}^+ - \widetilde{\boldsymbol{r}}^+\|$$

where $c_i = c_i(\rho, V, T, m)$, $i = 0, \cdots, n+1$ are constants, with appropriate radius m, $\|\boldsymbol{r}^+ - \widetilde{\boldsymbol{r}}^+\| < m$. Since $\varepsilon < 1$, we choose $\delta_0 = \delta_0(\rho, V, T, m)$ to have constant $\delta_0 = c_0 + \cdots + c_{n+1}$ such that

$$\|Q_{n+1}(\boldsymbol{r}^+,t) - Q_{n+1}(\widetilde{\boldsymbol{r}}^+,t)\| \leq \delta_0\|\boldsymbol{r}^+ - \widetilde{\boldsymbol{r}}^+\|.$$

The inequality (3.3.4), hence, becomes

$$\frac{d}{dt}\|\boldsymbol{r}^+ - \widetilde{\boldsymbol{r}}^+\| \leq \delta_0\|(\boldsymbol{r}^+ - \widetilde{\boldsymbol{r}}^+)\| + \|\boldsymbol{w}^+ - \widetilde{\boldsymbol{w}}^+\|. \tag{3.3.5}$$

In the same manner, for the second equation, we get

$$\frac{d}{dt}(\boldsymbol{w}^+ - \widetilde{\boldsymbol{w}}^+) = \frac{1}{\varepsilon}J(\boldsymbol{w}^+ - \widetilde{\boldsymbol{w}}^+) - (DQ_{n+1}(\boldsymbol{r}^+,t)\boldsymbol{w}^+ - DQ_{n+1}(\widetilde{\boldsymbol{r}}^+,t)\widetilde{\boldsymbol{w}}^+) + O(\varepsilon^{n+1}).$$

After writing the second term on the right hand side as

$$DQ_{n+1}(\boldsymbol{r}^+,t)(\boldsymbol{w}^+ - \widetilde{\boldsymbol{w}}^+) + (DQ_{n+1}(\boldsymbol{r}^+,t) - DQ_{n+1}(\widetilde{\boldsymbol{r}}^+,t))\widetilde{\boldsymbol{w}}^+,$$

the dot product with $(\boldsymbol{w}^+ - \widetilde{\boldsymbol{w}}^+)$ gives

$$\frac{d}{dt}\|\boldsymbol{w}^+ - \widetilde{\boldsymbol{w}}^+\| \leq \|DQ_{n+1}(\boldsymbol{r}^+,t)(\boldsymbol{w}^+ - \widetilde{\boldsymbol{w}}^+)\| + \|DQ_{n+1}(\boldsymbol{r}^+,t) - DQ_{n+1}(\widetilde{\boldsymbol{r}}^+,t)\|\|\widetilde{\boldsymbol{w}}^+\| + O(\varepsilon^{n+1}).$$

In Theorem 2.4.1, the term $\widetilde{\boldsymbol{w}}^+$ is bounded above by $O(\varepsilon^{n+1})$, and using Lemma 3.2.1, we get

$$\frac{d}{dt}\|\boldsymbol{w}^+ - \widetilde{\boldsymbol{w}}^+\| \leq \delta_1\|\boldsymbol{w}^+ - \widetilde{\boldsymbol{w}}^+\| + \delta_2\|\boldsymbol{r}^+ - \widetilde{\boldsymbol{r}}^+\| + O(\varepsilon^{n+1}). \tag{3.3.6}$$

The above two energy estimates in (3.3.5) and (3.3.6) could be reformulated as a linear system of inequalities,

$$\dot{\boldsymbol{z}}^+(t) \leq \delta A\boldsymbol{z}^+(t) + k,$$

where

$$z^+(t) = \begin{pmatrix}\|\boldsymbol{r}^+(t) - \widetilde{\boldsymbol{r}}^+(t)\| \\ \|\boldsymbol{w}^+(t) - \widetilde{\boldsymbol{w}}^+(t)\|\end{pmatrix}, \qquad A = \begin{pmatrix}1 & 1 \\ 1 & 1\end{pmatrix}, \qquad k = \begin{pmatrix}0 \\ \alpha\varepsilon^{n+1}\end{pmatrix}.$$

for two constants $\delta = \max\{\delta_0, \delta_1, \delta_2\}$ and $\alpha = \alpha(\rho, T, n, V)$. To bound this linear system, we use using the Gronwall-type inequality Proposition 3.2.2 and obtain

$$\boldsymbol{z}(t) \leq \boldsymbol{z}(0) + kt + \delta A \int_0^t \mathrm{e}^{\delta A(t-s)}(\boldsymbol{z}(0) + ks)ds\,. \tag{3.3.7}$$

Since the matrix A satisfies $A^n = 2^{n-1}A$, then we have

$$A\mathrm{e}^{\delta At} = \sum_{n=0}^{\infty} \frac{A^{n+1}(\delta t)^n}{n!} = A\sum_{n=0}^{\infty} \frac{(2\delta t)^n}{n!} = A\mathrm{e}^{2\delta t}\,,$$

and $A\mathrm{e}^{-\delta At} = A\mathrm{e}^{-2\delta t}$. The integral operator in (3.3.7) is handled as follows

$$\delta A \int_0^t \mathrm{e}^{\delta A(t-s)}ds = \delta A \int_0^t \mathrm{e}^{2\delta(t-s)}ds = \frac{1}{2}A(\mathrm{e}^{2\delta t} - 1)\,,$$

and in the same manner,

$$\delta A \int_0^t \mathrm{e}^{\delta A(t-s)}sds = \delta A \int_0^t \mathrm{e}^{2\delta(t-s)}sds = A\left(-\frac{t}{2} + \frac{\mathrm{e}^{2\delta t} - 1}{4\delta}\right)\,.$$

The system of inequalities (3.3.7), hence, becomes

$$\boldsymbol{z}^+(t) \leq P_1(t)\boldsymbol{z}^+(0) + P_2(t)k\,, \tag{3.3.8}$$

where

$$P_1(t) = I + \frac{1}{2}A(\mathrm{e}^{2\delta t} - 1) \quad \text{and} \quad P_2(t) = t(I - \frac{1}{2}A) - \frac{\mathrm{e}^{2\delta t} - 1}{4\delta}A\,.$$

with the explicit expressions of

$$P_1(t) = \frac{1}{2}\begin{pmatrix} \mathrm{e}^{2\delta t} + 1 & \mathrm{e}^{2\delta t} - 1 \\ \mathrm{e}^{2\delta t} - 1 & \mathrm{e}^{2\delta t} + 1 \end{pmatrix} \quad \text{and} \quad P_2(t) = \begin{pmatrix} -t/2 + \frac{\mathrm{e}^{2\delta t}-1}{4\delta} \\ t/2 + \frac{\mathrm{e}^{2\delta t}-1}{4\delta} \end{pmatrix} \alpha\varepsilon^{n+1}\,.$$

For the system evolving back in time, we get the following estimate

$$\boldsymbol{z}^-(t) \leq P_1(T-t)\boldsymbol{z}^-(T) + P_2(T-t)k\,. \tag{3.3.9}$$

Regarding the boundary conditions (3.3.1) and (3.3.2), the inequality (3.3.8) at $t = T$ becomes

$$\|\boldsymbol{w}^+(T) - \widetilde{\boldsymbol{w}}^+(T)\| \leq \frac{1}{2}(\mathrm{e}^{2\delta T} - 1)\|\boldsymbol{r}^+(0) - \widetilde{\boldsymbol{r}}^+(0)\| + \left(\frac{T}{2} + \frac{\mathrm{e}^{2\delta T} - 1}{4\delta}\right)\alpha\varepsilon^{n+1}\,, \tag{3.3.10}$$

and the inequality (3.3.9) at $t = 0$ takes the form

$$\|\boldsymbol{r}^-(0) - \widetilde{\boldsymbol{r}}^-(0)\| \leq \frac{1}{2}(\mathrm{e}^{2\delta T} - 1)\|\boldsymbol{w}^-(T) - \widetilde{\boldsymbol{w}}^-(T)\| + \left(-\frac{T}{2} + \frac{\mathrm{e}^{2\delta T} - 1}{4\delta}\right)\alpha\varepsilon^{n+1}\,. \tag{3.3.11}$$

As $\boldsymbol{r}^-(0) = \boldsymbol{r}^+(0)$ and $\widetilde{\boldsymbol{r}}^-(0) = \widetilde{\boldsymbol{r}}^+(0)$ at the linear end, we insert (3.3.11) in (3.3.10). To write the estimate for any iterate, we use Φ on $\boldsymbol{w}_m$ and $\widetilde{\boldsymbol{w}}_m$ and obtain the inequality (3.3.3) where

$$\sigma = \frac{1}{4}(1 - \mathrm{e}^{2\delta T})^2 \quad \text{and} \quad \beta = \alpha\frac{T}{4}(3 - \mathrm{e}^{2\delta T}) + \alpha\frac{\mathrm{e}^{4\delta T} - 1}{8\delta}\,. \tag{3.3.12}$$

□

Corollary 3.3.2. *Consider the assumptions and settings of Theorem 3.3.1. There exists a constant $\beta > 0$ for all $0 < \varepsilon < T$, and we have, for any $\delta < \frac{\ln 3}{2T}$,*

$$\limsup_m \|\boldsymbol{w}_{m+1} - \boldsymbol{w}_m\| \leq \frac{1}{1-\sigma}\beta\varepsilon^{n+1}\,. \tag{3.3.13}$$

Proof. We estimate the difference of two consecutive terms from above using (3.3.3):

$$\begin{aligned}\|\boldsymbol{w}_{m+1} - \boldsymbol{w}_m\| &= \|\Phi(\boldsymbol{w}_m) - \Phi(\boldsymbol{w}_{m-1})\| \\ &\leq \sigma\|\boldsymbol{w}_m - \boldsymbol{w}_{m-1}\| + \beta\varepsilon^{n+1} \\ &\leq \sigma^2\|\boldsymbol{w}_{m-1} - \boldsymbol{w}_{m-2}\| + (\sigma+1)\beta\varepsilon^{n+1} \\ &\;\;\vdots \\ &\leq \sigma^m\|\boldsymbol{w}_0 - \boldsymbol{w}_1\| + \frac{1-\sigma^m}{1-\sigma}\beta\varepsilon^{n+1}\,.\end{aligned}$$

When δ is smaller than $\frac{\ln 3}{2T}$, $\sigma < 1$, which with $m \to \infty$ gives the estimate (3.3.13). □

As it is shown in the above corollary, two consecutive fast iterates in the nudging scheme may not converge in mathematical sense, but these iterates can get closer up to a residual of algebraic order as powers of ε, which we call this behaviour as "quasi-convergence". If the term $\beta/(1-\sigma)\varepsilon^{n+1}$ vanishes, given the continuous dependence of Φ on $\boldsymbol{r}^*$, the sequence $(\boldsymbol{w}_m)$ would be characterised as a Cauchy sequence since the same term also appears for any two member of $(\boldsymbol{w}_m)$, and then, there would exist a unique solution. The estimate (3.3.13), nevertheless, does not say whether the numerical solution of the nudging scheme is unique or depends continuously on the base point.

Theorem 3.3.3. *Assume that a ramp function $\rho \in C^{n+1}[0,1]$ satisfies the algebraic order condition* (3.1)*. The backward-forward nudging scheme is used to create nudging iterates $(\boldsymbol{r}_m, \boldsymbol{p}_m)$. For any $T < \frac{\ln 3}{2\delta}$, there exist a constant $C = C(\rho, T, n, V)$ such that the nudging iterates gets closer to the slow manifold with the following residual,*

$$\limsup_m \|\boldsymbol{p}_m - Q_{n+1}(\boldsymbol{r}^*, T)\| \leq C\varepsilon^{n+1}\,. \tag{3.3.14}$$

Proof. Notice that

$$\begin{aligned}\|\boldsymbol{p}_m - Q_{n+1}(\boldsymbol{r}^*, T)\| &\leq \|\boldsymbol{p}_m - Q_{n+1}(\boldsymbol{r}_m, T)\| + \|Q_{n+1}(\boldsymbol{r}_m, T) - Q_{n+1}(\boldsymbol{r}^*, T)\| \\ &\leq \|\boldsymbol{w}_m\| + \delta\|\boldsymbol{r}_m - \boldsymbol{r}^*\|\,.\end{aligned}$$

For the second term $\|\boldsymbol{r}_m - \boldsymbol{r}^*\|$, we write the first component of the inequality (3.3.8) at $t = T$:

$$\|\boldsymbol{r}^+(T) - \widetilde{\boldsymbol{r}}^+(T)\| \leq \frac{1}{2}(\mathrm{e}^{2T}+1)\|\boldsymbol{r}^+(0) - \widetilde{\boldsymbol{r}}^+(0)\| + \left(-\frac{T}{2} + \frac{\mathrm{e}^{2\delta T}-1}{4\delta}\right)\alpha\varepsilon^{n+1}\,.$$

Using the inequality (3.3.11) with $\boldsymbol{r}^-(0) = \boldsymbol{r}^+(0)$, it becomes

$$\|\boldsymbol{r}^+(T) - \widetilde{\boldsymbol{r}}^+(T)\| \leq \eta\|\boldsymbol{w}^-(T) - \widetilde{\boldsymbol{w}}^-(T)\| + \varkappa\varepsilon^{n+1}\,,$$

where

$$\eta = \frac{1}{4}(\mathrm{e}^{4\delta T} - 1) \quad \text{and} \quad \varkappa = \alpha\frac{T}{4}(-3 - \mathrm{e}^{2\delta T}) + \alpha\frac{\mathrm{e}^{2\delta T}-1}{4\delta}\frac{\mathrm{e}^{2\delta T}+3}{2}\,.$$

In the nudging scheme, two consecutive cycles require an additional condition $\boldsymbol{w}^-(T) = \boldsymbol{w}^+(T)$ together with fixing the base point coordinate $\boldsymbol{r}^*$. The above inequality for the iterates $(\boldsymbol{r}_m, \boldsymbol{w}_m)$, then, becomes

$$\|\boldsymbol{r}_m - \boldsymbol{r}^*\| \leq \eta \|\boldsymbol{w}_m - \boldsymbol{w}_{m-1}\| + \varkappa \alpha \varepsilon^{n+1} .$$

As $m \to \infty$, the estimate (3.3.13) yields

$$\limsup_m \|\boldsymbol{r}_m - \boldsymbol{r}^*\| \leq \Big(\frac{\eta\beta}{1-\sigma} + \varkappa \Big) \alpha \varepsilon^{n+1} .$$

From Theorem 2.4.1, we know that $\|\boldsymbol{w}(T)\|$ is of order ε^{n+1}, and moreover for a constant $\alpha = \alpha(\rho, T, n, V)$,

$$\|\boldsymbol{w}_m\| \leq \alpha \varepsilon^{n+1} , \quad \forall m > 1 .$$

Then, the estimate (3.3.14) follows. □

In our main result, the "nudging balance error" is estimated as the closeness of the balanced state $(\boldsymbol{r}^*, \boldsymbol{p}_m)$ to a point $Q_{n+1}(\boldsymbol{r}^*, T)$ on the slow manifold of order $n+1$. This estimate has two error components: i) the balance error $\|\boldsymbol{w}_m\|$ and ii) the termination residual $\|\boldsymbol{r}_m - \boldsymbol{r}^*\|$. The balance error is bounded in Gottwald et al. (2017), can be explicitly seen in Theorem 2.4.1. As the termination residual and this error has the same order of ε^{n+1}, the nudging balance error immediately follows. Our numerical test cases for the rotating shallow-water flows will prove that optimal balance produces well-balanced states, albeit the nudging cycles are terminated by the sufficient closeness of consecutive iterates, quasi-convergence of iterates.

In practical use, the contribution of the termination residual could be cancelled by a proper BVP solver, for example, the simple shooting method applied by Gottwald et al. (2017). When finite-dimensional problems are considered, however, this type of solvers cost computationally high, because the fast time needs to be resolved properly, also discussed in Section 2.4.1. We, then, conjecture the requirement of a numerical scheme to solve the backward-forward nudging integrations such that the scheme is stable in the presence of the fast time but does not need to be accurate in that scale.

For the rest of this thesis, we clarify the terminology depending on our theoretical result presented here: By the convergence of the nudging scheme, we always mean its quasi-convergence without loss of generality.

Chapter 4

Optimal balance for shallow-water flows

In this chapter, we apply the method of optimal balance to the rotating shallow-water equations written in primitive height and velocity variables. After analysing time scales of the equations, the optimal balance BVP is explicitly defined with two boundary conditions. At the linear-end boundary, we require linear wave separation through spectral mode decomposition or partial differential equation (PDE) based approaches. At the nonlinear-end boundary, on the other hand, we require some kinematic transformations of the primitive variables and PV-inversion formulas. This chapter together with the following two chapters describe our original work already published in Masur and Oliver (2020).

4.1 Eulerian time scales of the shallow-water equations

The rotating shallow-water equations are fully described in Section 2.2.1, and here, we recall the equations as a brief reminder before building up the application of optimal balance: The nondimensional shallow-water model reads

$$\varepsilon\left(\partial_t \boldsymbol{u} + \boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{u}\right) + \boldsymbol{u}^{\perp} + \frac{1}{h_0}\frac{\mathrm{Bu}}{\varepsilon}\boldsymbol{\nabla} h = 0\,, \tag{4.1.1a}$$

$$\partial_t h + \boldsymbol{\nabla}\cdot(h\boldsymbol{u}) + h_0\boldsymbol{\nabla}\cdot\boldsymbol{u} = 0\,, \tag{4.1.1b}$$

with the Rossby number ε (2.2.3), the Burger number Bu (2.2.4) and the nondimensional layer depth h_0 (2.2.5). Two distinguished scaling limits give geostrophic balance (2.2.7) at leading order: the quasi-geostrophic and semi-geostrophic scaling limits, characterised by $h_0 = \varepsilon^{-1}$ and $h_0 = 1$, respectively. We will work with these two scaling regimes to analyse the performance of optimal balance.

The time scales of Eulerian dynamics is different in the quasi-geostrophic vs. the semi-geostrophic scaling limits. To determine these time scales, we substitute the balance relation $u_{\mathrm{g}}+O(\varepsilon)$ in (2.4.3) into the continuity equation (4.1.1b) with the respective scaling for h_0. In the quasi-geostrophic scaling, we obtain

$$\begin{aligned}\partial_t h &= -\boldsymbol{\nabla}\cdot(h\boldsymbol{u}) - 1/\varepsilon\boldsymbol{\nabla}\cdot\boldsymbol{u}\\ &= -\boldsymbol{\nabla}\cdot(h\ (\boldsymbol{u}_{\mathrm{g}} + O(\varepsilon))) - 1/\varepsilon\boldsymbol{\nabla}\cdot(\boldsymbol{u}_{\mathrm{g}} + O(\varepsilon))\\ &= -\boldsymbol{\nabla}\cdot(h\ \boldsymbol{u}_{\mathrm{g}}) - 1/\varepsilon\boldsymbol{\nabla}\cdot\boldsymbol{u}_{\mathrm{g}} + O(1)\end{aligned}$$

where the terms of $\boldsymbol{\nabla}\cdot\boldsymbol{u}_{\mathrm{g}}$ vanish. In the semi-geostrophic scaling, following the same manner, we obtain

$$\partial_t h = -\boldsymbol{\nabla}\cdot(h\boldsymbol{u}) - \boldsymbol{\nabla}\cdot\boldsymbol{u} = O(\varepsilon)\,.$$

The estimates yield that Eulerian dynamics evolve on a time scale of $O(1)$ in the quasi-geostrophic scaling and $O(\varepsilon^{-1})$ in the semi-geostrophic scaling. In the numerical studies of optimal balance, we scale time interval of dynamical evolution by ε^{-1} in the semi-geostrophic scaling, and we choose the time interval independent of ε in the quasi-geostrophic scaling. Similar analysis for L1-balance model in the semi-geostrophic scaling is already carried out by Dritschel et al. (2017).

4.2 Optimal balance in primitive variables

The theory of optimal balance has been applied in different approaches due to the equations of motion and associated variables (Section 2.4). The difference in approaches comes from the deformation type of nonlinear interactions: The implicit deformation in the optimal PV balance (Viúdez and Dritschel, 2004) and the explicit deformation in optimal balance (Cotter, 2013; Gottwald et al., 2017). We, now, apply optimal balance the rotating shallow-water model (4.1.1). The formulation of the model allows the use of explicit deformation, then the optimal balance BVP becomes

$$\varepsilon\,(\partial_\tau \boldsymbol{u} + \rho(\tau/T)\boldsymbol{u}\cdot\boldsymbol{\nabla}\boldsymbol{u}) + \boldsymbol{u}^\perp + \boldsymbol{\nabla} h = 0\,, \tag{4.2.1a}$$

$$\partial_\tau h \;+\; \rho(\tau/T)\boldsymbol{\nabla}\cdot(h\boldsymbol{u}) + h_0\boldsymbol{\nabla}\cdot\boldsymbol{u} = 0\,, \tag{4.2.1b}$$

which needs to be closed at the temporal-end points.

At the linear end, $\tau = 0$, the exact decomposition of fast and slow dynamics is achieved by the boundary condition,

$$\mathbb{P}_{\mathrm{GW}}(\boldsymbol{u}, h)\big|_{\tau=0} = 0\,, \tag{4.2.2}$$

where the matrix $\mathbb{P}_{\mathrm{GW}}$ projects the linear flow onto its fast imbalanced component, gravity wave, and therefore, only slow balanced component, Rossby wave, remains on trivial slow manifold Γ_0. At the nonlinear end, $\tau = T$, a prescribed base-point coordinate is applied. As motivated in Section 2.4.1, the choice of base points are the PV q or the height h fields. We fix the base-point coordinates,

$$q(T) = q^* \quad \text{or} \quad h(T) = h^*\,. \tag{4.2.3}$$

with specific q^* or h^* fields on an approximate slow manifold Γ_T.

The optimal balance BVP (4.2.1) is, hence, set up with its boundary conditions, (4.2.2) and (4.2.3). We need to built the concrete structure of end boundaries: the derivation of $\mathbb{P}_{\mathrm{GW}}$ and some useful transformations for base point q, to be introduced in the rest of the chapter.

4.3 Linear wave separation

The optimal balance BVP (4.2.1) at $\tau = 0$ is reduced to the linear rotating shallow-water equations (2.2.8) which dynamically preserve the *linear potential vorticity*,

$$q_{\mathrm{lin}} = \varepsilon\,\zeta - h/h_0\,. \tag{4.3.1}$$

This linear quantity can be derived by linearising the nonlinear PV of the full shallow-water model (4.1.1), which is to be introduced in the upcoming section, as well as, a direct computation from the linear system (2.2.8). We apply the linear-end boundary condition (4.2.2) to the model (4.2.1) with two methods: a normal-mode decomposition and an equivalent PDE-based approach.

4.3.1 Normal-mode decomposition

The linear-end boundary condition can be applied via the oblique projector $\mathbb{P}_{\boldsymbol{k}}^{\mathrm{RW,obliq}}$ and orthogonal projector $\mathbb{P}_{\boldsymbol{k}}^{\mathrm{RW,orth}}$ derived in Section 2.2.2. For wave-number vector $\boldsymbol{k} = (k, l)$ with k in x direction and l in y direction, these spectral projectors are

$$\mathbb{P}_{\boldsymbol{k}}^{\mathrm{RW,obliq}} = \frac{1}{h_0\,|\boldsymbol{k}|^2 + \varepsilon^{-1}} \begin{pmatrix} h_0\,l^2 & -h_0\,kl & -\mathrm{i}l/\varepsilon \\ -h_0\,kl & h_0\,k^2 & \mathrm{i}k/\varepsilon \\ \mathrm{i}h_0\,l & -\mathrm{i}h_0\,k & 1/\varepsilon \end{pmatrix} \tag{4.3.2}$$

and

$$\mathbb{P}_{\boldsymbol{k}}^{\mathrm{RW,orth}} = \frac{\boldsymbol{\upsilon}_{\boldsymbol{k},0}\,\boldsymbol{\upsilon}_{\boldsymbol{k},0}^*}{\|\boldsymbol{\upsilon}_{\boldsymbol{k},0}\|^2} = \frac{1}{|\boldsymbol{k}|^2 + 1} \begin{pmatrix} l^2 & -kl & -\mathrm{i}l \\ -kl & k^2 & \mathrm{i}k \\ \mathrm{i}l & -\mathrm{i}k & 1 \end{pmatrix}, \tag{4.3.3}$$

where $\boldsymbol{\upsilon}_{\boldsymbol{k},0}$ is the eigenvector corresponding the Rossby-wave mode, $\lambda_{\boldsymbol{k}}^{\mathrm{RW}} = 0$, and $\boldsymbol{\upsilon}_{\boldsymbol{k},0}^*$ denotes the Hermitian conjugate.

To analyse the oblique and orthogonal projectors, we compare the projectors, and the position of the Rossby-wave and gravity-wave modes in respect of each other. The orthogonal projector acts independent of ε and h_0; however, the characteristics of the oblique projector change depending on these parameters. In the quasi-geostrophic scaling, where $h_0 = \varepsilon^{-1}$, as the subspaces of Rossby-wave and gravity-wave modes are orthogonal, the oblique projector corresponds to the orthogonal one. In the semi-geostrophic scaling, where $h_0 = 1$, the subspaces of the modes are oblique, then taking the oblique projector as reference, we examine the behaviour of the orthogonal projector in details in Section 6.4.

4.3.2 PDE-based approach

The approach of building projection matrices by normal-mode decomposition is easy on the periodic domain, but this approach is fluid-model and fluid-domain dependent, and it might be tiresome or it might not be, even, practically available for more general ocean models and more general domains. In these general cases, we require PDE-based approaches which implicitly apply the linear-end projection. Since the linear boundary preserves slow dynamics spanned via the Rossby modes and the geostrophic balance (2.2.7) supplies the exact Rossby-mode representation for the linearised system, we can use this balance for the implicit PDE-based approaches.

First, the most natural choice is to "preserve h" and compute $\boldsymbol{u}$ by the geostrophic balance (2.2.7). The condition is advantageous to be easily computed by any other scheme rather than spectral differentiation. After application of the condition, we evolve the optimal balance problem (4.2.1) starting from the linear system (2.2.8), where the geostrophic $\boldsymbol{u}$ field is placed, and we experience the loss of order one derivative: As a disadvantage, "preserve h" gives rise to "loss of derivatives" in each recursive iteration which injects small-scale noise and, ultimately, results in failure of convergence in the backward-forward nudging scheme. Moreover, for large ε values, we expect the convergence issue become stronger regardless the choice of the base point, since in the large-ε case, the h field involves higher amplitude of variations mostly with steep gradient.

Second, to deal with the issue of "loss of derivatives", we offer an approach to preserve vorticity $\zeta = \boldsymbol{\nabla}^{\perp} \cdot \boldsymbol{u}$ for the given $\boldsymbol{u}$ field. The h field is obtained in the manner of a

stream function by solving

$$\Delta h = \zeta\,, \tag{4.3.4}$$

then the $\boldsymbol{u}$ field is constructed only by its divergence-free component – the curl-free component is removed– using (2.2.7). In this formulation, one derivative for ζ and double integral for h followed by another derivative for $\boldsymbol{u}$: No order of derivatives is lost. We theoretically expect better convergence in "preserve ζ" than "preserve h"; however, this alternative approach do not perform well in our numerical results as it is foreseen.

Third, as the nonlinear PV can be a candidate as base point, we can use the corresponding linear PV q_{lin} (4.3.1) at the linear end and construct a projector to "preserve q_{lin}". The linear quantity is advectively conserved by the linear system, while the nonlinear one is materially conserved by the nonlinear system. The q_{lin}-preserving projector is constructed in the following way: We take a new state $(\boldsymbol{u}', h')$ to be determined which satisfies the geostrophic balance,

$$\boldsymbol{u}' = \boldsymbol{\nabla}^{\perp} h', \tag{4.3.5}$$

and $(\boldsymbol{u}', h')$ is restricted to preserve q_{lin} obtained by an already-known state $(\boldsymbol{u}, h)$,

$$q_{\text{lin}} = \varepsilon\, \boldsymbol{\nabla}^{\perp} \cdot \boldsymbol{u} - h/h_0\,. \tag{4.3.6}$$

We obtain h' from q_{lin} by

$$\varepsilon\, \Delta h' - h'/h_0 = q_{\text{lin}}\,,$$

or in other words, h' is found by solving the constant-coefficient Helmholtz operator,

$$h' = (-h_0^{-1} + \varepsilon\, \Delta)^{-1} (\varepsilon\, \boldsymbol{\nabla}^{\perp} \cdot \boldsymbol{u} - h/h_0)\,. \tag{4.3.7}$$

The computation of $\boldsymbol{u}'$ follows from the assumption (4.3.5). Since the above equations include linear operators, q_{lin} (4.3.6) is written in spectral representation as follows

$$q_{\text{lin},\boldsymbol{k}} = \begin{pmatrix} -\varepsilon\, \mathrm{i} h_0\, l & \varepsilon\, \mathrm{i} h_0\, k & -h_0^{-1} \end{pmatrix} \boldsymbol{z}_{\boldsymbol{k}}\,,$$

while the equations in (4.3.7) and (4.3.5) correspond to

$$h'_{\boldsymbol{k}} = \frac{-1}{h_0^{-1} + \varepsilon\, |\boldsymbol{k}|^2}\; q_{\text{lin},\boldsymbol{k}}\,, \quad \text{and} \quad \boldsymbol{u}'_{\boldsymbol{k}} = \begin{pmatrix} -\mathrm{i} l \\ \mathrm{i} k \end{pmatrix} h'_{\boldsymbol{k}}\,,$$

and hence, $z_{\boldsymbol{k}} = (\boldsymbol{u}'_{\boldsymbol{k}}, h'_{\boldsymbol{k}})$ is easily written as

$$\boldsymbol{z}'_{\boldsymbol{k}} = \frac{1}{h_0\, |\boldsymbol{k}|^2 + \varepsilon^{-1}} \begin{pmatrix} -\mathrm{i} l \\ \mathrm{i} k \\ 1 \end{pmatrix} \begin{pmatrix} \mathrm{i} h_0\, l & -\mathrm{i} h_0\, k & \varepsilon^{-1} \end{pmatrix} \boldsymbol{z}_{\boldsymbol{k}}\,.$$

The matrix operator applied to $\boldsymbol{z}_{\boldsymbol{k}}$ is the same as the oblique projector in (4.3.2). The oblique projector is explained in terms of the PDEs in (4.3.5) and (4.3.7), which can be used without available explicit mode decomposition. This equivalence becomes visible while formulating a projector to preserve the existing linear PV. In the results, we notice the advantages of PV-based projectors at both linear and nonlinear ends, which supports the use of PV in numerical balance models.

4.4 Nonlinear-end boundary condition

The problem (4.2.1) at $\tau = T$ becomes fully nonlinear, and we apply the nonliner-end boundary (4.2.3) to this nonlinear system. When h is considered as base point, the application of the boundary condition, $h(T) = h^*$, is simple. When q is set as base point, it requires some kinematic equations. The base point h neglects these equations from the backward-forward nudging scheme, and the simplified scheme involves only the primitive $\boldsymbol{u}$-h variables.

4.4.1 Geostrophic-ageostrophic variables

The fluid dynamics in our setting are evolved in the $\boldsymbol{u}$-h variables. Preservation of base point q, however, requires to convert the given $(\boldsymbol{u}, h)$ fields into the geostrophic-ageostrophic variables: PV q, velocity divergence δ, and ageostrophic vorticity γ. With the relative vorticity $\zeta = \boldsymbol{\nabla}^\perp \cdot \boldsymbol{u}$, the q-δ-γ variables are defined by

$$q = \frac{\varepsilon\zeta + 1}{h_0 + h}\,, \qquad \delta = \boldsymbol{\nabla} \cdot \boldsymbol{u}\,, \qquad \gamma = \zeta - \Delta h\,, \tag{4.4.1}$$

and we call them the transformations for geostrophic-ageostrophic variables. If the $\boldsymbol{u}$-h fields are geostrophically balanced, the ageostrophic δ-γ variables vanish at leading order, since $\boldsymbol{\nabla} \cdot \boldsymbol{\nabla}^\perp h = 0$ and inserting $\zeta = \Delta h$ in the γ equation cancels these terms.

4.4.2 PV-inversion equations

The flow states in the (q, δ, γ) variables need to be inverted into the corresponding $(\boldsymbol{u}, h)$ due to the formulation of the fluid dynamics. These inverse transformations are called PV-inversion equations. First, we want to recover velocity $\boldsymbol{u}$ field. The (q, δ, γ) definitions in (4.4.1) do not carry the information of mean velocity $\bar{\boldsymbol{u}}$, so that $\boldsymbol{u}$ can be established up to its mean $\bar{\boldsymbol{u}}$. In our case, $\bar{\boldsymbol{u}}$ is known due to the evolution of fluid dynamics in the $\boldsymbol{u}$-h variables, in other words, $\bar{\boldsymbol{u}}$ is also preserved along with base point q. To determine the full velocity field $\boldsymbol{u}$, we use the Helmholtz decomposition which splits the mean free velocity, $\boldsymbol{u} - \bar{\boldsymbol{u}}$, into its divergence-free and curl-free components,

$$\boldsymbol{u} = \boldsymbol{\nabla}^\perp \psi + \boldsymbol{\nabla}\phi + \bar{\boldsymbol{u}}, \tag{4.4.2}$$

with the stream function ψ and the velocity potential ϕ. These functions are obtained by solving two Poisson equations, $\Delta\psi = \zeta$ and $\Delta\phi = \delta$, on the doubly periodic domain.

Second, we insert ζ obtained by the γ definition into the q definition and obtain the following problem,

$$(-q + \varepsilon\Delta)h = -\varepsilon\gamma + qh_0 - 1\,. \tag{4.4.3}$$

To recover the height field h, this problem is treated as the Helmholtz equation with a constant-coefficient operator. We split q into its mean $\bar{q}$ and mean-free $q - \bar{q}$ components, and the constant-coefficient $(-\bar{q}+\varepsilon\Delta)$ remains on the left hand side. The obtained problem is solved iteratively and quickly converges provided that $\bar{q} > 0$ and $q - \bar{q}$ is sufficiently small. At the end, we, in total, solve three different linear elliptic equations to obtain $(\boldsymbol{u}, h)$ from given (q, δ, γ).

Chapter 5

Experimental set-up

The numerical implementation of the optimal balance method for shallow water flows is presented in this chapter. For this implementation, we modified an existing rotating shallow-water code, which is the open source code PyRsw by Poulin (2016). Our modified numerical code is also accessible (Masur, 2022). In numerical experiments, the BVP uses different ramp functions and initial conditions. In the backward-forward nudging scheme, the solution of the BVP, or a balanced state, is found as a point close to the given base-point coordinate within some tolerance. The quality of the balanced state is determined by a special diagnostics, the diagnosed imbalance. Different than our work Masur and Oliver (2020), the complete set-up of the nudging scheme is extensively explained, supported by a schematic for better comprehension. Through this chapter, we consider all time variables as in time scale of the quasi-geostrophic scaling, which is of $O(1)$. To adjust them to the semi-geostrophic scaling, we scale the physical time t and the artificial time τ variables with ε^{-1}.

5.1 Numerical implementation

The numerical model is set up to evolve either in artificial time τ or in physical time t. Through τ evolution, we solve the optimal balance BVP (4.2.1), where the ramp function is employed. Although the scheme is introduced in Section 2.4.1, we briefly remind: The scheme integrates the problem backward in time as a final value problem starting from $\tau = T$, and forward in time as an initial value problem starting from $\tau = 0$. At the boundaries, the respective boundary conditions, (4.2.2) and (4.2.3), are imposed on shallow-water flows which are linear at $\tau = 0$ and nonlinear at $\tau = T$. The scheme together with setting the boundaries goes on iteratively until getting sufficiently close to a fixed base-point coordinate and, consequently, reaches a balanced state. Throughout the t evolution, the ramp function is turned off, so we solve the nonlinear shallow-water equations (4.1.1) starting from $t = 0$ up to some time $t = t'$ in a usual manner.

The numerical settings in the model are as follows: The 2π-doubly-periodic domain keeps the same grid resolution $n = 256$ in the x and y directions. The spatial derivatives are computed in the spectral space, while the nonlinear products are computed in the physical space. To remove aliasing, we use a spectral filter with full dealising according to the $2/3$ rule; then, the maximal resolved wave-number becomes $n/3$. The flow model is adaptively

evolved by the third-order Adams-Bashforth method, and time step is computed by

$$\Delta t_{\boldsymbol{u}} = n_{\text{cfl}} \frac{\Delta x}{|u|_{max} + 2\varepsilon^{-1}}$$

with the CFL number $n_{\text{cfl}} = 0.5$. For the purpose of substantial analysis, the model set-up gives permission to reach geostrophic-ageostrophic q-δ-γ fields via the equations (4.4.1) at each time step. We, additionally, checked other time steppers and time steppings for ageostrophic fields, Δt_δ and Δt_γ, where Δt_γ turned the shortest time step. Our results do not, nevertheless, depend on the choice of time stepper and time stepping.

Our numerical model is integrated over the ramp-time length T and the physical-time length t' which are of order of one eddy turnover time or less, so that dissipation does not play a crucial role through the evolution of the model. Though the dissipation is not essential, some test cases are, still, carried out including the non-dimensional viscosity term, $\varepsilon/\text{Re}\Delta\boldsymbol{u}$ where Reynolds number is

$$\text{Re} = \frac{UL}{\nu} \tag{5.1.1}$$

with the kinematic viscosity ν. Activated in the numerical model, the viscosity term is treated as any other nonlinear term, i.e., it is deformed by the ramp function over the τ evolution. The model is stable, when the viscosity term is applied dissipatively in the direction of integration. It means that the viscosity term has a reverse sign in front to keep dissipative manner through the integration backward in time. Moreover, optimal balance supports our choice of the viscous dissipation, as it executes primarily slow fields by its setting, and enstrophy transfer to small scales is prevented. Eddy viscosity or turbulent viscosity is, therefore, unnecessary.

5.2 Ramp functions

Different ramp functions took place in the optimal balance method up to so far. In the optimal PV balance, Viúdez and Dritschel (2004) used the cosine ramp function

$$\rho(\theta) = (1 - \cos(\pi\theta))/2\,; \tag{5.2.1}$$

in optimal balance, Gottwald et al. (2017) implemented ramp functions of the form

$$\rho(\theta) = \frac{f(\theta)}{f(\theta) + f(1-\theta)}\,, \tag{5.2.2}$$

where the choice $f(\theta) = \theta^k$ corresponds to a method of algebraic order k and the exponential ramp with $f(\theta) = \exp(-1/\theta)$. Being different than the former work, the analytical aspects of ramp functions are studied in the later one which decides the order of balance, see in Section 2.4.2. The cosine ramp (5.2.1) is determined as second order with respect to its analytical aspect. In our numerical implementation, we investigate the effect of ramp functions on optimal balance with the given choices above.

5.3 Initial conditions

The initial configurations to start numerical test cases are, randomly, generated such that the features of these configurations in spectral energy can be regulated. At every wave number $\boldsymbol{k}$, spectral energy density of h is described by

$$\mathcal{S}_h = \frac{|\boldsymbol{k}|^7}{(|\boldsymbol{k}|^2 + ak_0^2)^{2b}}$$

with $a = 4b/7 - 1$ and $b = (7 + d)/4$, so that $S_h \sim k^{-d}$ when $k \to \infty$. The parameter k_0 sets the maximum of the spectral energy density at $|\boldsymbol{k}| = k_0$, while d sets the order of spectral decay. Unless we analyse the effect of initial-condition structure on optimal balance, we create a random h perturbation with $k_0 = 6$ and $d = 6$. We scale this h perturbation to deviate from the equilibrium with $|h| < 1/5$. The total height $h + h_0$ is, then, built by adding the mean-free h component to the mean height: $h_0 = \varepsilon^{-1}$ for the quasi-geostrophic scaling and $h_0 = 1$ for the semi-geostrophic scaling.

After h is randomly generated, we choose to build the velocity fields $\boldsymbol{u}$ of the initial configuration in two ways: setting geostrophic velocity $\boldsymbol{u} = \boldsymbol{u}_\mathrm{g}$ by (2.2.7) and setting zero velocity $\boldsymbol{u} = 0$. The majority of our test cases are initialised by the geostrophic configuration $(\boldsymbol{u}_\mathrm{g}, h)$. When optimal balance is run using the base point h, we expect that test runs with these two configurations converge the same optimally balanced state, where the speed of convergence may not be similar. When it is run using base point q, the test runs differ in both balanced states and the convergence speed due to having different PV q fields.

5.4 Diagnosed imbalance

The quality of balanced states can be assessed by the balance error, for base point q

$$\|(\delta_T, \gamma_T) - G_n(q^*)\|\,, \tag{5.4.1}$$

with the reference slow vector field G_n, see Section 2.4.1. At the end of each nudging iteration, we get iterates $(\boldsymbol{u}_T, h_T)$, where $\tau = T$ is written as subscript for simplicity. As our numerical setting transforms between two different type of fields, the iterate can be represented by $(q_T, \delta_T, \gamma_T)$. For base point h, $G_n(q^*)$ is replaced by $G_n(h^*)$. The description of the slow manifold G_n field is not directly accessible, so that the balanced state obtained by optimal truncation of G_n is not available. To estimate the balance error, we adopt a special representation called *diagnosed imbalance* in the work of Gottwald et al. (2017).

The diagnosed imbalance to the rotating shallow-water equations executes the following steps:

i) Balance the initial flow $(\boldsymbol{u}^*, h^*)$ by solving the optimal balance BVP (4.2.1) over the ramp time T at $t = 0$ and obtain the balanced state $(\boldsymbol{u}_T, h_T)$.

ii) Evolve the balance flow $(\boldsymbol{u}_T, h_T)$ through the shallow-water equations (4.1.1) starting from $t = 0$ up to the physical time t' and the flow at final time is indicated by $(\boldsymbol{u}', h')$.

iii) Rebalance the evolved flow $(\boldsymbol{u}', h')$ by optimal balance at $t = t'$ and reach the new balanced state $(\boldsymbol{u}'_T, h'_T)$.

iv) Using the imbalance fields of the evolved flow $(\boldsymbol{u}', h')$ and those of the rebalanced flow $(\boldsymbol{u}'_T, h'_T)$, define the diagnosed imbalance by the following symmetrised relative error:

$$I = \frac{\|(\delta', \gamma') - (\delta'_T, \gamma'_T)\|}{\frac{1}{2}\left(\|(\delta', \gamma')\| + \|(\delta'_T, \gamma'_T)\|\right)}\,, \tag{5.4.2}$$

where $\|\cdot\|$ denotes the Euclidean norm.

The absolute error is used in the diagnosed imbalance (5.4.2) as an alternative to is replaced by the relative error, and the diagnosed imbalance becomes

$$I = \|(\delta', \gamma') - (\delta'_T, \gamma'_T)\|\,. \tag{5.4.3}$$

These two errors can be also formulated for the velocity $\boldsymbol{u}$ and the height h separately. Although the choice of the diagnostic norm might alter the computation of imbalances by order, we anticipate similar scaling of the diagnosed imbalanced in numerical tests results, see Section 6.7.2.

5.5 Stopping criterion

The stopping criterion is a condition to stop the iterative nudging scheme, when the chosen nonlinear-end boundary condition is fulfilled up to some tolerance. With the choice of base point q, in particular, the nudging iterates q_T is expected to converge to the given base point q^*. In the nudging scheme, the speed of convergence differs for some combinations of the linear-end and the nonlinear-end boundary conditions. We, therefore, let the shallow-water flow get out of the iterative loop, when the relative difference between two successive iterates of the nudging scheme are below the prescribed tolerance. The value of this tolerance is set to terminate each combinations of boundary conditions. The stopping criterion is, then, determined as follows:

$$\frac{\|q_T(n+1) - q_T(n)\|}{\frac{1}{2}\left(\|q_T(n+1)\| + \|q_T(n)\|\right)} \leq \kappa\,, \tag{5.5.1}$$

where $q_T(n)$ and $q_T(n+1)$ denote successive iterates, and the convergence tolerance parameter is $\kappa = 10^{-4}$.

When h is used as base-point coordinate, we replace h instead of q in the relative error (5.5.1). Keeping the same tolerance of base point q is, however, not feasible for the current base point, it needs to be updated. The height field h includes slow and fast dynamics at the same time unlike the slow variable q. The nudging scheme, thus, provides very slow convergence for $\varepsilon \approx 1$, and the tolerance κ is forced to be 0.02, which is gradually decreasing to 10^{-4} when ε gets smaller.

The stopping criterion includes two different parameters to be tested: the computational norm and the tolerance κ. We perform both parameters only for base point q, because the κ values for base point h are determined in a way that no further convergence are observed. First, as done in the diagnosed imbalanced, the absolute error is used in the criterion as follows

$$\|q_T(n+1) - q_T(n)\| \leq \kappa\,, \tag{5.5.2}$$

Second, the smaller κ values can be investigated to terminate the nudging scheme. As a result, both parameters give reason to the same impact, increase in the number of

iterations. The quality of convergence to the prescribed base points is affected by these parameters; however, the quality of balance is independent of them. Further explanation and numerical test cases can be seen in Section 6.6.

5.6 The complete set-up

The complete set-up of our numerical implementation consists of the backward-forward nudging scheme and the diagnosed imbalance, which respectively provide the balanced states and computes the error of these balanced states. We, here, describe this set-up as an computational algorithm and, further, schematise it in Figure 5.1, where the quasi-geostrophic scaling limit and base point q are considered.

The balancing procedure includes backward and forward integrations, and the application of the linear-end and the nonlinear-end boundary conditions, which is the nudging scheme itself. At the frozen real time $t = 0$ (Step i)), the optimal balance BVP (4.2.1) initialised with $(\boldsymbol{u}^*, h^*)$ is integrated backward in the artificial time starting from $\tau = T$ down to $\tau = 0$, and through ramping the nonlinear terms in the system, the problem becomes linear at final step. The gravity-wave component in the linear flow is vanished by the application of the linear-end boundary condition (4.2.2). The slow flow on a trivial slow manifold Γ_0 is integrated forward up to $\tau = T$, and the problem becomes fully nonlinear again on an approximate slow manifold Γ_T.

The nonlinear-end boundary condition (4.2.3) is set according to the PV q^* or height h^* of the initial flow $(\boldsymbol{u}^*, h^*)$. When base point q is chosen, we derive geostrophic-ageostrophic variables by the kinematic transformations (4.4.1): $(q^*, \delta^*, \gamma^*)$ of the initial configuration $(\boldsymbol{u}^*, h^*)$ and $(q_T, \delta_T, \gamma_T)$ of the nudging iterate $(\boldsymbol{u}_T, h_T)$. Depending on q^* and q_T, we check if the stopping criterion (5.5.1) is satisfied to stop the balancing loop. The stopping criterion is defined for two successive nudging iterates, $q_T(n)$ and $q_T(n+1)$, but the schematics in Figure 5.1 represents, for simplicity, only the first iteration.

a) If the criterion fails, q_T is set to q^* while preserving the ageostrophic δ_T-γ_T variables, i.e., the flow $(q^*, \delta_T, \gamma_T)$ is obtained. By the PV-inversion equations, we derive the corresponding $(\boldsymbol{u}, h)$ variables and start a new iteration with this flow.

b) If the criterion is satisfied, the optimally balanced flow $(\boldsymbol{u}_T, h_T)$ is reached.

On the other hand, when base point h is chosen, the transformations and the PV-inversion equations are completely removed from the algorithm, and then, the stopping criterion is checked for h^* and h_T. The failed criterion leads to setting h_T as h^* while preserving $\boldsymbol{u}_T$, and a new iteration starts with the flow $(\boldsymbol{u}_T, h^*)$. The satisfied criterion ends the nudging scheme with $(\boldsymbol{u}_T, h_T)$.

The propagation of $(\boldsymbol{u}_T, h_T)$ provides the flow $(\boldsymbol{u}', h')$ at $t = t'$ (Step ii)). At the frozen time $t = t'$, $(\boldsymbol{u}', h')$ is rebalanced by optimal balance in the same manner as in the balancing procedure, and the rebalanced flow $(\boldsymbol{u}'_T, h'_T)$ is received (Step iii)). At the end of the algorithm, we diagnostically quantify spontaneously generated imbalances depending on the corresponding ageostrophic components of $(\boldsymbol{u}', h')$ and $(\boldsymbol{u}'_T, h'_T)$, which are (δ', γ') and (δ'_T, γ'_T), by the diagnosed imbalance (5.4.2) (Step iv)). If optimal balance provides a well-balanced flow, the evolution of $(\boldsymbol{u}_T, h_T)$ up to t' remain in the close neighbourhood of Γ_T. Less imbalances are, thus, excited, and the evolved and the rebalanced flow shall not have much difference. In the numerical experiments, we analyse the performance of optimal balance in this computational algorithm.

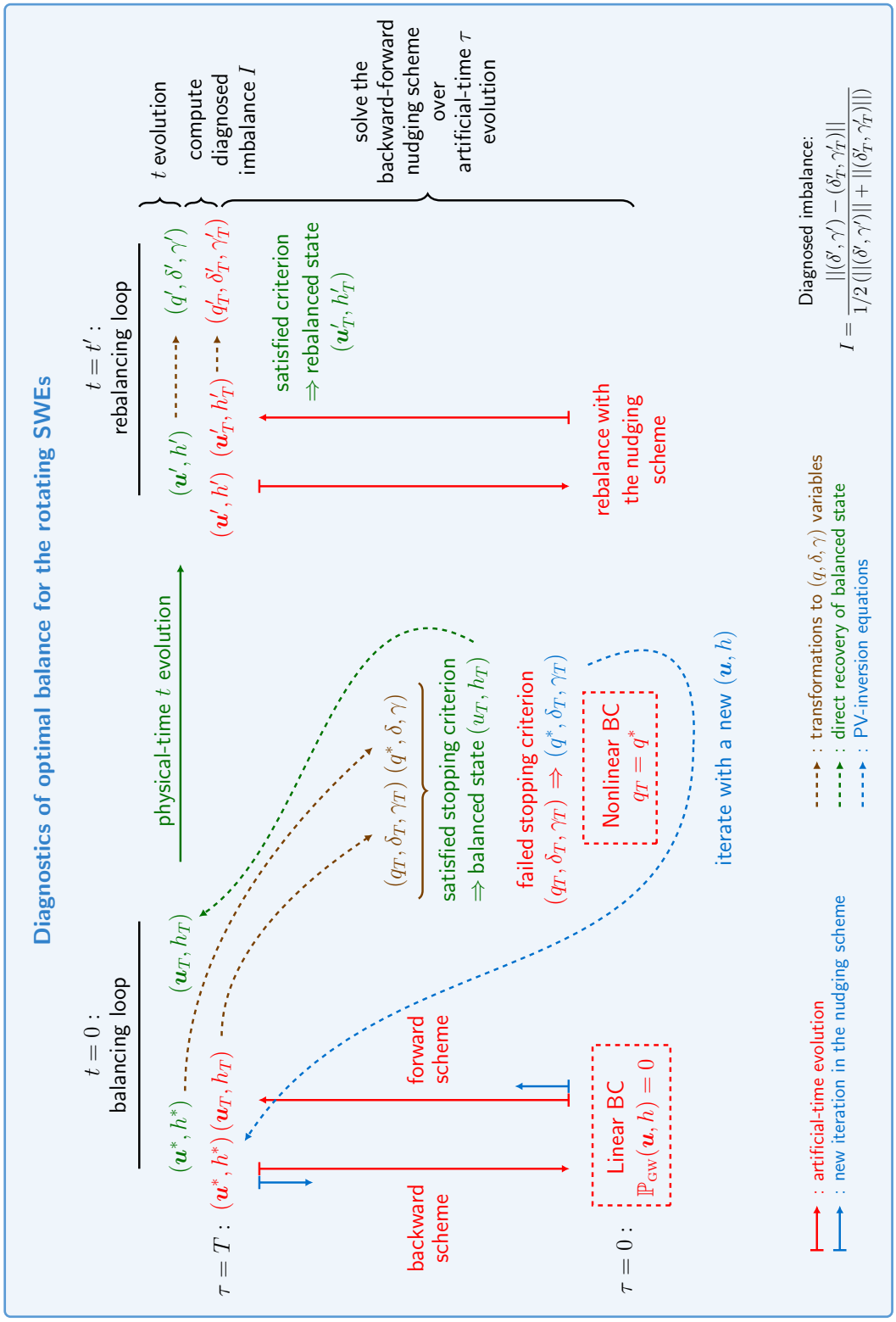

Figure 5.1: Optimal balance applied on the rotating shallow-water equations is studied in a special computational algorithm presented in the schematic form, and the diagnosed imbalance I (5.4.2) is used to determine the quality of optimal balance. The balancing procedure runs at $t = 0$, and the physical-time t evolution up to $t = t'$ is followed by the rebalancing procedure at final time step. The calculation of the diagnosed imbalance I ends the algorithm. The quasi-geostrophic scaling in time variables and base point q are set in the schematics.

Chapter 6

Numerical implementation results

After describing the analytical and experimental set-up, we explore the quality of optimal balance with different test cases and its dependency on several design parameters. First, we visually demonstrate the performance of optimal balance applied on shallow-water flows. For some test cases, we check the quality of balanced states a priori. Although the viscosity term, not eddy viscosity in the context of turbulence closure, is not active in simulations, it is still studied in our setting. We, later on, investigate the convergence of the backward-forward nudging scheme for different combinations of boundary conditions. For further simulations, we set reasonable length of physical and ramp times to use in the diagnostics. With the obtained time lengths, the parameters of the computational algorithm are tested, which are convergence tolerance, linear-end boundary condition, base-point coordinate, and ramp function. We, further, compare the quantification of imbalances at linear end vs. the nonlinear end, and diagnostic error computed by relative error vs. absolute error. Finally, the chapter is closed with the effect of the initial-condition structure on optimal balance.

The base settings of optimal balance are decided, based on our preliminary observations, as follows: Our test cases are initialised at geostrophic balance with random flow fields $(\boldsymbol{u}_{\mathrm{g}}, h)$, explained in Section 5.3. The oblique projector, the preservation of the linear PV q_{lin}, is used as the linear-end boundary condition, and the PV q is chosen as the base-point coordinate. The exponential ramp function, (5.2.2) with $f(\theta) = \exp(-1/\theta)$, is employed. We focus on the wide range of Rossby number, $\varepsilon = 2^{-m/2}$, where $m = 2, \ldots, 11$, and for some specific test trials where ε is fixed, we choose $\varepsilon = 0.1$ within this range. The convergence tolerance κ is fixed, $\kappa = 10^{-4}$ for base point q; it is forced to be increased for base point h concerning the parameter ε values, and in the specific tests, for $\varepsilon = 0.1$, we set $\kappa = 10^{-3}$. We remind that the viscosity is inactive in the scheme. These base settings are kept unless otherwise stated. Each test case is run for both quasi-geostrophic and semi-geostrophic scaling regimes, and we mostly present both results to observe the differences. We, explicitly, refer parts of Appendix A for the results not displayed in this chapter.

6.1 Qualitative analysis

Optimally balanced states of shallow-water flows are visualised starting at two different initial configurations with the same randomly generated h field: a flow near balance, $(\boldsymbol{u}_{\mathrm{g}}, h)$, and a second flow far from balance, $(0, h)$. Both configurations are introduced in

Section 5.3. We apply optimal balance in two ways: i) directly on the initial configurations, ii) on the evolved states of the initial configurations.

In this section, we present some important test results (balanced states) only in the semi-geostrophic regime, and all other results executed in both regimes, mainly in the quasi-geostophic ones, can be found in Appendix A.1. For $\varepsilon = 0.1$, the shallow-water fields show free surface h, mean-free PV $\hat{q} = q - \bar{q}$, velocity divergence δ, and ageostrophic vorticity γ. The ageostrophic δ-γ fields are zero when the test cases are started near balance, and the velocity divergence δ field is zero when they are started with unbalanced initialisation, a flow with zero velocity. Optimal balance employs the ramp time $T = 0.1/\varepsilon$. The base settings are kept the same except the choice of base point, which are explicitly stated.

First for test case i), the direct application on the initial configurations show the effect of base-point choice on balanced states. When base point q is chosen, the nearly balanced and the unbalanced initialisations correspond to different balanced states in characteristics and magnitude, as the initial configurations hold different q fields, see the second column in Figures A.1 and A.2 in the appendix, in order, for the semi-geostrophic regime. When base point h is chosen, both initialisations should correspond to identical balanced states because of holding the same h fields. Due to the issue of convergence and its tolerance, however, the balanced states possess only moderately alike appearance in large scales in both regimes. Although the stopping criterion is also satisfied with a smaller tolerance κ for the considered ε parameter, slow convergence of h in the nudging scheme is the obstacle to improve the balanced states by much.

The choice of base-point coordinate can be recognised by its preserved structure, which is up to some tolerance. For the unbalanced initialisation, the choice is clearly visible due to qualitative differences in the balanced state. For the nearly-balanced initialisation $(\boldsymbol{u}_{\mathrm{g}}, h)$, nevertheless, as the γ field becomes zero, the PV-inversion equation (4.4.3) to recover h field matches with the definition of q in (4.4.1): Using base point q returns the same h, which is used to define q. As a consequence, it gives unsubstantial differences between balanced parts provided by both base points, and for base point h, the convergence issue can be distinguished at the PV $\hat{q}$ field. This test case returns the same observations for the quasi-geostrophic regime, can be seen in Figures A.3 and A.4.

Second for test case ii), we aim to generate more typical flows, so the configurations first evolved up to the respective physical time t' of the scaling regimes; then, optimal balance is applied on the evolved states for the frozen t' value. For $\varepsilon = 0.1$, we present initial fields at $t = 0$, evolved fields at $t = t'$ and balanced parts at $t = t'$. The physical-time length t' is chosen longer to evolve dynamics of the nearly-balanced initial flow in considerable extent. Regarding the chosen $T = 0.1/\varepsilon$ values, we have $t' = 1/\varepsilon$ and $t' = 0.1/\varepsilon$ for the nearly balanced and unbalanced initialisations, respectively.

Starting with the nearly-balanced initialisation produces less imbalances through the physical-time t evolution, see in Figure 6.1. These imbalances are extracted by optimal balance procedure using base points q and h, especially from ageostrophic δ-γ fields, see first two columns in Figure 6.2. The balanced states obtained using both base points keep similar structure, but they have non-zero residual, see the last column in the latter figure.

Starting with the initial flow holding zero velocity gives rise to an evolved flow dominated by higher imbalances, which are especially shock waves, noisy small scale structures, see Figure 6.3. The shock waves are visible in all fields except in the PV q, and these waves are only obtained in the semi-geostrophic regime, for the other regime, see Figure A.7 in

	base point q		base point h	
iteration	$\|\delta_g\|$	$\|\gamma_g\|$	$\|\delta_g\|$	$\|\gamma_g\|$
1	142	2113	142	2113
2	0.032	0.139	185	2575
3	$8 \cdot 10^{-6}$	$3 \cdot 10^{-5}$	181	2580
⋮			⋮	⋮
85			145	2579

Table 6.1: The convergent nudging scheme provides decreasing norm of unbalanced components at the linear end in each iteration. In the table, we present the norm of unbalanced components of δ-γ variables, $\|\delta_g\|$ and $\|\gamma_g\|$, for both base points. The scheme using base point q produces less imbalances and finished in 3 iterations, but on the other hand, the scheme using base point h cannot decrease the excitation of imbalances through 85 iterations.

Appendix A.1. Optimal balance using base point q projects onto the full phase space, which is smooth and small in magnitude, see first column in Figure 6.4. The h field does not have a regular structure, yet optimal balance can use h as base point. In this case, optimal balance does not converge to a balanced state; it only executes a flow state disturbed by shock waves, see the second column in the later figure. We check the convergence of consecutive nudging iterates up to a tolerance to terminate the nudging scheme, therefore h in the last iterate can be notably different than the base point h^*. A thorough analysis is needed to specify proper test results where optimal balance can employ h as base-point coordinate, which is carried out in the upcoming section.

6.2 A priori quality check

The quality of the optimal balance can be detected a priori before receiving balanced states as an outcome. For this examination, we consider the test case initialised with the unbalanced flow in the semi-geostrophic regime, displayed in Figure 6.3. The shown evolved flow is balanced using both base points, and the low quality of balance obtained employing base point h can be distinguished earlier, see second column in Figure 6.4.

In the nudging scheme, for a priori quality check, we consider an additional criterion at the linear end, where the norm of unbalanced components of δ-γ fields is analysed, denoted as $\|\delta_g\|$ and $\|\gamma_g\|$ in Table 6.1. The same analysis can be achieved for the $\boldsymbol{u}$-h variables too. When optimal balance uses base point q, $\|\delta_g\|$ and $\|\gamma_g\|$ decrease quickly at each nudging iteration, and the nudging scheme terminates after the third iteration. When optimal balance uses base point h, we however observe no improvement at these norms despite of higher number of iterations, that is 85 iterations until being stopped owing to generating sufficiently close nudging iterates. We, then, stop the balancing procedure for base point h, if the procedure executes norm-wise non-decreasing sequence of imbalanced components and conclude that applying base point h fails to provide a convergent nudging scheme in the current setting. In our upcoming test cases, flows with shock waves do not appear, so that we do not apply this criterion at the linear end in optimal balance.

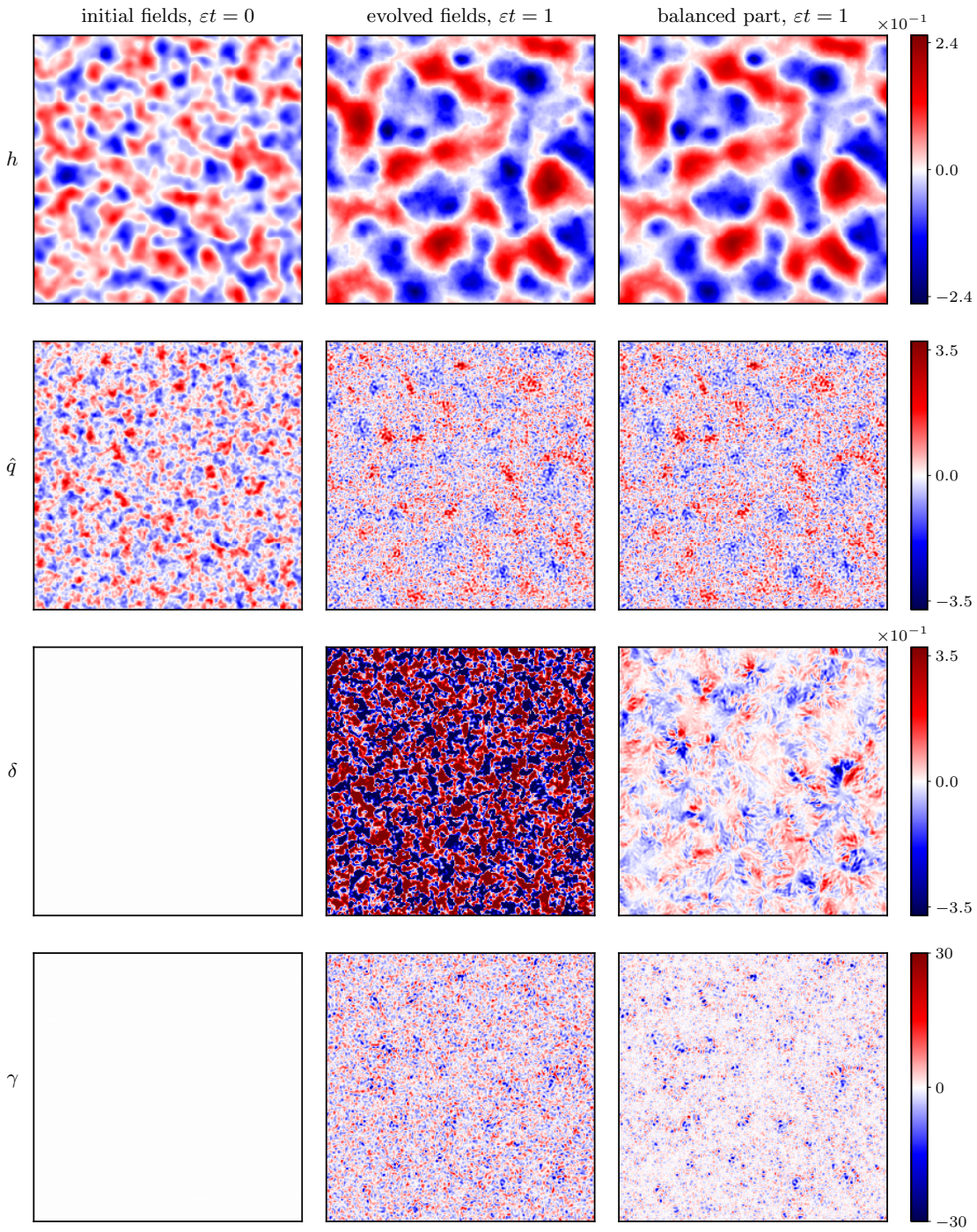

Figure 6.1: The nearly-balanced initial flow is balanced via optimal balance after its evolution while preserving PV q and vanishing imbalances from ageostrophic δ-γ fields in the semi-geostrophic regime. For $\varepsilon = 0.1$, we present the initial shallow-water flow fields in geostrophic balance (left column), the evolved fields (middle column) of the initial flow at $t = 1/\varepsilon$ and the optimally balanced part (right column) of the evolved flow using the oblique projector, base point q, and ramp time $T = 0.1/\varepsilon$.

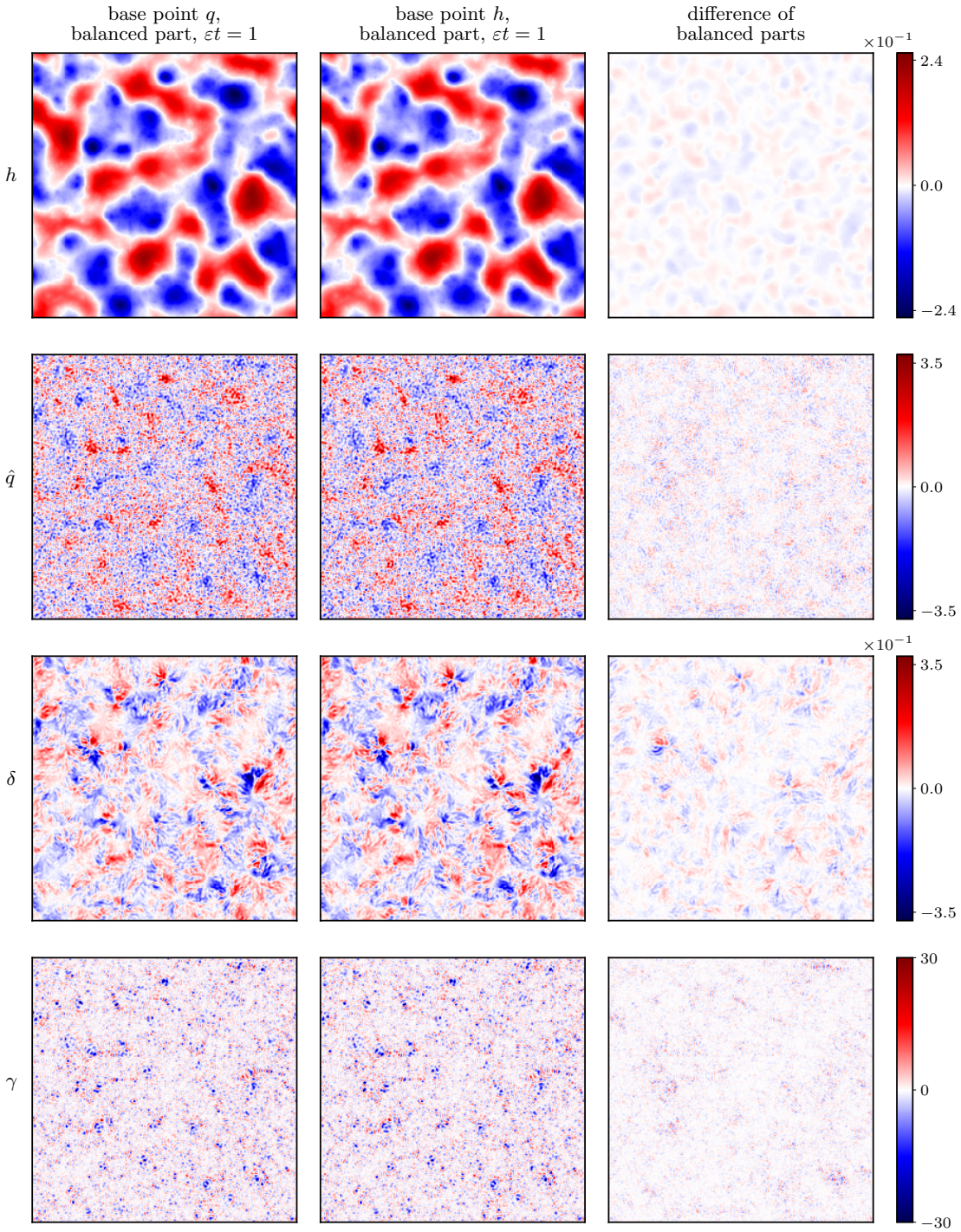

Figure 6.2: The evolved flow in Figure 6.1 is optimally balanced using base point h besides base point q. Both balanced parts show similar pattern, but they relatively differentiate from each other (right column).

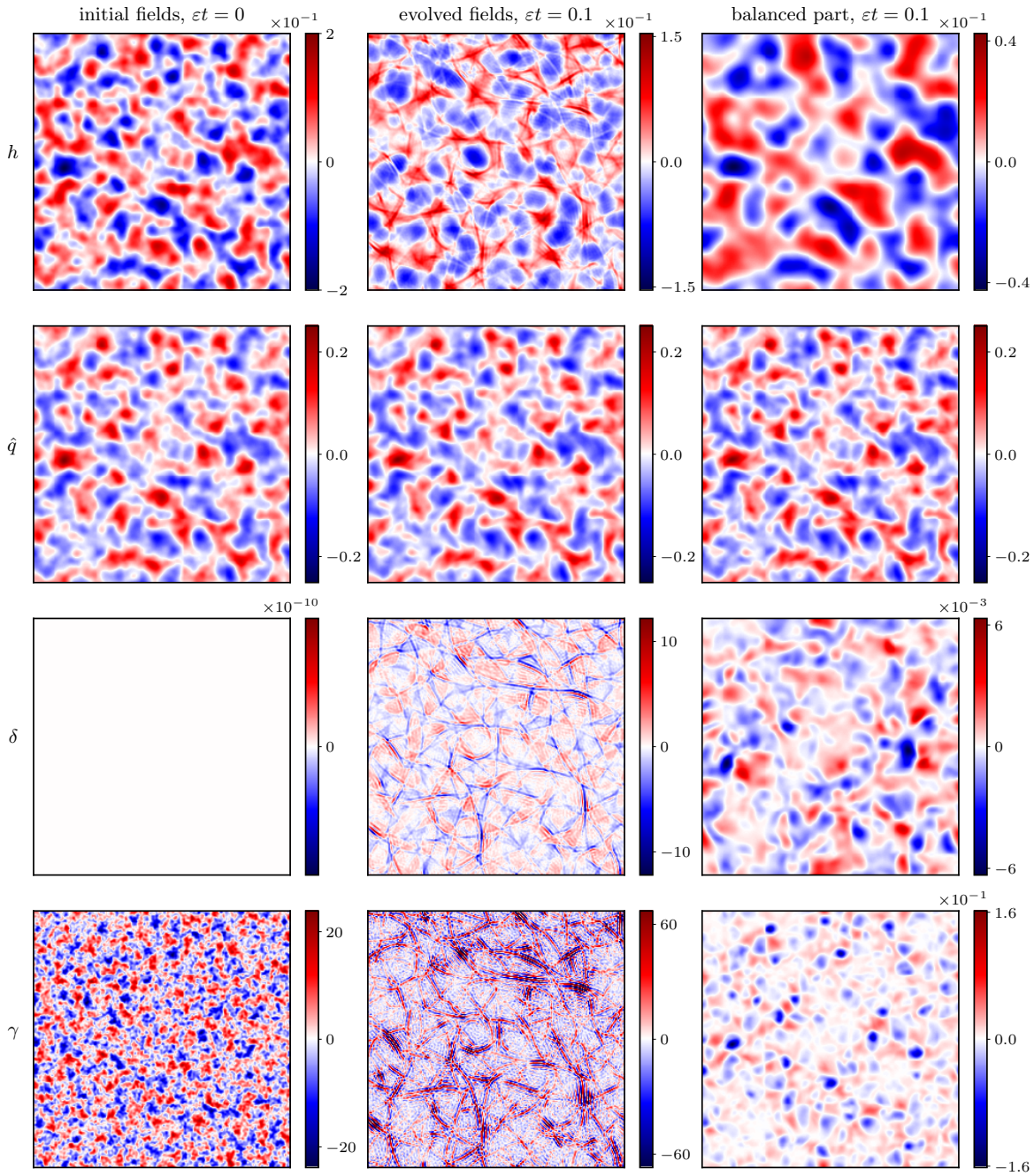

Figure 6.3: With the initial fields holding zero velocity instead of geostrophic balance, we carried out the same test case as in Figure 6.1 for shorter physical-time length $t = 0.1/\varepsilon$ in the semi-geostrophic regime. The change in the initial fields results in unphysical structures in the evolved fields; however, optimal balance smooths these imbalances fixing base-point coordinate q.

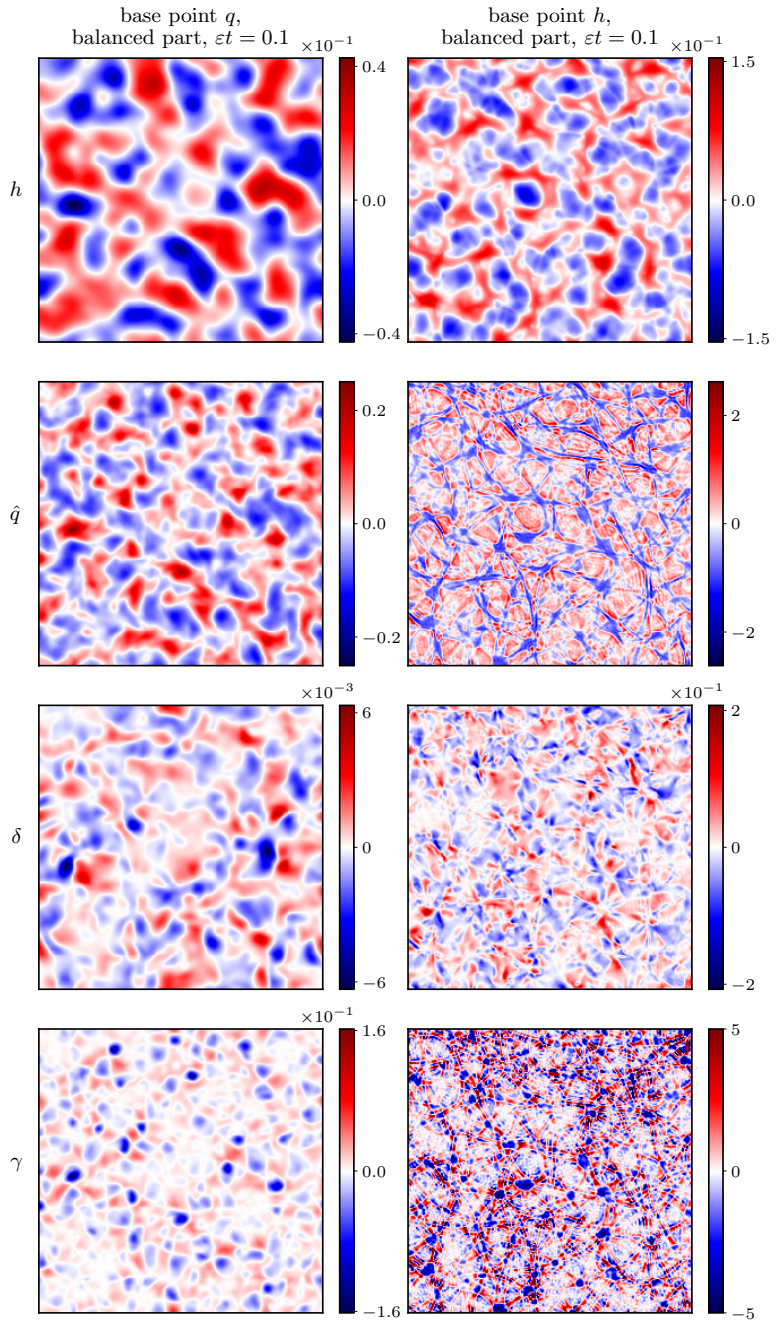

Figure 6.4: Optimal balance employing base point h is also applied on the evolved fields in Figure 6.3, and the nudging scheme preserving h does not converge to a balanced state. The obtained state disturbed by shock waves is indeed not a balanced state, albeit the backward-forward nudging scheme is stopped by the criterion at the nonlinear end.

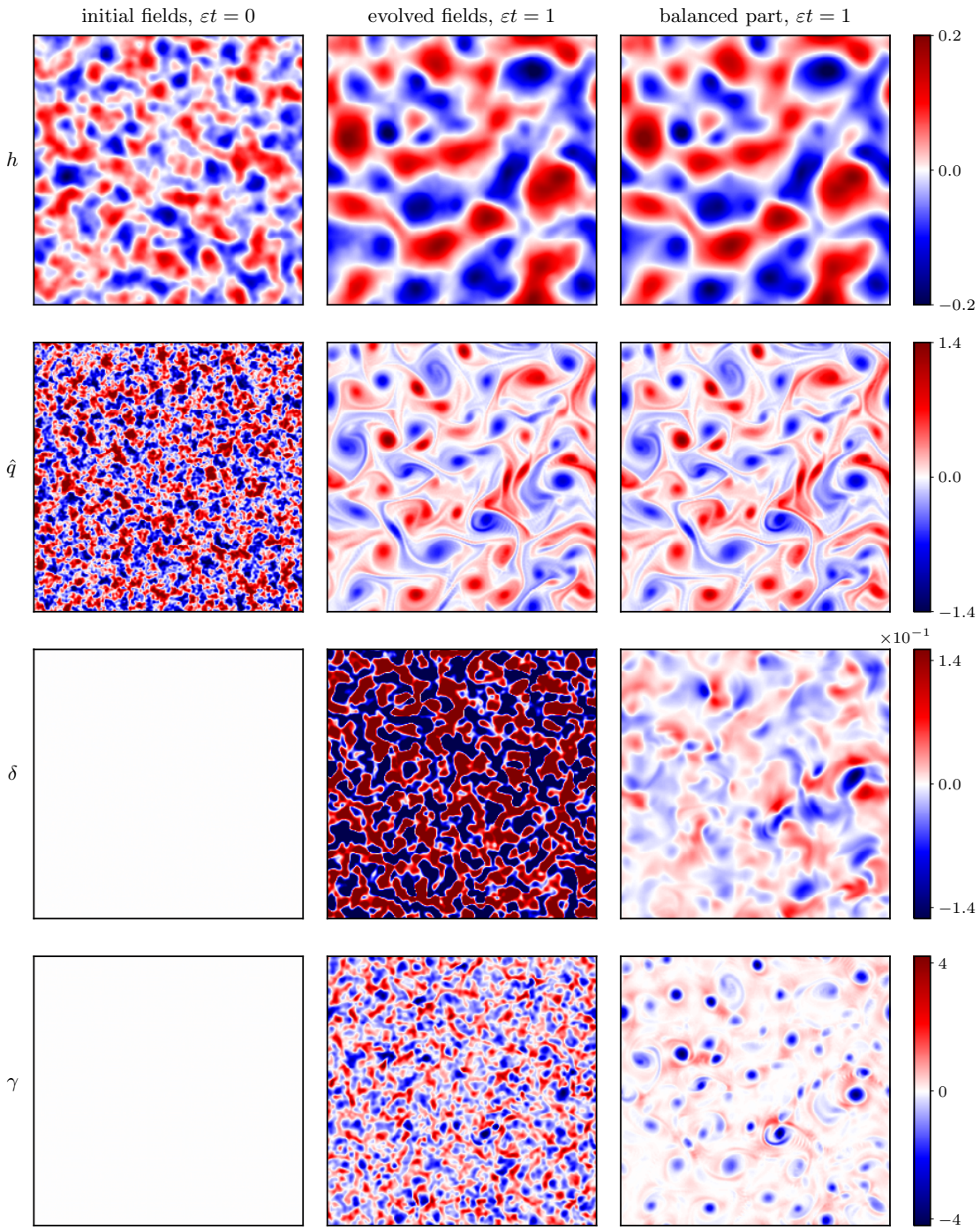

Figure 6.5: The viscosity term dissipates energy at small scales, so flow fields have smooth pattern being significantly void of small-scale structures. The figure displays the effect of the viscosity term with $\mathrm{Re} = 3 \times 10^3$ in the semi-geostrophic regime. The test case is the same as in Figure 6.1, which is run without the viscosity term.

6.3 Viscosity

Dissipation, and so the viscosity term, is not a crucial concept in our numerical model, as the artificial-time and the physical-time lengths are of order one or less than eddy turnover time in test cases, where we analyse the performance of optimal balance in the computational algorithm. The viscosity term, thus, remains deactivated in test cases; still, we investigate a possible application of viscosity in the artificial-time τ and the physical-time t integrations, simultaneously.

By running a long-time simulation of shallow-water model without using optimal balance, the viscosity coefficient, Reynolds number in (5.1.1), is determined. We discovered $\mathrm{Re} = 3 \times 10^3$, which is large enough to prevent pile-up of spectral energy in higher wave numbers near the resolution scale. When the viscosity term with the current Re value is applied in both integrations dissipatively, details in Section 5.1, energy decreases in all evolved fields, where the effect is mainly seen on small-scale energy and, as a result, on balanced states.

To analyse the effect of viscosity, we rerun the test case in Figure 6.1, that is in the semi-geostrophic regime, using the viscosity term. As the t-integration length is chosen longer than one eddy turnover time in the qualitative analysis, the considerable amount of large-scale energy has decreased, see in Figure 6.5. The test case in the quasi-geostrophic regime can be seen in Figure A.9 in the appendix. By this result, we justified the use of viscosity in optimal balance, when the notion of dissipation is unavoidable to stabilise numerical models.

6.4 Convergence of the nudging scheme

The convergence of the backward-forward nudging scheme is analysed in this section by plotting energy spectra of the nudging iterates for different combinations of linear-end and nonlinear-end boundary conditions. The nudging scheme is, here, run without being terminated by the stopping criterion. The iterates are selected among the first 40 iterations, but if the manner of convergence is unclear, we consider up to 300 iterations. In our other test cases, in the case of convergent nudging scheme, the balancing procedure is terminated before reaching the last iterate.

The parameters are in our base configuration: The scheme is initialised near balance and uses the exponential ramp function. Both geostrophic scaling regimes are tested with their respective ramp times $T = 1$ and $T = 0.1/\varepsilon$ for three different ε values, $\varepsilon = 0.5, 0.1, 0.03$. We compare the four choices of the linear-end boundary condition: the oblique, orthogonal, h-preserving and ζ-preserving projectors, all have been introduced in Section 4.3. Our observation on convergence indicates both scaling regimes unless there exist different behaviour. We simply refer the row numbers in the related figures: Figures 6.6 and 6.8 when optimal balance uses base point q; Figures 6.7 and 6.9, when optimal balance uses base point h. The quasi-geostrophic regime is presented in Figures 6.6 and 6.7; the semi-geostrophic one is in Figures 6.8 and 6.9.

First, preserving h (first rows in the figures) at the linear end diverges immediately in both regimes, when h is applied as base point at the nonlinear end. We present until the last iterations before the scheme is completely disturbed by noise, see all other flow fields in the semi-geostrophic regime in Appendix A.3. When q is taken as base point, for $\varepsilon = 0.5$, the scheme diverges again after the first iteration. For smaller ε values, nevertheless, the

Quasi-geostrophic regime

	base point q		base point h	
linear-end boundaries	small ε	large ε	small ε	large ε
preserve h	✓			
preserve ζ	✓	✓		
oblique projector	✓	✓	✓	
orthogonal projector	✓	✓	✓	

Table 6.2: In the quasi-geostrophic regime, the oblique and orthogonal projectors coincide with each other and give convergent iterates except for large ε values, when base point h is applied in the nudging scheme. The table shows the behaviour of the nudging scheme using different combinations of the linear-end boundary and the nonlinear-end boundary (base points) conditions. The (✓) and (✓) symbols stand for the combinations resulting in convergent schemes, which are displayed in Figures 6.6 and 6.7; however, we approve only the combinations indicated with the (✓) symbol, and we reject the combinations indicated with the (✓) symbol due to slower rate of convergence. If a combination holds no symbol, then the combination causes a divergent scheme.

nudging iterates converge to balanced states, where the spatial spectra are very close to the lowest energy level but at a slower rate relative to other linear boundary conditions. Given that the h-preserving projector works only with base point q and given its slow convergence at smaller parameter range are enough to reject this projector as the linear-end boundary condition.

Second, preserving ζ (second rows) is suggested as an alternative to preserving h. The ζ-preserving projector alongside base point q gives a convergent nudging scheme. The balanced states are reached in a few iterations, but they have higher energy level than those of obtained by the oblique and orthogonal projectors, which can be recognised better for smaller ε values. When base point is switched to h, though, using preserving ζ provides diverging iterates, which is delicate to notice at early iterations. The energy in base point h is not properly damped, and especially in large scales, the projector fails to act on some specific wave numbers. In the semi-geostrophic regime, the effect of this failure on other fields can be seen $\varepsilon = 0.1$ in Appendix A.3. The ζ-preserving projector is, hence, another poor choice because of its divergence when base point h is applied and the higher energy level of its converged states when base point q is applied.

Third, the spectral oblique and orthogonal projectors (third and fourth rows, in order) have different features, which fundamentally arise from the angle between the Rossby-wave and gravity-wave subspace as a result different performance in the nudging scheme. In the quasi-geostrophic regime, these subspaces are orthogonal, so the projectors are identical giving the same convergence speed; in the semi-geostrophic regime, they are non-orthogonal, so the behaviour of the projectors changes under the effect of ε and wave numbers. As a general overview in both regimes, smaller ε values excite less imbalances as expected, and the fast convergence is notable relative to larger ε values for the nudging schemes converging balanced states.

In the semi-geostrophic regime, for $\varepsilon \approx 1$, the subspaces are close to orthogonality, especially more orthogonal at large wave numbers. The oblique projector behaves as a rough orthogonal projector. Independent of the base-point choice, consequently, qualitatively close balanced states are produced, and for $\varepsilon = 0.5$, each nudging iteration piles

Semi-geostrophic regime

	base point q		base point h	
linear-end boundaries	small ε	large ε	small ε	large ε
preserve h	✓			
preserve ζ	✓	✓		
oblique projector	✓	✓	✓	
orthogonal projector	✓	✓		

Table 6.3: In the semi-geostrophic regime, the oblique projector provides convergent nudging scheme except when base point h is employed with large ε values. The table is prepared in the same format of Table 6.2: the (✓) and (✓) symbols for the combinations providing convergent schemes, which can be seen in Figures 6.8 and 6.9, and the (✓) symbol for the accepted combinations.

up more energy at small scales. For $\varepsilon \ll 1$, more oblique subspaces are obtained, and the angle becomes smaller particularly at small wave numbers. The orthogonal projector provides Rossby-wave components with less energy at the linear end, and the rate of the energy decreases while ε getting smaller. Choosing q as base point balances the energy changes in the Rossby components, and yet, the orthogonal projector converge at a slower rate than the oblique one. Base point h, nevertheless, is unsuccessful to maintain regularity. The orthogonal projector cannot progressively reduce the amplitude of the spectral energy at each iterate, and this leads to straight divergence at large scales. The same type of divergence is also observed for the ζ-preserving projector, although the underlying reason might be different. The oblique projector, however, generates convergent iterates with slower rate of convergence, which come down to almost zero at small scales, in comparison with the case applying base point q. The last observation is also obtained in the quasi-geostrophic regime. Given the divergence in the current setting, the orthogonal projector is unacceptable as a linear-end boundary. On the other hand, the weakness of the oblique projector arises only when base point h is used with larger ε values.

Our work on all combinations of the linear-end and nonlinear-end boundary conditions for different ε values is summarised in Tables 6.2 and 6.3 for the related regimes. The diverging nudging schemes occur as a result of i) using base point q with the h-preserving projector for large ε values in both regimes and ii) using base point h with the linear-end conditions except the oblique and orthogonal projectors in the quasi-geostrophic regime and the oblique projector in the semi-geostrophic regime for small ε values. The oblique projector is, hence, the most robust choice for the linear-end boundary condition providing convergence, and base point q as the nonlinear-end boundary condition makes this convergence stronger. Besides the robust convergence, the oblique projector is recovered as the linear PV q_{lin}-preserving projector formulated by a PDE–based approach in Section 4.3.2. Using the oblique projector and base point q together also maintain the existing PV at both temporal-end points. On the other hand, we shall not discard directly the choice of base point h and shall analyse it further in the diagnosed imbalanced, as base point h returns reasonable decrease of energy in large scales while optimal balance employs the oblique projector for smaller ε values.

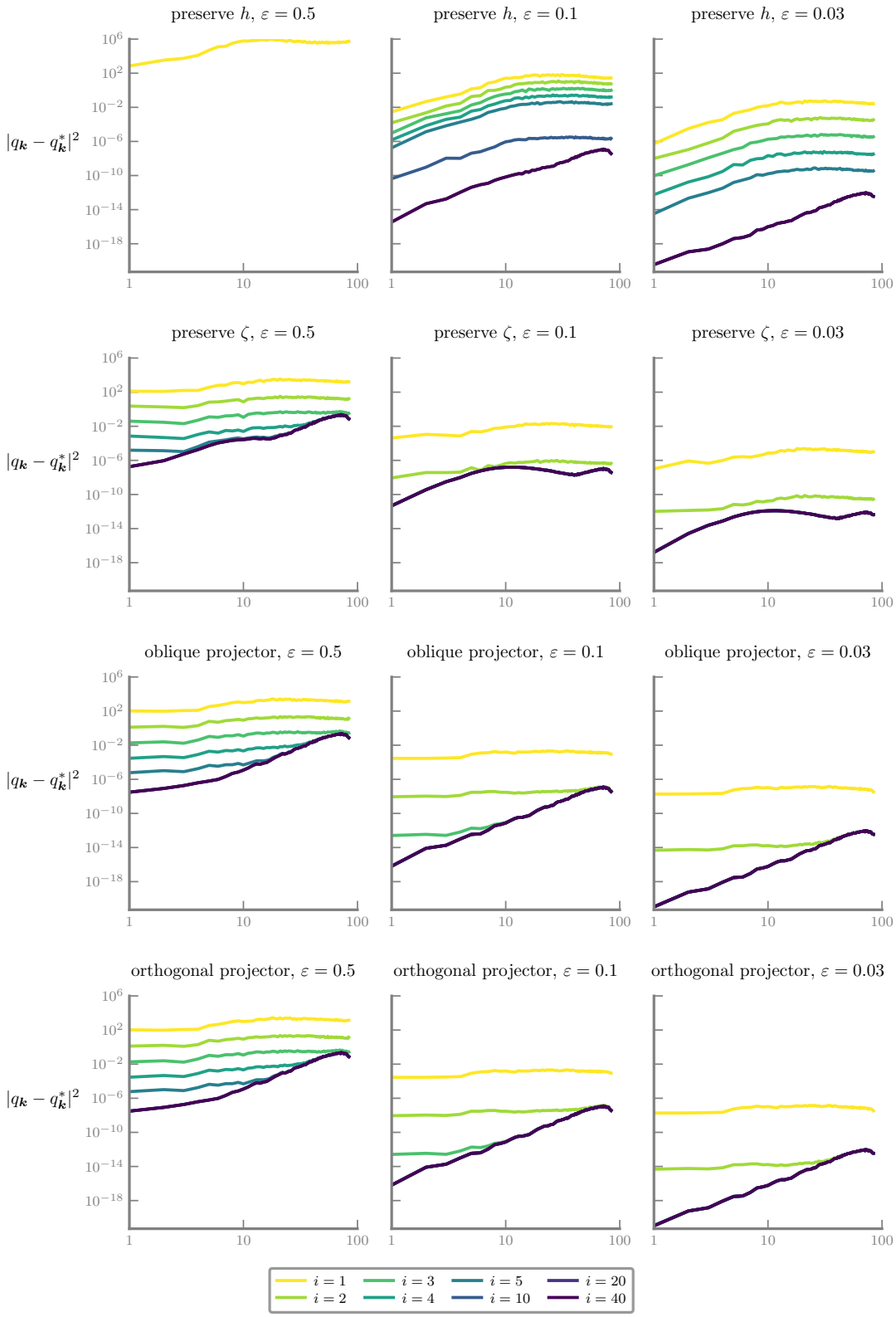

Figure 6.6: The convergence speed of the backward-forward nudging scheme is affected by the linear-end boundary conditions and the Rossby number ε values in the quasi-geostrophic regime. Energy spectra of selected nudging iterates are presented for four linear-end boundaries and three ε values. Optimal balance initialised near balance uses the base point q, the ramp time $T = 1$, and the exponential ramp function.

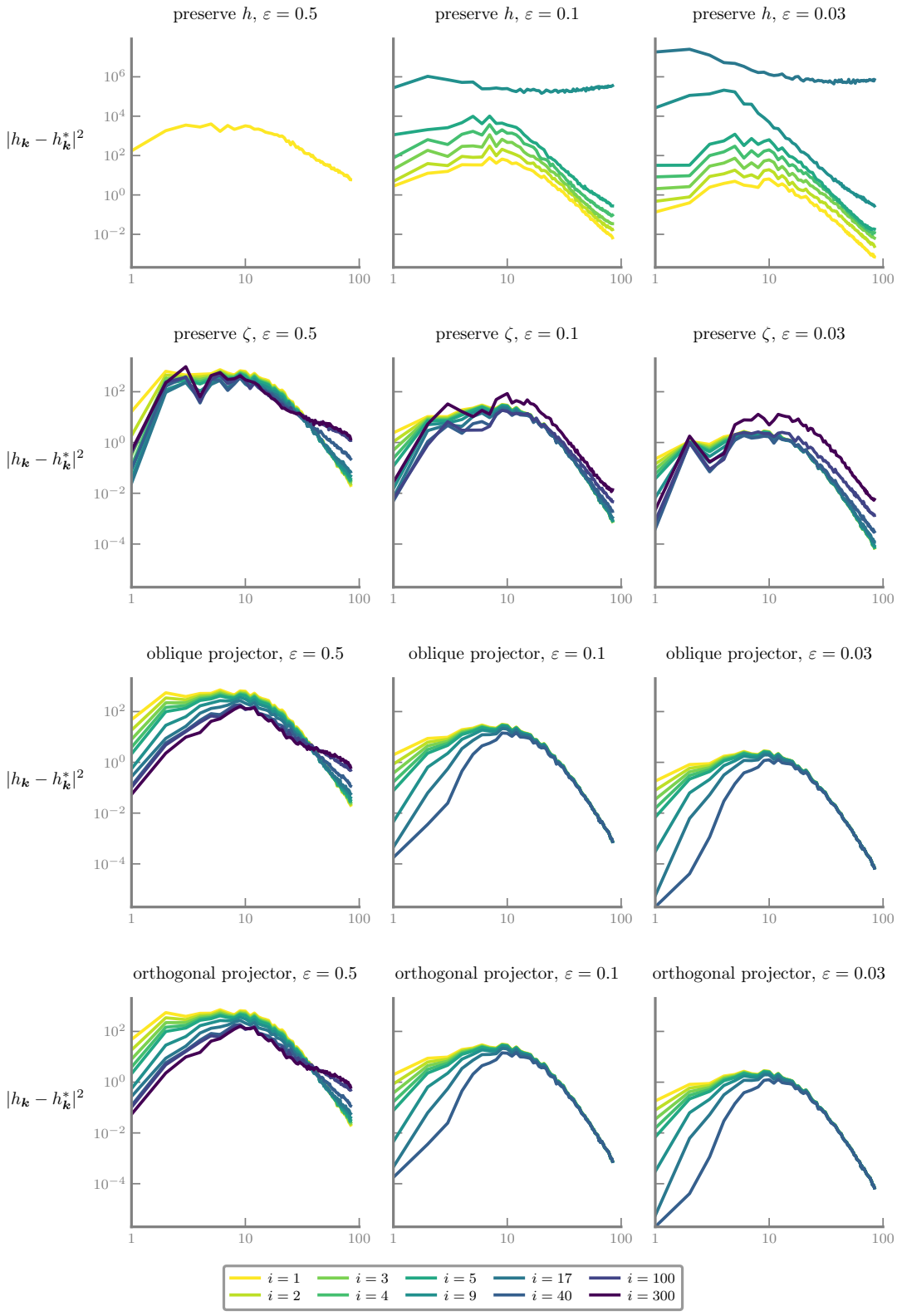

Figure 6.7: The base point h slows down the convergence speed of the nudging scheme for all linear-end boundary conditions over all ε values. Due to resulting into divergent schemes, the linear-end conditions preserving h and ζ are excluded. The energy spectra as in Figure 6.6 is executed for base point h with the same settings.

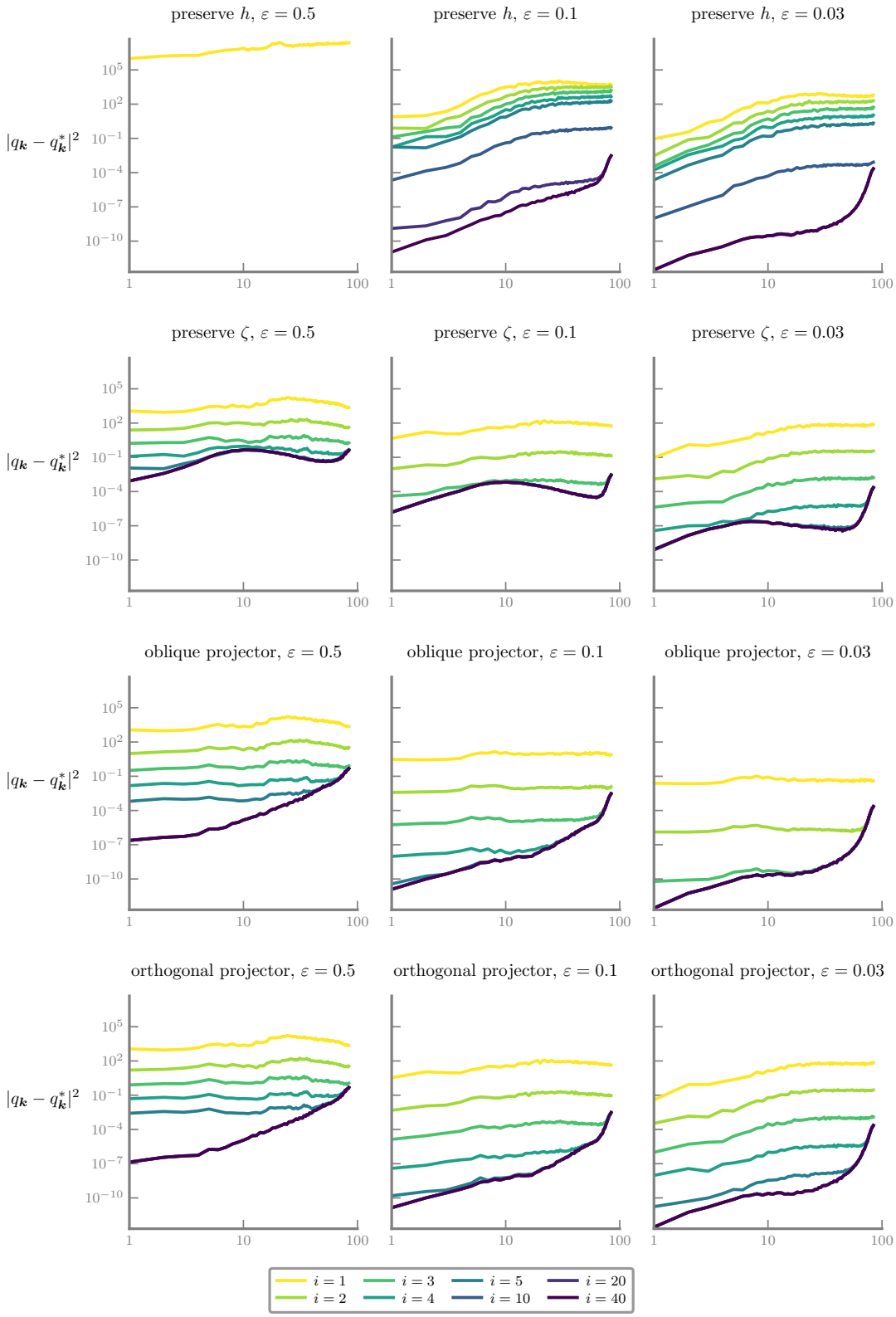

Figure 6.8: The nudging scheme using base point q has convergent iterates for all linear-end boundary conditions in the semi-geostrophic regime, except the h-preserving projector for the larger ε-range values, like the case in the quasi-geostrophic regime. The test case in Figure 6.6 is, here, performed in the semi-geostrophic regime. All other settings are kept the same, but the ramp time is $T = 1/\varepsilon$.

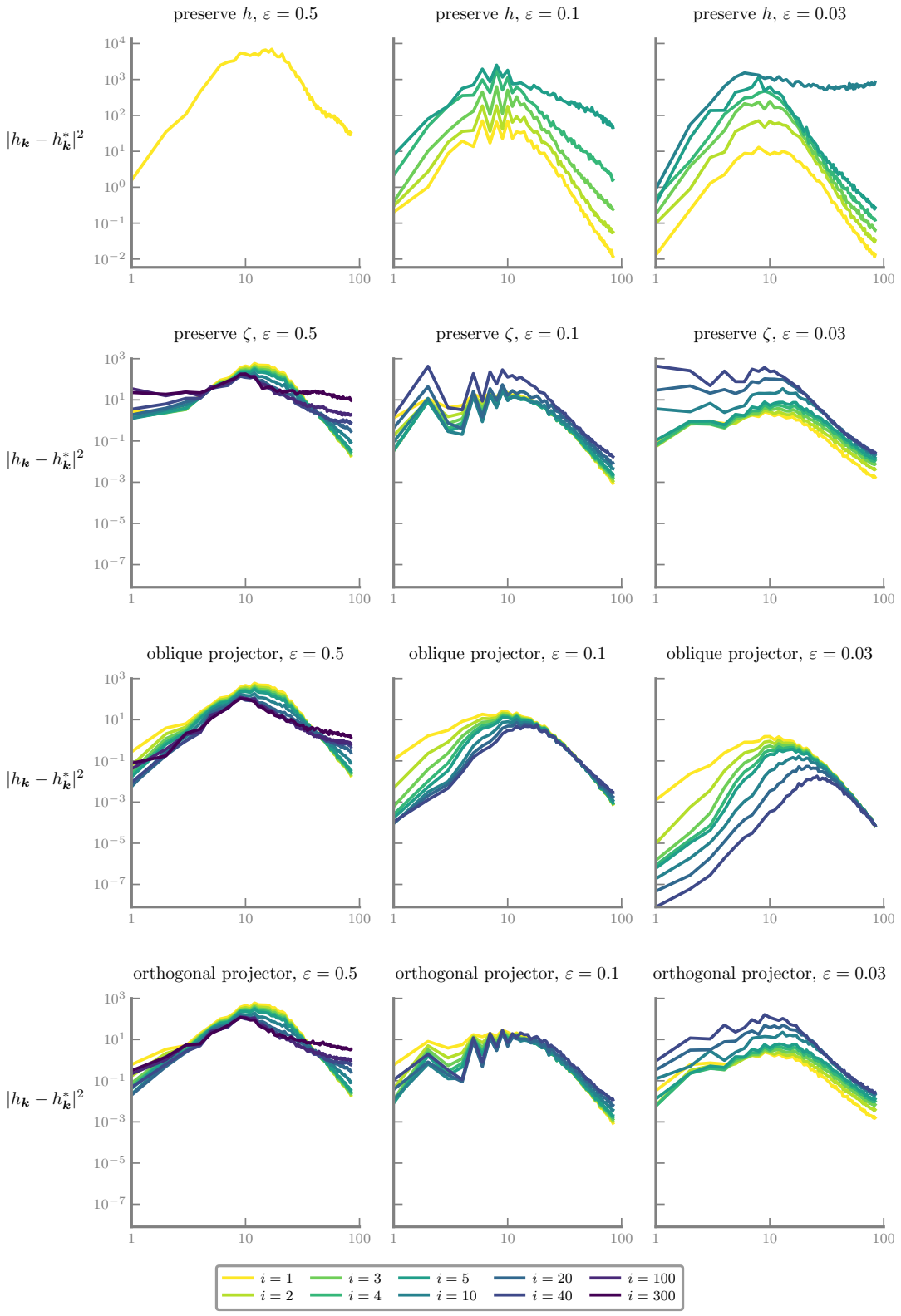

Figure 6.9: The nudging iterates of the ζ-preserving and the orthogonal projectors diverge for smaller ε values in the semi-geostrophic regime being different than the result of the quasi-geostrophic one as in Figure 6.7. The oblique projector shows clear divergence with energy pile-up not only at small scales but also at large scales for $\varepsilon = 0.5$. The test case in Figure 6.8 is, here, executed for base point h.

6.5 Optimal integration time scales

The lengths of ramp-time τ and physical-time t in the computational algorithm in Section 5.6 are important to provide reliable quantitative analysis of optimal balance. To determine good choices, we tested diagnosed imbalance as a function of either ramp-time length T or physical-time length t' for a fixed $\varepsilon = 0.1$ in both scaling regimes. All results of diagnosed imbalance, which are presented here or in the forthcoming sections, are plotted on a logarithmic scale for both axes. As earlier, we preserve the base settings of optimal balance using the oblique projector, base point q and the exponential ramp function.

The range of ramp-time length T are decided from Figure 6.10 and Figure A.12 in the appendix for the quasi-geostrophic and semi-geostrophic regimes, respectively. The sensitivity of the diagnosed imbalance to different t' values become distinguishable if T is long enough. After the least imbalance is diagnosed, where the algorithm takes the best ramp time, more imbalances are produced longer the T values gets, and the sensitivity to t' decreases. Based on these results, the computational algorithm can be studied for the T values, that are smaller than or equal to the best ramp time: $T = 0.5, \ldots, 2$ for the quasi-geostrophic regime and $\varepsilon T = 0.05, \ldots, 0.2$ for the semi-geostrophic regime.

The physical time t' values are selected by studying the diagnosed imbalance as a function of t', see Figure 6.11 and Figure A.13 in the appendix. For different T values, as expected, shorter physical-time integration produces comparable amount of imbalances. After the t integration becomes long enough to excite different amount of imbalances; however, longer it is more the diagnosed imbalance is. To select the t' lengths, we focus on the range, where the diagnosed imbalance is nearly insensitive to slight changes in t', and the reasonable choices are, in order, $t' = 0.5$ and $t' = 0.05/\varepsilon$ for the quasi-geostrophic and the semi-geostrophic regimes.

To examine the effect of ramp time T choice further, we execute the diagnosed imbalance across the range of ε, $\varepsilon = 2^{-m/2}$ with $m = 2, \ldots, 11$, for three different ramp times T in both scaling regimes, see Figures 6.12 and 6.13. Longer ramp times are advantageous in optimal balance, and flows that are initially well balanced in the balancing procedure stay near an approximate slow manifold through the t evolution, i.e, less imbalances are produced. The advantage is visible across the whole range of ε, but for the small ε-range, optimal balance reaches physical limits due to numerical limitations, especially for the semi-geostrophic regime. In our upcoming test cases, we display only ramp times $T = 1$ and $T = 0.1/\varepsilon$ for the corresponding scaling regimes. Test cases can be, still, successfully carried out for the determined range of ramp times, without any convergence problem.

Intentionally left blank.

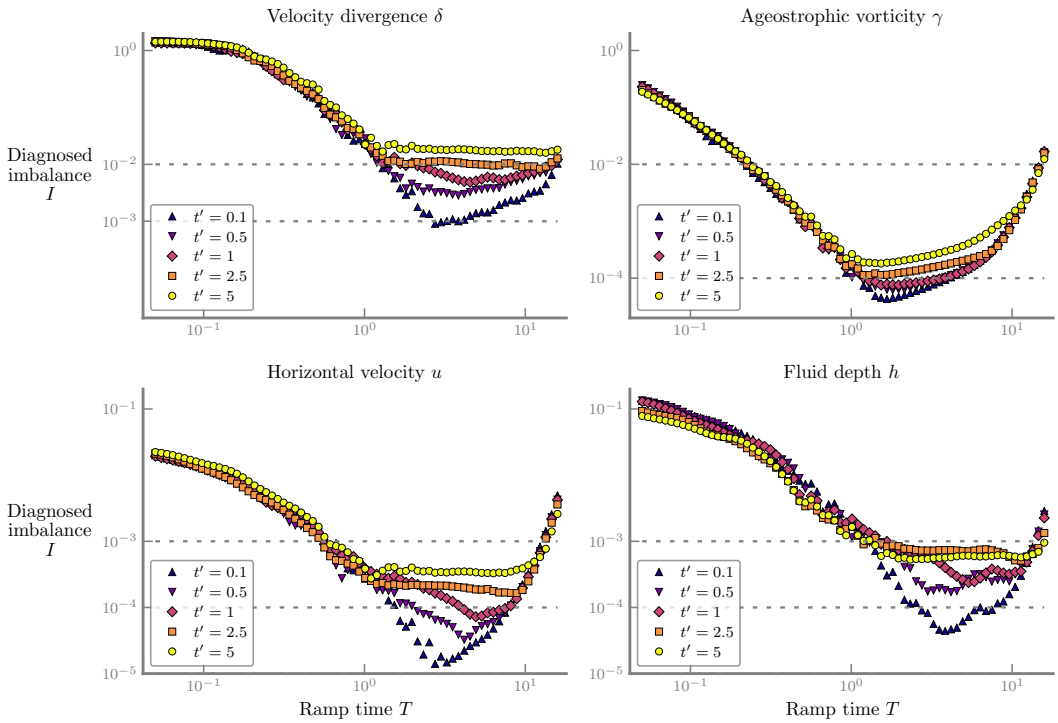

Figure 6.10: Diagnosed imbalance I has a particular range of ramp time to study diagnostic analysis. The figure presents the diagnosed imbalance as a function of ramp-time length T, $I(T)$, in the quasi-geostrophic regime for $\varepsilon = 0.1$. Optimal balance uses the oblique projector, base point q and the exponential ramp.

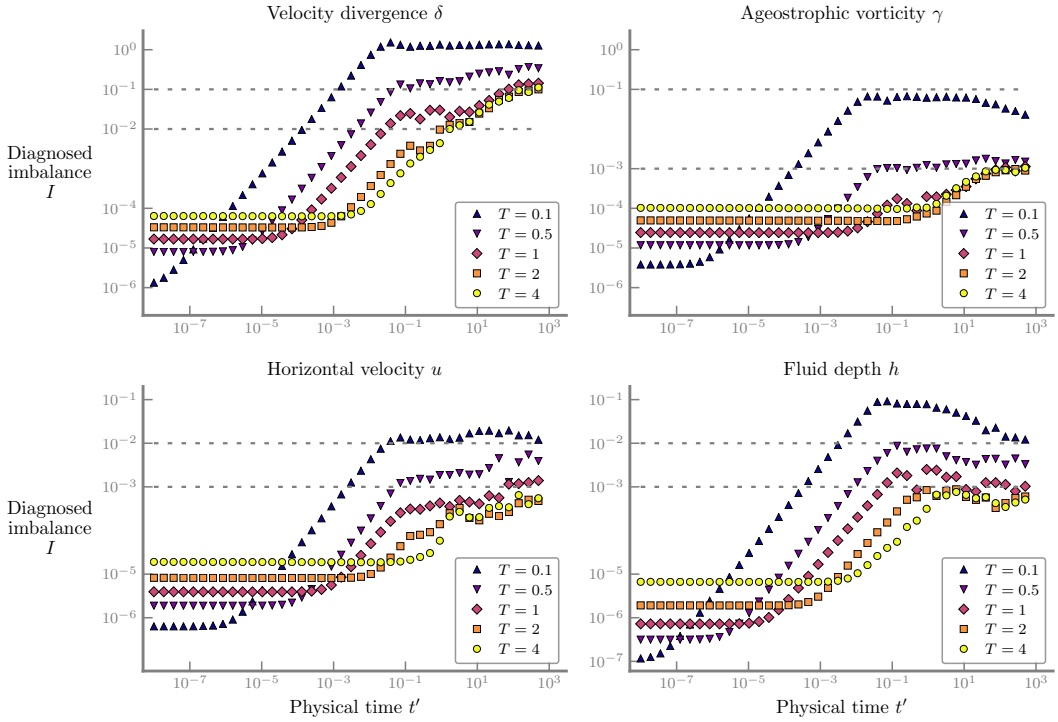

Figure 6.11: Diagnostic analysis also requires appropriate physical-time lengths t'. We display $I(t')$ in the quasi-geostrophic regime for $\varepsilon = 0.1$. The same optimal balance settings as in Figure 6.10 are maintained.

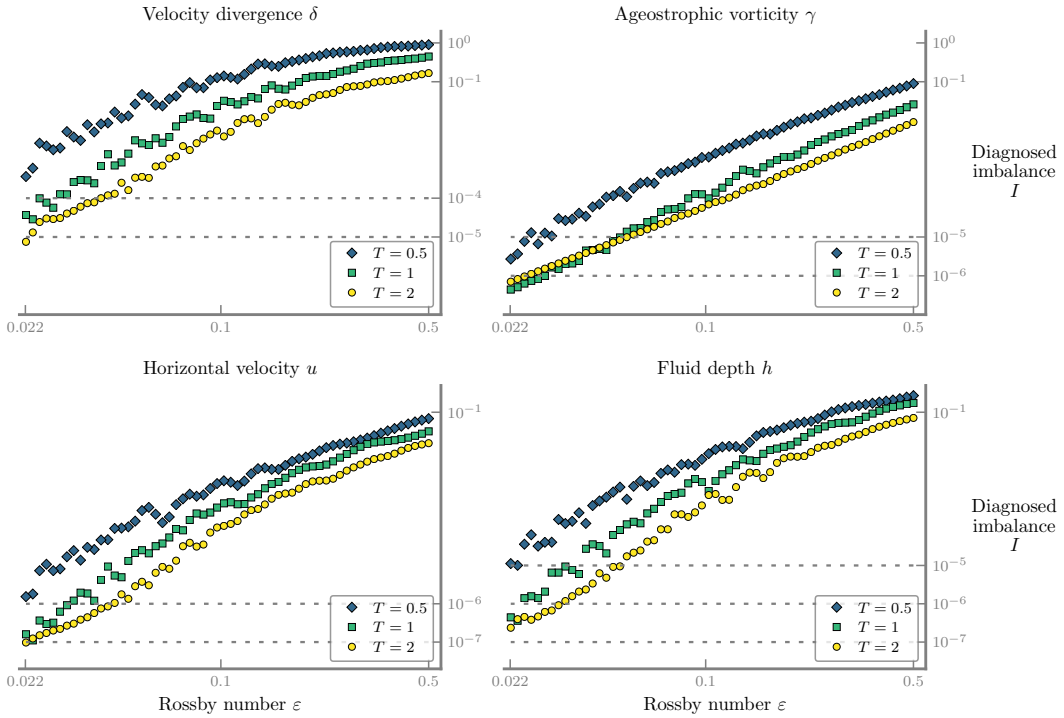

Figure 6.12: Longer ramp time T provides smaller diagnosed imbalances in the quasi-geostrophic regime. The figure shows the diagnosed imbalances I as a function of ε for different ramp times. The base settings are used with the physical-time length $t' = 0.5$.

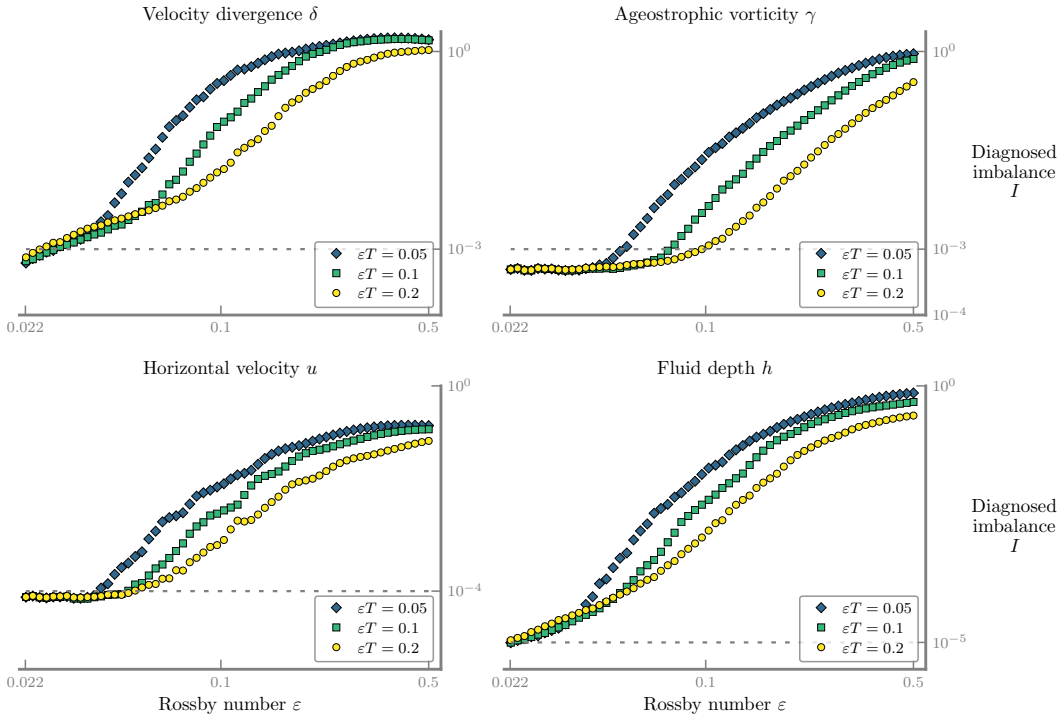

Figure 6.13: The same observation as in Figure 6.12 is received in the semi-geostrophic regime. The test is run with the same settings except the artificial-time length indicated in the figure and the physical-time length $t' = 0.05/\varepsilon$.

6.6 Systematic exploration of the algorithm parameters

The quality of optimal balance is explored depending on the choice of design parameters: convergence tolerance, linear-end boundary condition, base-point coordinate and ramp function. We, then, discover the "best" option in terms of the excitation of less imbalances. The interested range of Rossby number stays the same, $\varepsilon = 2^{-m/2}$ with $m = 2, \ldots, 11$. If not specifically stated, the parameter comparison is mainly carried out with the combination of the oblique projector at the linear end and base point q at the nonlinear end, when test cases start with a geostrophically balanced flow.

6.6.1 Convergence tolerance

In the stopping criterion (5.5.1), the convergence tolerance κ is decided as $\kappa = 10^{-4}$, when base point q is used. When the κ value is decreased to $\kappa = 10^{-8}$, we get no improvement in the quality of balance despite of more iterations, since the former tolerance is enough to provide sufficient convergence to base point q and to lead a balanced state. The diagnosed imbalance is, hence, κ-insensitive, see Figures A.14 and A.15 in the appendix. The κ-sensitivity test, on the other hand, cannot be applied to base point h due to the limitation on the speed of convergence. The different κ values are chosen across ε values in a way that no further convergence is possible: $\kappa = 0.02$ for $\varepsilon \approx 1$ and decreasing to $\kappa = 10^{-4}$ for $\varepsilon \ll 1$. We can also report the effect of ε values on the speed of convergence for base point q, but it is weak to limit the κ value.

6.6.2 Linear-end boundary condition

The linear-end boundary condition is another parameter be explored. When q is chosen as base point, the ζ-preserving, oblique and orthogonal projectors provide a converging scheme over the determined ε-range, while the h-preserving projector yields a diverging scheme for large ε values. Even if some projectors are rejected to be used as the linear-end boundary, see in Section 6.4, we analyse them for convergent cases.

Using different linear-end conditions impacts the characteristics of the computational algorithm. The balancing procedure provides different balanced states, which are shown in Section 6.4, but the same amount of imbalances are diagnosed, see Figures 6.15 and 6.16 for the respective regimes. We choose the longest ramp times, $T = 2$ and $T = 0.2/\varepsilon$, to observe the characteristics, when optimal balance uses the spectral projectors, which are the oblique and the orthogonal ones, and the ζ-preserving projector. In the semi-geostrophic regime, we recall that small ε values characterise the angle between Rossby-wave and gravity-wave subspaces as far from being orthogonal. The orthogonal projector, therefore, requires more iterations in the nudging scheme for initial balancing, right figure in Figure 6.14. The increase in the number of iterations is also visible in the ζ-preserving projector. In the quasi-geostrophic regime, nevertheless, the oblique and the orthogonal projectors requires the same number of iterations because of being identical, shown under the label "spectral projectors" in the left figure in Figure 6.14. We again observe that the ζ-preserving projector needs more iterations while ε gets smaller.

In the context of diagnosed imbalance, the h-preserving projector is studied within the different ε range, $\varepsilon = 2^{-m/2}$ with $m = 6, \ldots, 11$, due to the convergence issue of the projector for larger ε values. The scaling of the diagnosed imbalance is qualitatively similar to those of other projectors, see Figures A.16 and A.17 in the appendix. By these

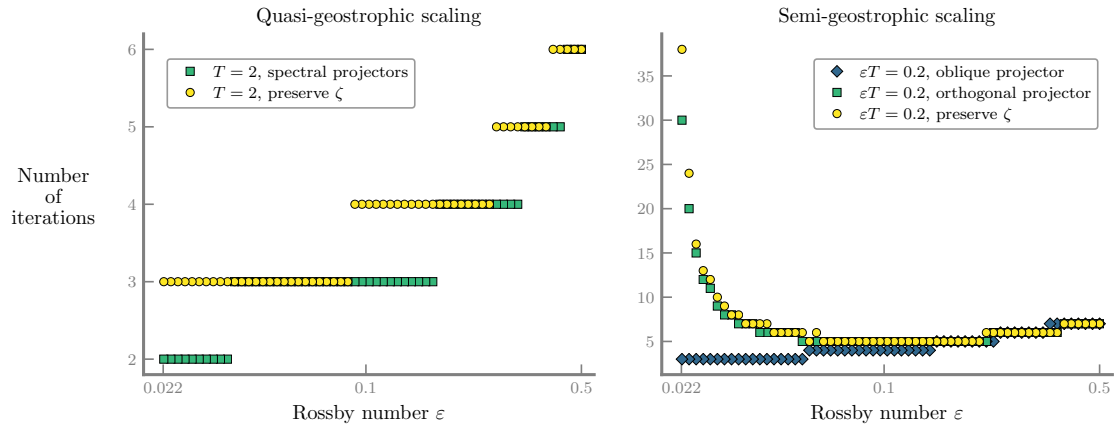

Figure 6.14: The linear-end boundary conditions conclude the nudging scheme with different characteristics. The figure shows the number of iterations in the quasi-geostrophic regime (left) and in the semi-geostrophic regime (right) for the linear-end boundary conditions: the spectral projectors, which are the oblique and the orthogonal projectors, and the ζ-preserving projector. Optimal balance uses $T = 2$ and $t' = 0.5$ for the quasi-geostrophic regime, and $T = 0.2/\varepsilon$ and $t' = 0.05/\varepsilon$ for the semi-geostrophic regime.

results, the diagnosed imbalance is emphasised as being independent of the choice of the linear-end boundary condition.

6.6.3 Base-point coordinate

In this section, we explore the response of optimal balance to the choice of base point. The quality of balanced states obtained using both base points are quantitatively close to each other in terms of the diagnosed imbalance except using base point q comes with slight better results in the γ field, see in Figures 6.17 and 6.18. In the presence of convergent nudging iterates, we opt for base point q to be used in optimal balance. Base point q is physically grounded because of its slow nature in linear case and its consistency in convergence, but it comes with a drawback of an computational algorithm with the PV-inversion equations. Providing the possible choice of base point h is, however, a strong result, and base point h fits the formulation of the model dynamics. It also simplifies the application of the computational algorithm by omiting the inversion equations, but as a drawback, it requires more nudging iterations.

The diagnosed imbalance of base point q (not shown here) is few order smaller than that of other fields, but we do not observe this for base point h. The slow convergence of base point h limits the convergence to a fixed point h^*, while the nudging iterates gets closer to each other, and then, the scheme is terminated due to sufficiently small norm of the iterates. This slower convergence also reasons to the requirement of more iterations for base point h. In our published work Masur and Oliver (2020), the diagnosed imbalance of h takes total height $h + h_0$ into account instead of only free surface h. The imbalance provides smaller error of two order in h, as anticipated, but this small order emerge only due to slightly different definition.

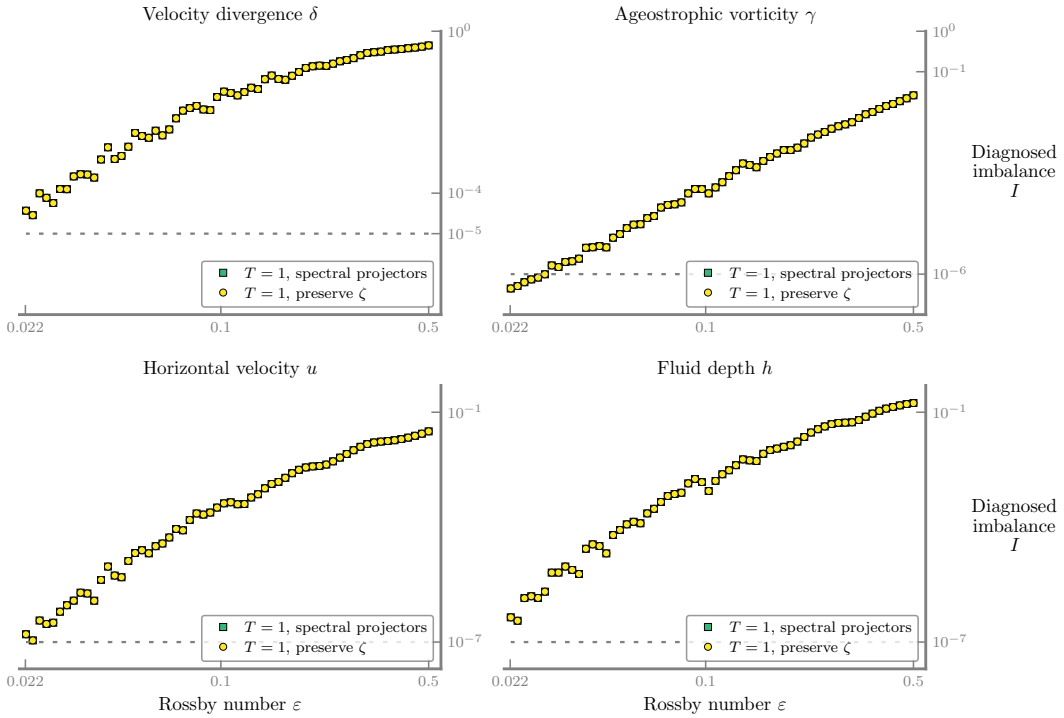

Figure 6.15: The liner-end boundary conditions give the same quality of balance in the quasi-geostrophic regime. Optimal balance uses the spectral projectors, the oblique and orthogonal ones and the ζ-preserving projectors with $T = 1$ and $t = 0.5$.

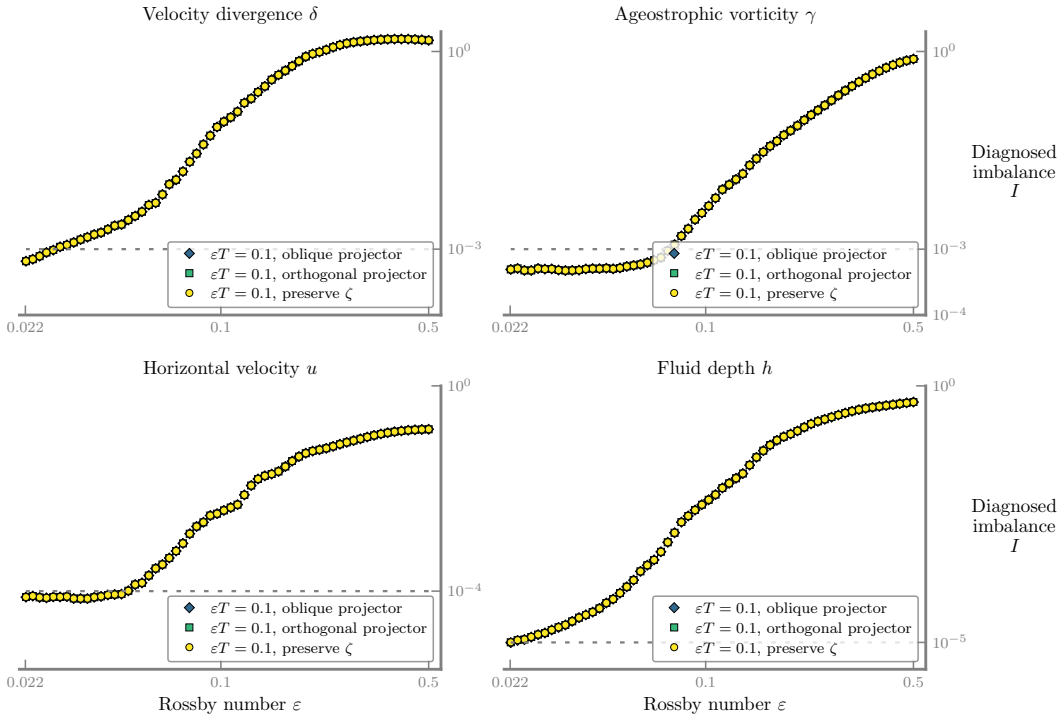

Figure 6.16: In the semi-geostrophic regime, the test case in Figure 6.15 is performed with $T = 0.1/\varepsilon$ and $t' = 0.05/\varepsilon$. The diagnosed imbalance is independent of the choice of the linear projectors, which give a convergent nudging scheme across the whole ε-range.

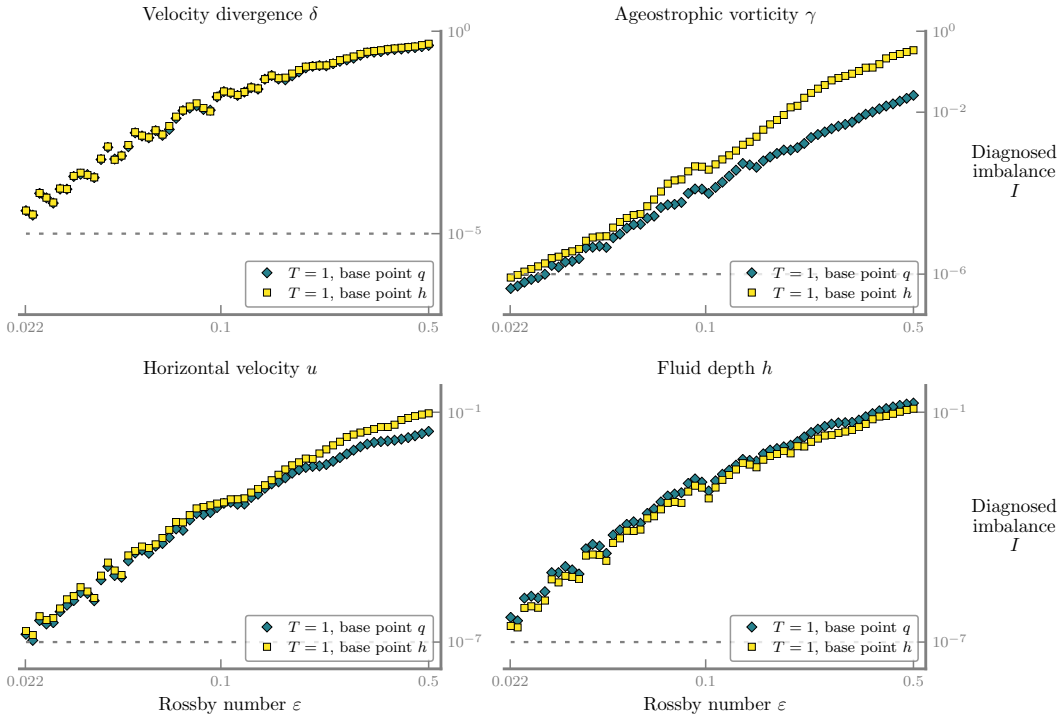

Figure 6.17: The base-point coordinates provide quantitatively closer diagnosed imbalances in the quasi-geostrophic regime. The diagnosed imbalance as $I(\varepsilon)$ is presented for base points q and h, when optimal balance uses $T = 1$ and $t' = 0.5$, and the exponential ramp function.

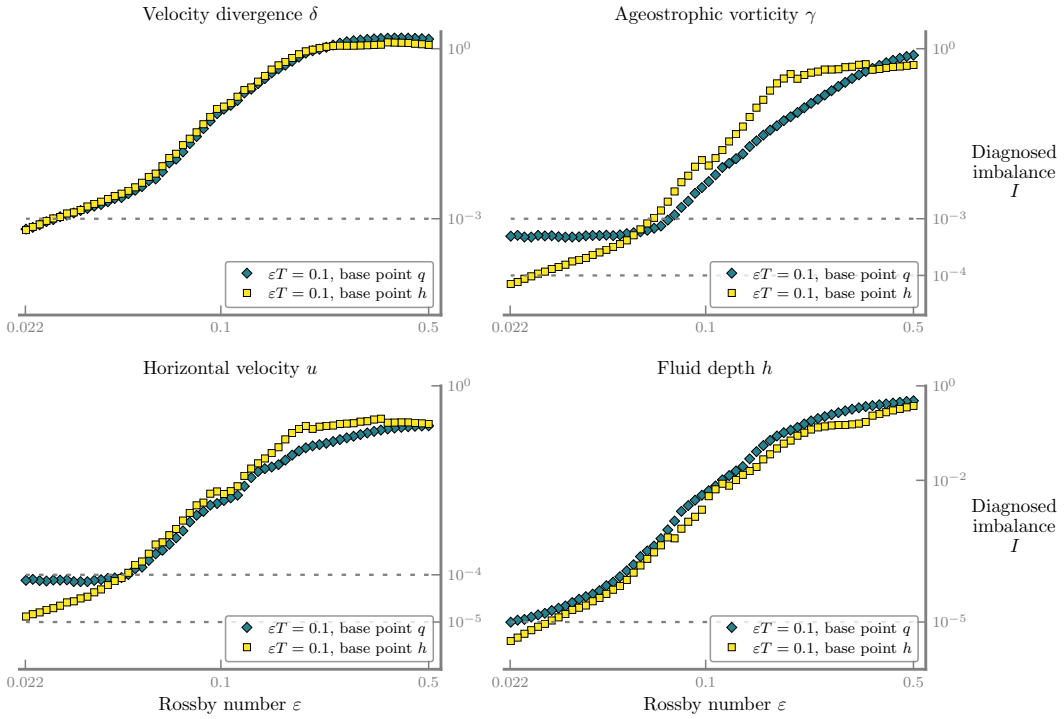

Figure 6.18: The test case in the above figure is run for the semi-geostrophic regime, when the settings kept the same except $T = 0.1/\varepsilon$ and $t' = 0.05/\varepsilon$. The quality of balance is comparable for both base points as in the quasi-geostrophic case.

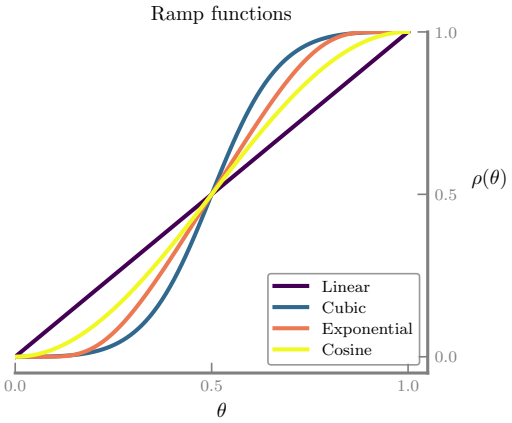

Figure 6.19: Ramp functions helps to build homotopy between the nonlinear equations and its linear form. The figure displays four different ramp functions ρ: i) in the form of the function (5.2.2) with $f(\theta) = \theta$, $f(\theta) = \theta^3$ and $f(\theta) = \exp(-1/\theta)$, which are respectively called the "linear", "cubic" and "exponential" functions, and ii) the "cosine" ramp function in (5.2.1).

6.6.4 Ramp function

The choice of ramp function is the last parameter to be tested for the quality of balance. We work on four possible choices of ramp functions: three of them are formulated in the expression (5.2.2) with $f(\theta) = \theta$ ("linear"), $f(\theta) = \theta^3$ ("cubic") and $f(\theta) = \exp(-1/\theta)$ ("exponential"), and the fourth one is the "cosine" ramp (5.2.1) of second order, which are drawn in Figure 6.19. The exponential ramp results in exponential decay; the other functions give algebraic decay as regards to their order, see Section 2.4.2.

The order of balance error is correlated with two components: i) a "dynamic" component determined by the derivatives of the ramp function and ii) a "boundary" component dependent on the vanishing derivatives of the ramp function at the end points. When ε is large, the dynamic component becomes dominant in the balance error. As the cosine ramp has smaller component, it gives better balance, which is more visible in the semi-geostrophic regime, see in Figures 6.20 and 6.21. A rapid change in the balance errors are visible starting from a specific ε value, where the type of the dominant error component changes. When ε is small, the boundary component, if not zero, becomes important, then it restricts the quality of balance. The small-ε asymptotics is clear in the quasi-geostrophic regime; the error due to the spatial resolution might conceal the impact of boundary component in the semi-geostrophic regime, see the Section 6.8 for the spatial resolution.

Intentionally left blank.

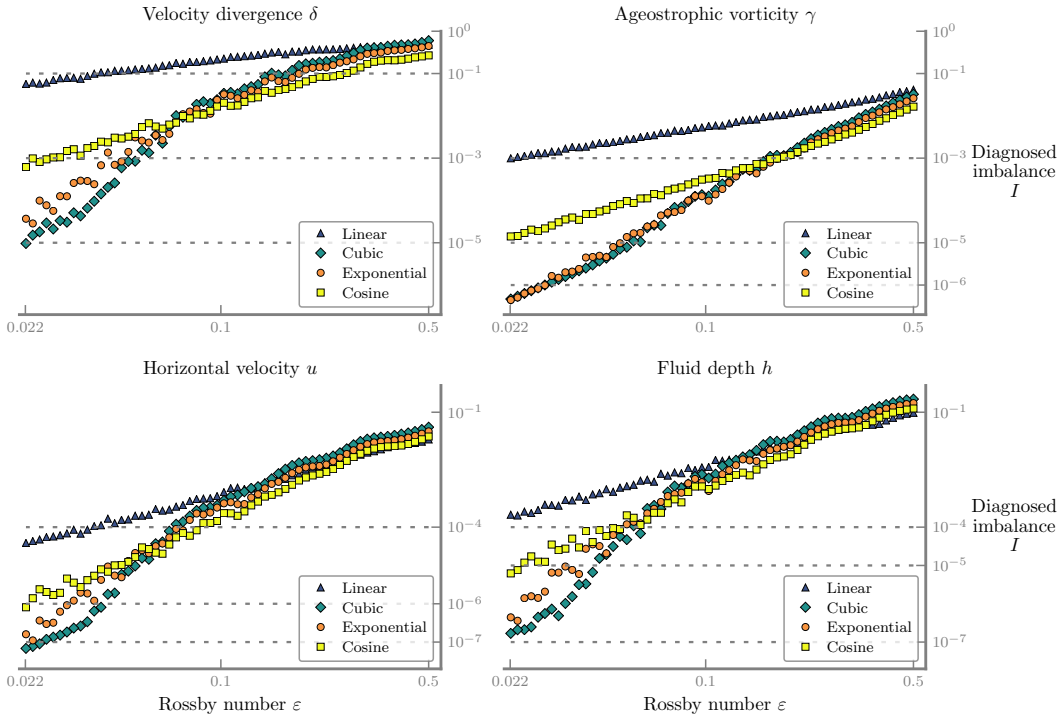

Figure 6.20: The cosine ramp function gives slightly less imbalances over practically important ε-range in the quasi-geostrophic regime. The figure shows diagnosed imbalances across the ε-range for different ramp functions. The base settings are preserved, and the integration lengths are $T = 1$ and $t' = 0.5$.

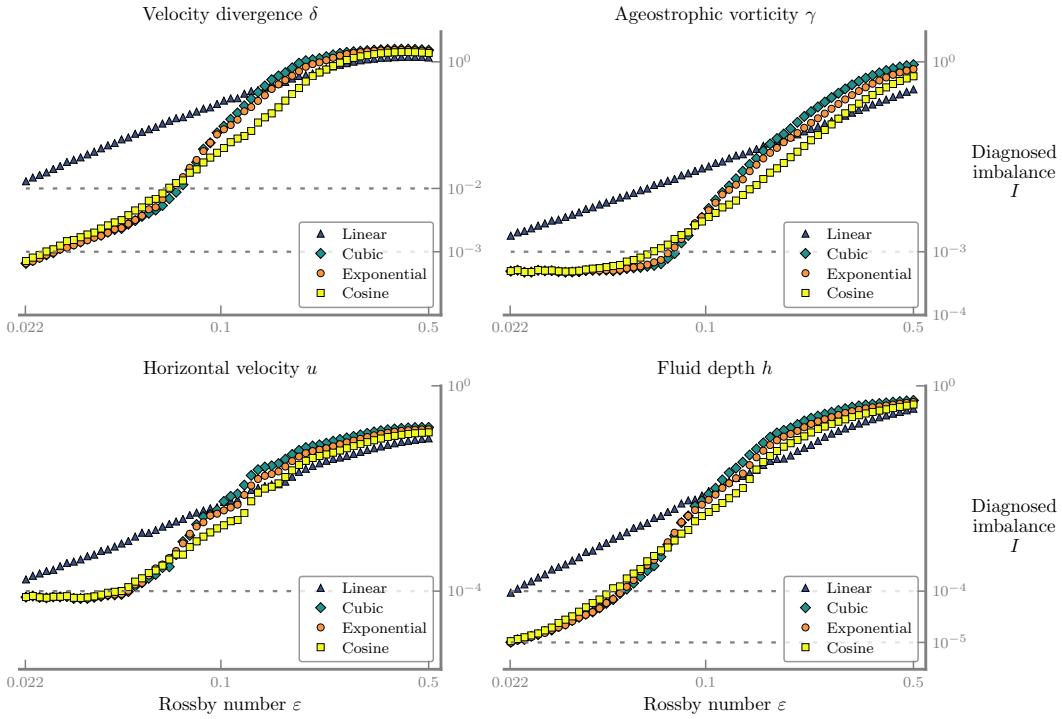

Figure 6.21: In the semi-geostrophic regime, smaller diagnosed imbalances using the cosine ramp are, here, more distinguishable over the same ε-range as in Figure 6.20. The test is executed with the base settings when $T = 0.1/\varepsilon$ and $t = 0.05/\varepsilon$.

6.7 Systematic exploration of diagnostics

The diagnosed imbalance in Section 5.4 is investigated by the comparison of diagnostics at linear and nonlinear ends and the type of error computation used in the diagnostics. We, still, preserve the base settings in optimal balance: the oblique projector, base point q and the exponential ramp function.

6.7.1 Diagnostics at linear vs. nonlinear end

In the computational algorithm, the diagnosed imbalance is quantified on nonlinear dynamics. We justify the possibility of comparable diagnostics at the linear end by introducing an alternative algorithm. The quantification of imbalances at the linear end eliminates the necessity of the full backward-forward nudging scheme in the rebalancing procedure. After the physical-time evolution up to $t = t'$, the flow $(\boldsymbol{u}', h')$ is integrated backward to the linear end, and the quantification follows in two ways: i) the norm of the gravity-wave component of a linear flow and ii) the relative difference between the linear flow and its Rossby-wave component as in our actual diagnosed imbalance (5.4.2). In Figures 6.22 and 6.23, we use the label of "linear end $\|.\|_g$" for the error in i), "linear end" for the error in ii), and "nonlinear end" for our actual error.

The balance error of the $\boldsymbol{u}$-h fields are analogous at "nonlinear end" and "linear end" in both regimes. The relative error in the δ-γ fields at "linear end", nonetheless, are constant across different ε values, since these fields have no contribution to the linear Rossby-wave component. This problem can be solved using the norm of linear gravity-wave components at "linear end $\|.\|_g$". The latter norm also supplies close proportional scaling to the error at the "nonlinear end" for all fields. We, hence, verified reasonable diagnostics of imbalance at the linear end by introducing a new algorithm, which is advantageous to computational time.

6.7.2 Diagnostics error type

Our main settings include the stopping criterion (5.5.1) and the diagnosed imbalance (5.4.2) with the relative error. To observe the effect of diagnostic error type, we choose the absolute error instead, and the referred equations are replaced with (5.4.3) and (5.5.2), respectively. The absolute error inevitably reasons more iterations to terminate the nudging scheme for the same κ value, which does not enhance the quality of a balanced state: It has been already observed in the test of κ-sensitivity. We, then, see the effect of error mainly on the computation of the diagnosed imbalance.

The effect of the error type becomes visible on scaling the order of terms, which occurs in the relative error. The δ-γ fields, same for δ'-γ', are decomposed as follows $\delta' = O(\varepsilon)+$ small imbalanced contributions (see Section 2.4), and in the rebalanced flow, the imbalanced contributions become exponentially small. The $\boldsymbol{u}$-h fields, however, has $O(1)$-term additionally. Regarding both regimes, the scaling of the diagnosed imbalances differentiates in the δ-γ fields, especially for large ε values; it is quantitatively comparable in the $\boldsymbol{u}$-h fields, see in Figures 6.24 and 6.25. The choice of error type, hence, changes only the computation of diagnostics without any effect on the quality of balance. This test is only possible to perform while preserving base point q, because base point h holds some limitations on convergence.

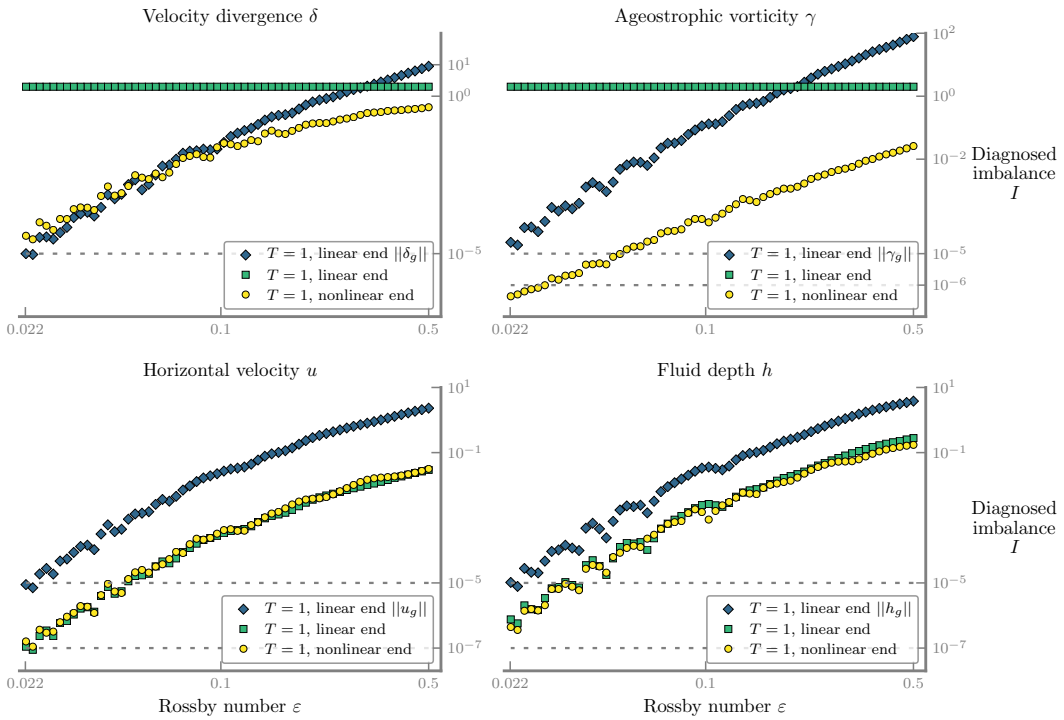

Figure 6.22: The quantification of imbalances at the linear and nonlinear ends give similar observation in the quasi-geostrophic regime. These quantifications are done by the norm of linear gravity waves (linear end $\|.{}_g\|$), the relative norm of the linear flow and its Rossby wave (linear end), and the relative norm of nonlinear flow (nonlinear end). The time lengths are $T = 1$ and $t' = 0.5$.

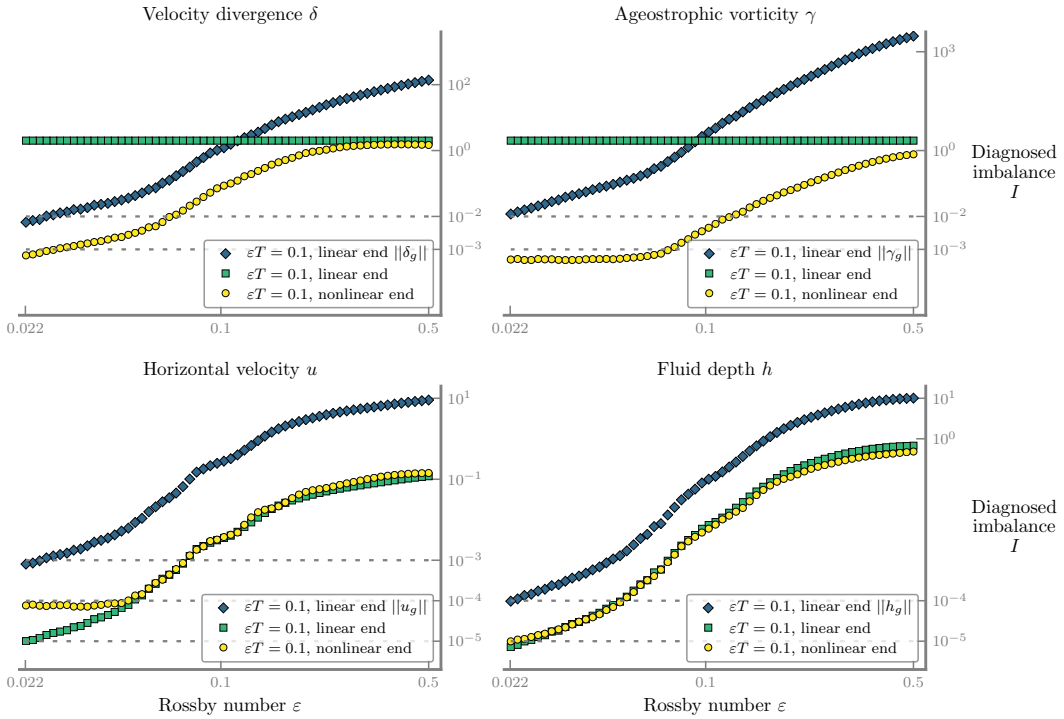

Figure 6.23: The above test is performed in the semi-geostrophic regime using $T = 0.1/\varepsilon$ and $t' = 0.05/\varepsilon$. The diagnostics of imbalance is possible at the linear end, as well as at the nonlinear end.

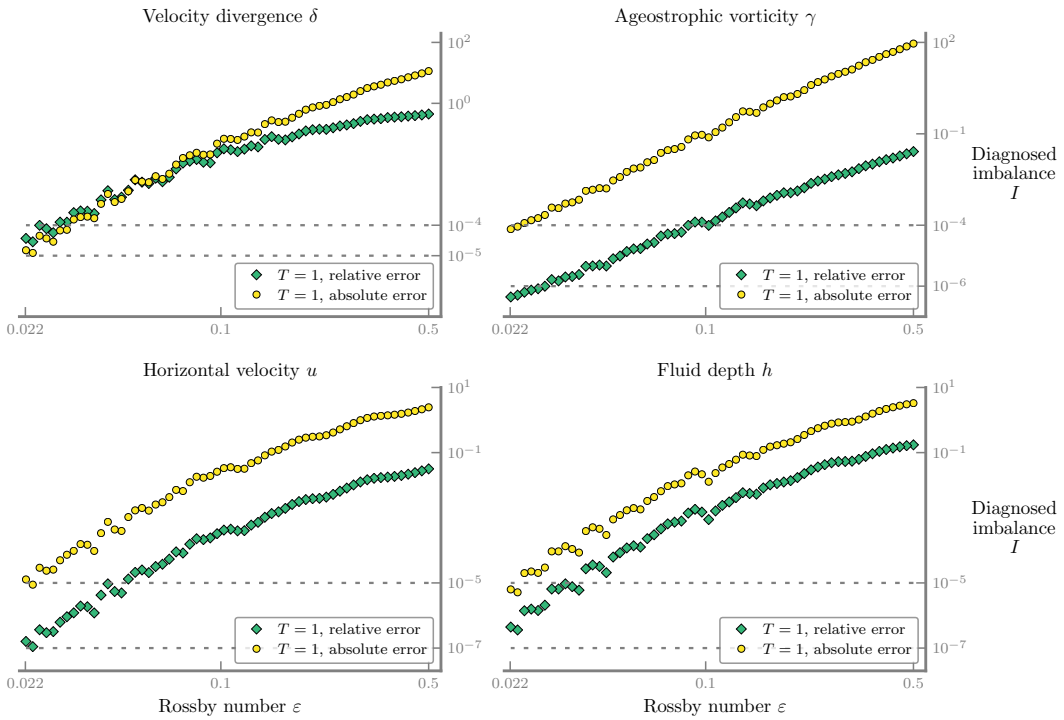

Figure 6.24: Using the absolute norm in the diagnosed imbalance and the stopping criterion gives reliable quantification of imbalances in the quasi-geostrophic regime. The figure shows the diagnosed imbalances for the absolute and the relative norms, and time lengths are $T = 1$ and $t' = 0.5$ in the simulations.

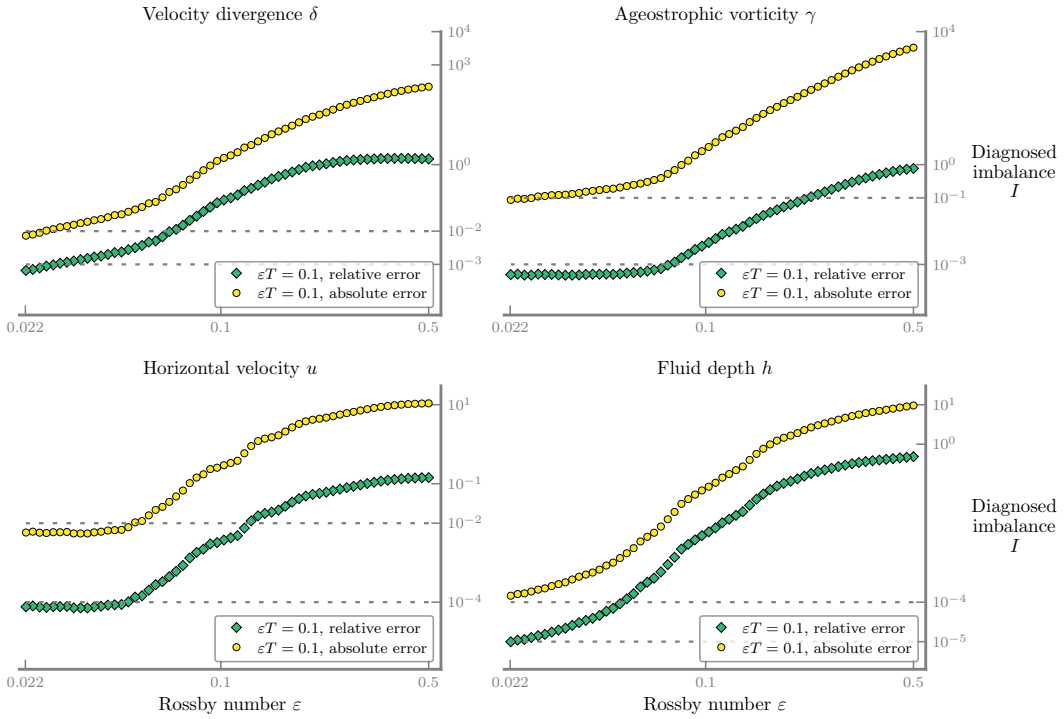

Figure 6.25: The above test case is executed in the semi-geostrophic regime, while time lengths are $T = 0.1/\varepsilon$ and $t' = 0.05/\varepsilon$. The diagnostics of imbalances is also carried out using the absolute norm.

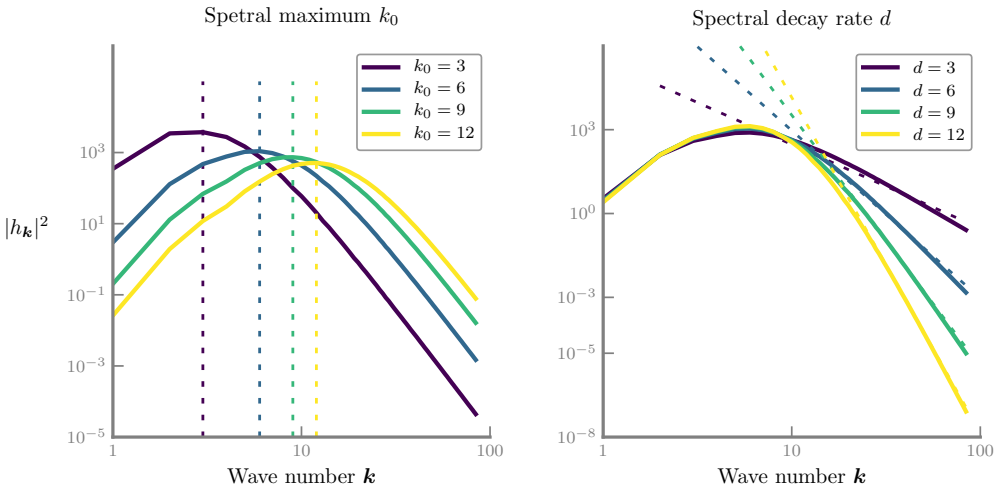

Figure 6.26: The randomly generated height h field has different energy distribution at a particular spatial scale depending on spectral maximum k_0 and spectral decay rate d. In the left figure, $d = 6$ is fixed; k_0 gets different values. In the right figure, $k_0 = 6$ is fixed; d gets different values.

6.8 Initial condition structure

Optimal balance provides different balanced states depending on the structure of the initial condition. Our commonly used initial conditions consist of the randomly created h field with spectral maximum $k_0 = 6$ and spectral decay rate $d = 6$, and $\boldsymbol{u}$ field is set by the geostrophic balance. To test the dependence of the quality of balance on the structure of initial conditions, we produce several different initial conditions: When one of the parameters k_0 and d stayed fixed, say 6, and the other takes values among $3, 6, 9, 12$. These initial flows have energy spectra as displayed in Figure 6.26.

The test cases for different k_0 and d indicate the impact of spatial resolution on optimally balanced states: Smaller k_0 and larger d are analogous to finer resolution in the model. In the quasi-geostrophic regime, our general initial data is already well resolved, so smaller k_0 and larger d values put forward no significant progress in balance, see Figures A.18 and A.19 in the appendix. In the semi-geostrophic regime, nevertheless, finer spatial resolution improves the quality of balance especially for small ε values. This improvement can be noticed after a cutover in the ageostrophic δ-γ fields, see Figures 6.27 and 6.28. A finer resolution also lowers the ε limit that can be studied without any numerical restriction as expected; hence, it advanced the quality of balance in the semi-geostrophic scaling regime.

To stress the reliability of optimal balance, we perform the same experiment for turbulent flows using he "Kolmogorov -5/3 spectrum" which predicts the energy spectrum of $\boldsymbol{u}$ field in a turbulent flow. In our settings, this flow is a geostrophic $\boldsymbol{u}$ field with the spectral decay $d = 5/3$. To derive it, we work with randomly created h fields, where $d = 11/3$ for different k_0 values. Depending on the above result, the sensitivity of the diagnosed imbalance on the turbulent flow can be anticipated, but yet, we separately present the results to emphasise some restrictions and diverging cases. The spatial energy distribution of turbulent flows and their diagnosed imbalances are displayed in Appendix A.7.

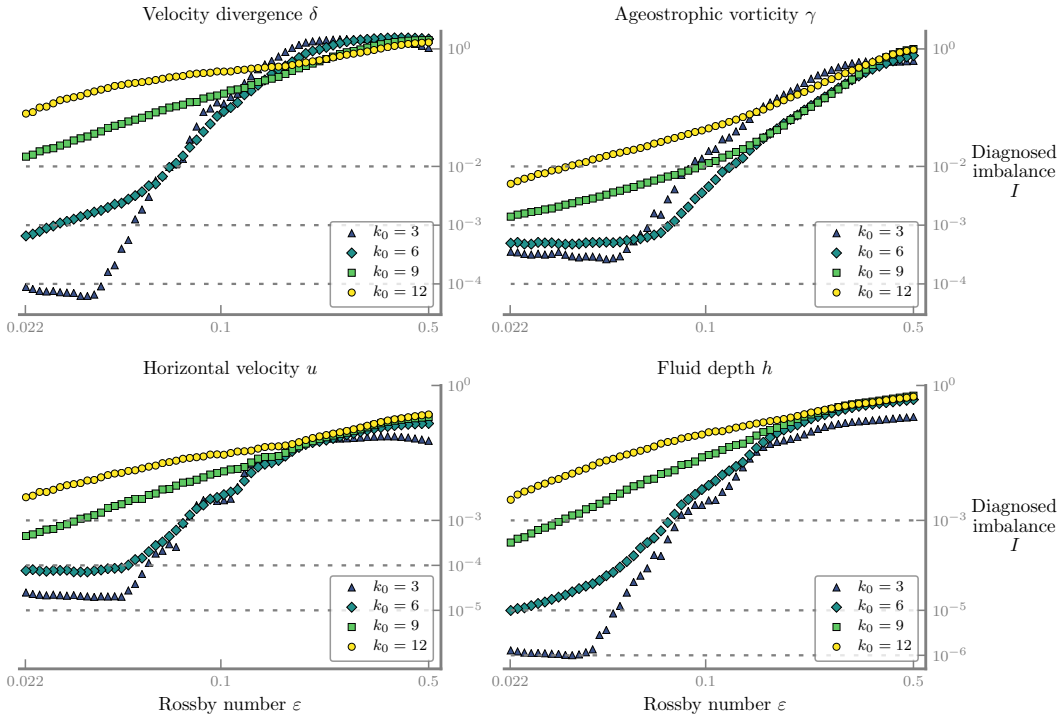

Figure 6.27: Holding the spectral decay $d = 6$ and the spectral maximum k_0 at large scales, initial configurations result into higher quality of balance especially for small ε values. This test case uses the PV-based boundaries and the integration lengths: $T = 0.1/\varepsilon$ and $t' = 0.05/\varepsilon$.

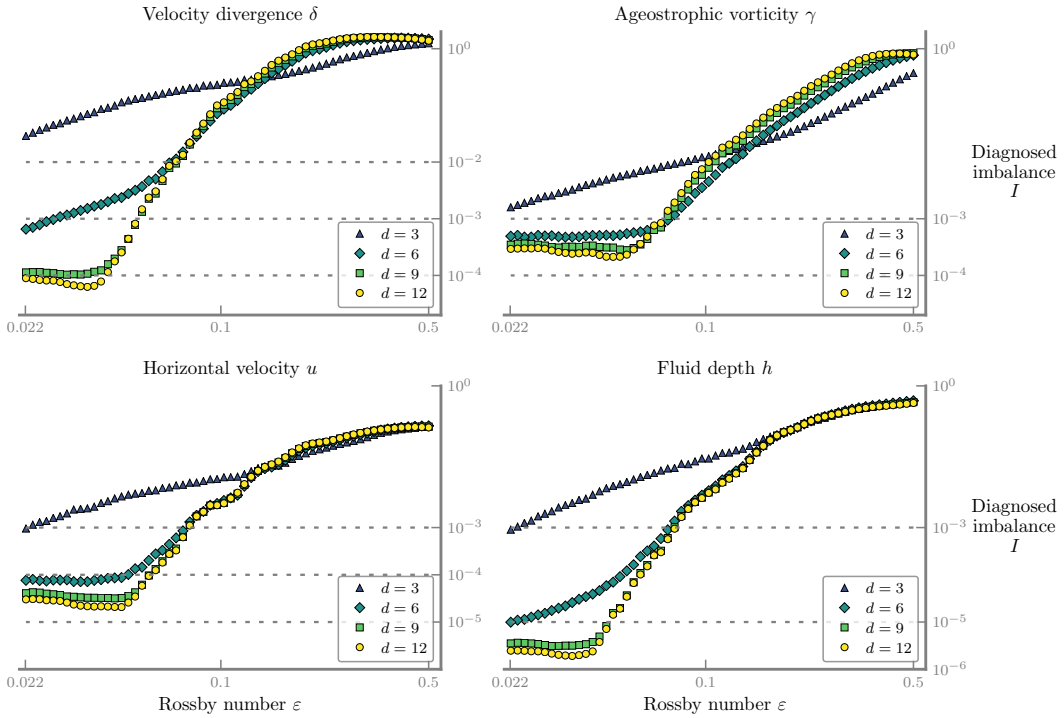

Figure 6.28: Initial condition with steeper spectral tail increases the quality of balance especially smaller ε-range in the semi-geostrophic regime. The settings as in Figure 6.27 are reused.

Chapter 7

Discussion and Conclusion

The method of "optimal balance", in this thesis, is numerically investigated in a single-layer rotating shallow-water model on the f-plane. The novelty of our work lies under the application of optimal balance in the primitive velocity-height variables and the simplicity of its application. Implementing at the top of an existing code, optimal balance provides good-quality balance without any special treatment and expensive computational cost. This method has already numerically studied by Viúdez and Dritschel (2004) in a special Lagrangian setting, but as the majority of global models is operated in Eulerian primitive variables, it was essential for us to work in these variables.

The application of optimal balance returns a boundary value problem in time solved iteratively by a backward-forward nudging scheme, and an obtained solution is an optimally balanced state. We, in this thesis, tested methodically the quality of balance for the effect of several design parameters in our numerical setting. Besides the numerical implementation, we have also provided "quasi-convergence" of the nudging scheme in a finite-dimensional model, which has similar dynamics to the evolution of rotating shallow-water model in the semi-geostrophic scaling limit. In the upcoming sections, we summarise the key findings of this thesis under two aspects: theoretical and numerical ones.

7.1 Theoretical aspects

The convergence in the nudging scheme is characterised by the closeness of two following iterates. For the finite-dimensional model (3.1.1), the difference between the successive iterates decreases only up to a small, non-vanishing residual: There is no convergence to the solution of the BVP, when the number of nudging iterations are getting larger. The nudging scheme, therefore, has no unique solution and we call this convergence behaviour "quasi-convergence".

The nudging balance error between the obtained balanced state and the slow manifold is determined by the quality of the optimal balance scheme and the quality of convergence of the nudging implementation. This gives rise to two components of the overall error, the balance error and the termination residual. The balance error comes from the direct transition of the system from the trivial slow manifold to an approximate nonlinear slow manifold, and it is already bounded by $O(\varepsilon^{n+1})$ by Gottwald et al. (2017). The termination residual is the contribution of the nudging scheme to the overall error, proved in Chapter 3. As it is of the same order as the balance error, it has no impact on the asymptotics of the overall error. In this result, it is assumed that the ramp function holds the algebraic order

condition. For future work, the convergence of the nudging scheme should be analysed for the ramp function satisfying the exponential order condition, see Masur et al. (2022). Another further step could be a rigorous analysis in infinite dimensions, e.g., for the rotating shallow-water equations.

7.2 Numerical aspects

The quality of optimally balanced states of the shallow-water flows is explored in the definition of the diagnosed imbalance. When the best parameters are chosen in the computational algorithm, the balance error is limited by small-scale dynamics resolved by the spatial grid scale. By testing the structure of the initial data, we observed: In the quasi-geostrophic regime, the initial condition employed in the test cases is well-resolved. In the semi-geostrophic regime, initial conditions with less small-scale dynamics near the grid resolution decrease the excitation of imbalances especially for the small-ε range. It is concluded that the spatial resolution influences the quality of balance more in the semi-geostrophic regime. Through the diagnostics, therefore, smaller imbalances are obtained for smoother flows, though optimal balance is also very successful to balance highly non-smooth flows.

Providing better balance states is also correlated by the properties of continuous deformation. The choice of longer ramp time, not longer than the optimal value, makes transition between the linear and nonlinear dynamics smoother and produces better balance. As another parameter, the choice of ramp function also improves balance slightly due to the effect on the deformation. The cosine ramp is a good choice for the large-ε range, that is practically used. For small-ε range, the numerical restrictions, like grid scale, play role in the semi-geostrophic regime, and the cubic ramp gives better balance in the quasi-geostrophic regime.

The most decisive parameters in the algorithm are the linear projector acting on a linear flow and the nonlinear projector, so-called base point. The best combination of these parameters is using base point q and the oblique projector together. Though both choices of base point balance flows in close quality, base point q is preferred to base point h due to its faster convergence and larger basin of convergence. On the other hand, using base point q brings about a more complicated computational algorithm, where the PV-inversion equations become necessary, albeit the model dynamics is free of the PV field. As an important aspect of our work, the variable h can be, however, replaced as base point in convergent cases. Base point h is more in the line with using the primitive variables. This yields a simpler algorithm without need of inversion equations, but there is a compromise in the number of nudging iterations.

The choice of linear projector depends critically on the angle between the Rossby-wave and gravity-wave subspaces. Since smaller-ε values make the subspaces more non-orthogonal, the oblique projector stands out by its better convergence in every case. This boundary condition projects the linear phase space onto the Rossby modes in spectral space, but it is also represented by some elliptic PDEs solving a linear PV-inversion preserving the linear PV. By this formulation, the oblique projector becomes advantageous to apply on general models in general geometries. In those general settings, when base q is also included, the computational effort to solve elliptic PDEs for the linear and nonlinear inversion equations might need a careful care. As a result, the robust convergence is provided by the projectors preserving linear and nonlinear PVs in the algorithm, and

this result supports strongly the idea of working with PV-based balancing algorithms in geophysical fluid dynamics.

Our implementation is based on the simple f-plane rotating shallow water model. As next step, the performance of optimal balance can be investigated on the β-plane in the equatorial region or on the sphere. Getting closer to the equatorial region where the mixed Rossby-gravity waves becomes visible in the domain of dynamics, the time scale of the Rossby and gravity-wave modes are poorly separated. In this case, the effect of the design parameters can be more perceivable especially the choice of boundary conditions. When stratified models are concerned, a reference stratification must be the part of the formulation while ramping its perturbation. As the main target is to apply optimal balance to full atmosphere and ocean models, a significant difficulty is not foreseen. There is, nevertheless, a need to formulate the linear-end and nonlinear-end boundary conditions carefully. Due to large model sizes, the choice of parameters of optimal balance might need careful tuning, but the method itself is a diagnostic tool that can be switched on if necessity, possibly even locally in parts of the domain.

Appendix A

Additional figures

This appendix includes unshown figures in Chapter 6.

Intentionally left blank.

A.1 Qualitative analysis

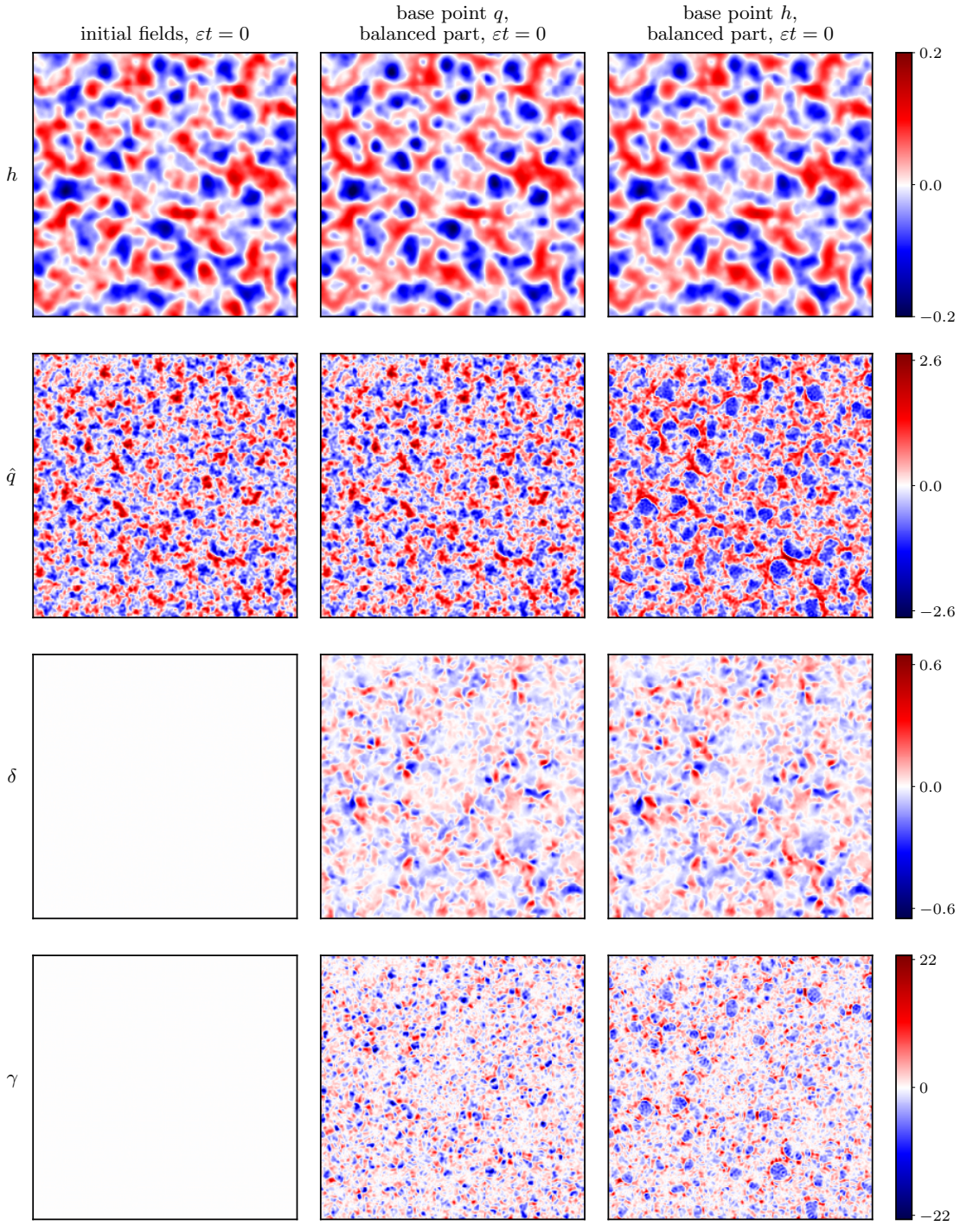

Figure A.1: Optimally balanced parts of the nearly-balanced flow show rather similar characteristics for both base points in the semi-geostrophic regime. The figure presents, for $\varepsilon = 0.1$, the initial shallow-water flow fields in geostrophic balance (left column), and their balanced parts obtained using base point q (middle column) and by base point h (last column) with the oblique projector and ramp time $T = 0.1/\varepsilon$. The fields are shallow water free surface h (first row), mean-free potential vorticity (PV) $\hat{q} = q - \bar{q}$ (second row), velocity divergence δ (third row), and ageostrophic vorticity γ (last row).

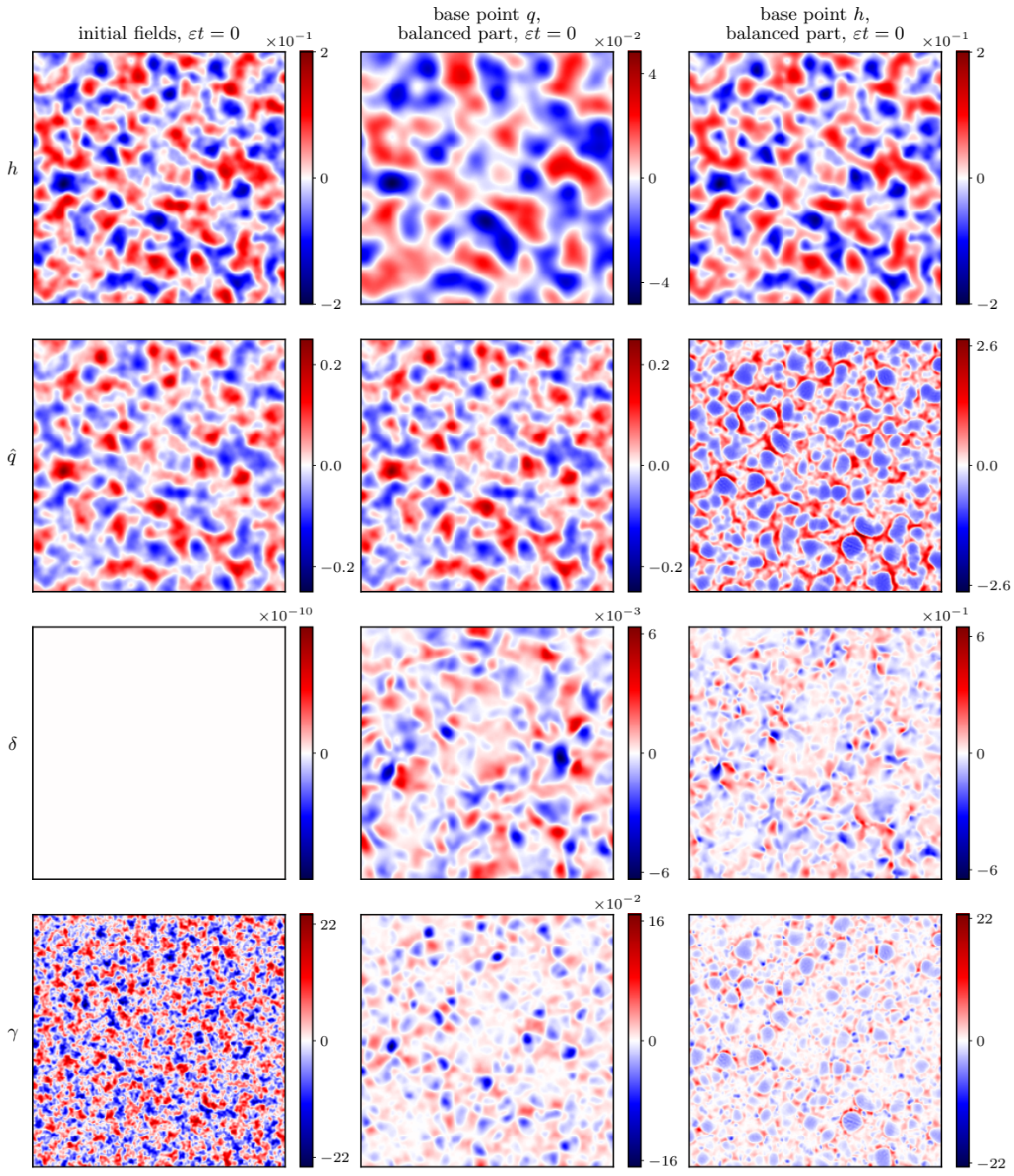

Figure A.2: The same test case as in Figure A.1 is performed with zero velocity instead of holding geostrophic balance in the initial flow. Starting with this unbalanced flow, optimal balance executes balanced parts with dissimilar quality for base point q and h.

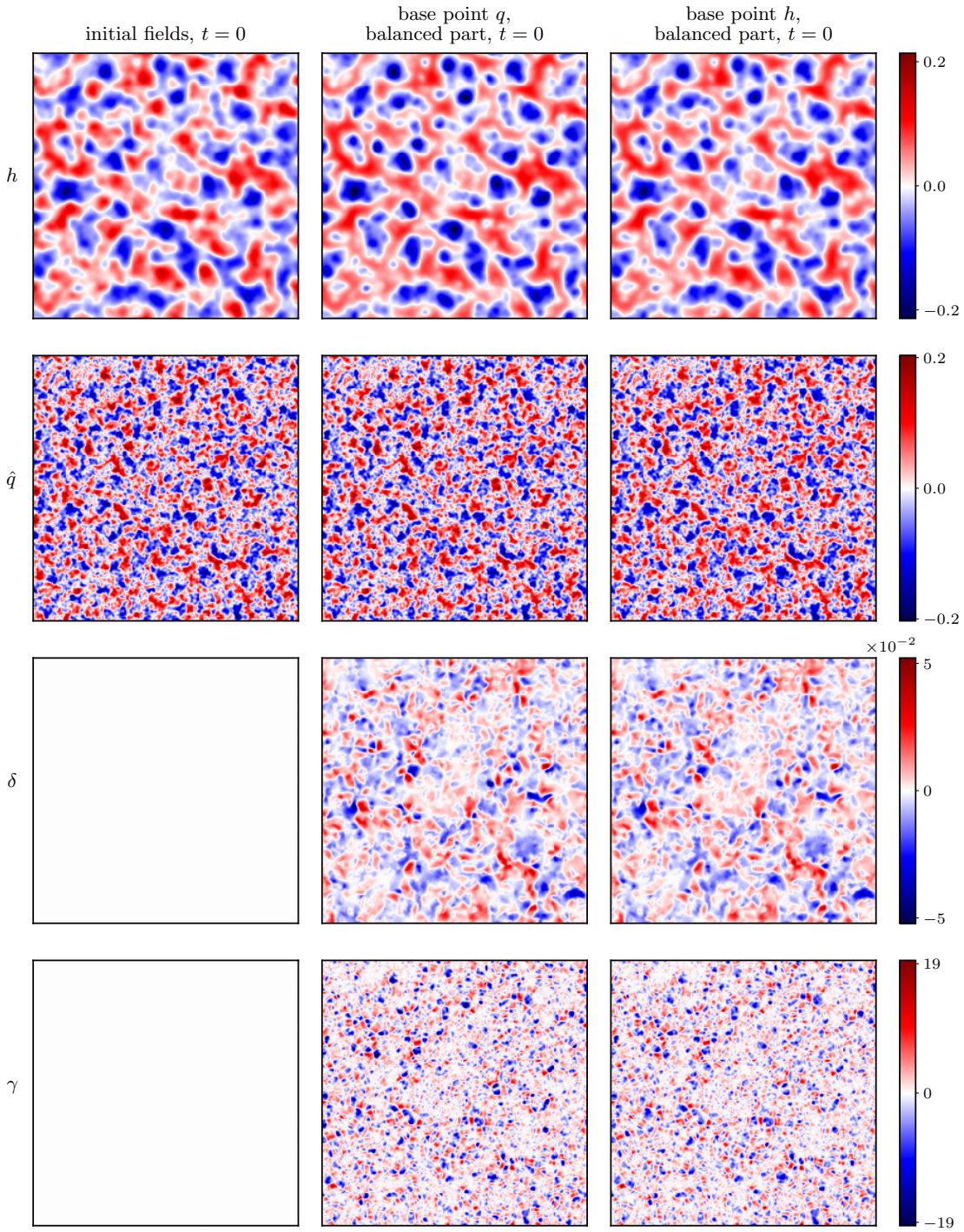

Figure A.3: Starting with the initial condition near balance, optimal balance with both base points converges to very close balanced states. The test case in Figure A.1 is executed, here, in the quasi-geostrophic regime. While keeping all other settings the same, optimal balance employs ramp time $T = 1$.

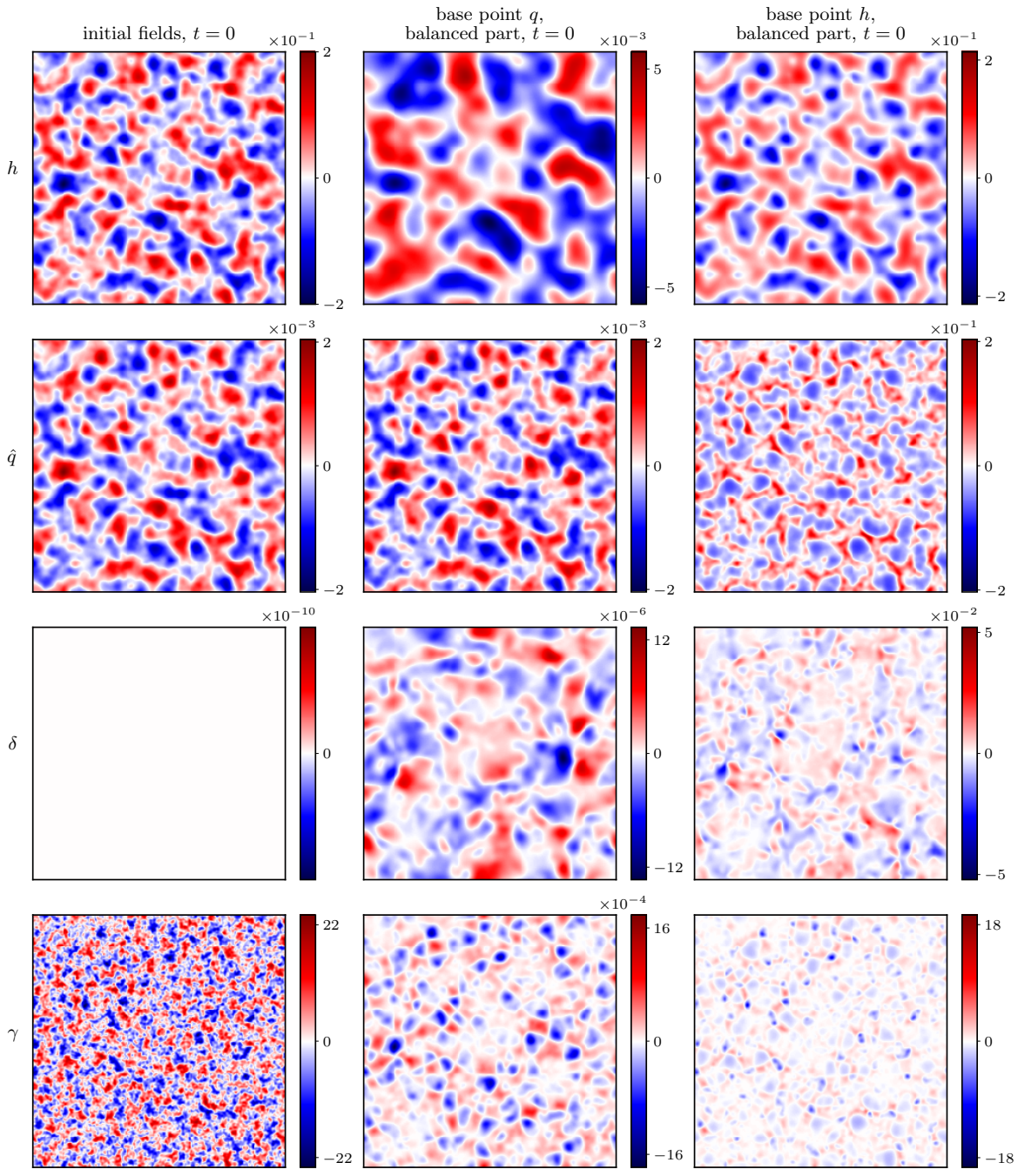

Figure A.4: By keeping all the settings as in Figure A.3, the test case is run starting with the initial flow fields involving zero velocity. In this case, optimal balance for different base points produces qualitatively different balanced states.

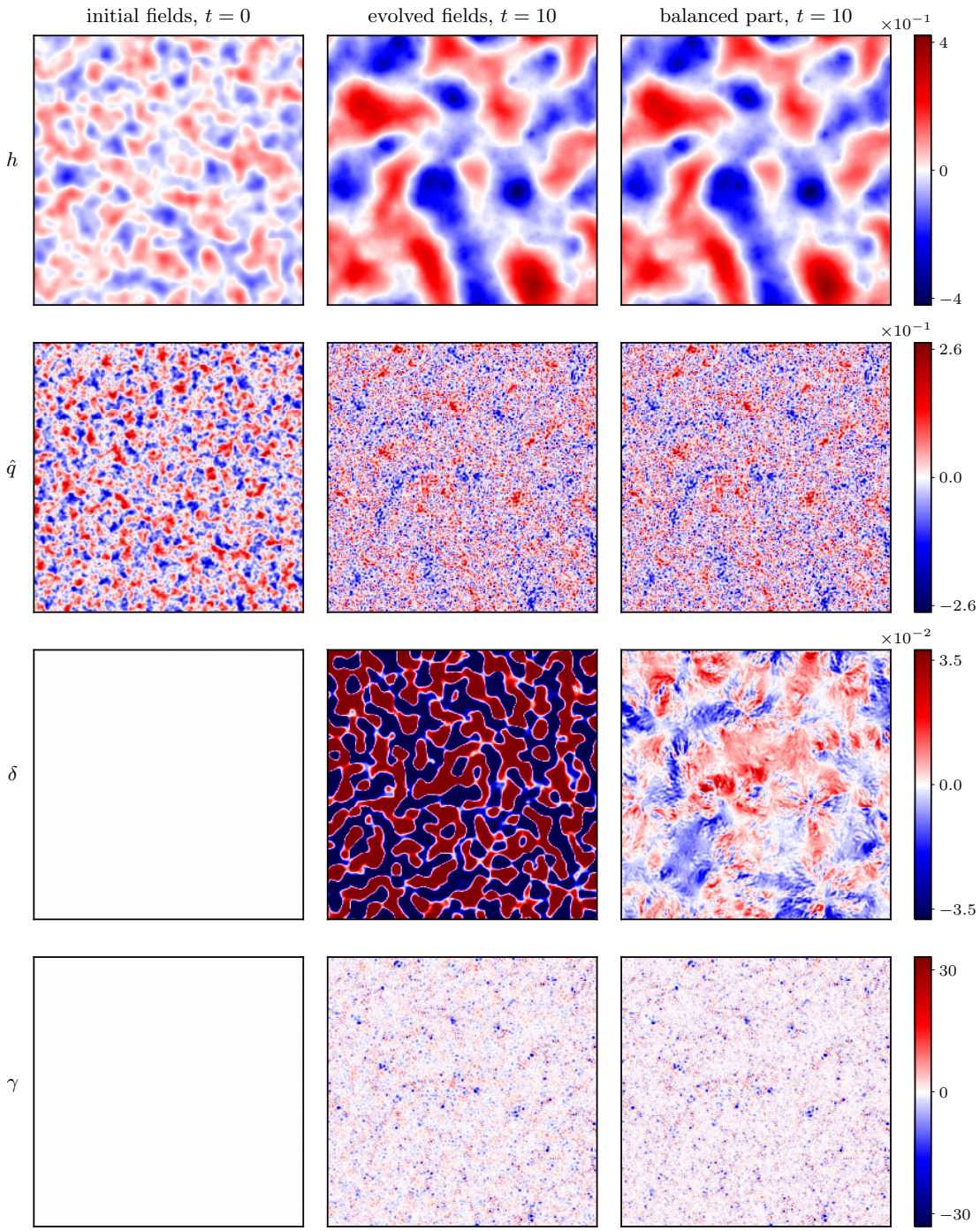

Figure A.5: Optimal balance employing base point q is used to balance the evolved fields (middle column) in the quasi-geostrophic regime, and it is successful to remove excessive parts of ageostrophic fields (see last column). The evolved fields, for $\varepsilon = 0.1$, are generated by evolving the initial fields (first column) up to $t = 10$. Optimal balance uses ramp time $T = 1$ and base point q.

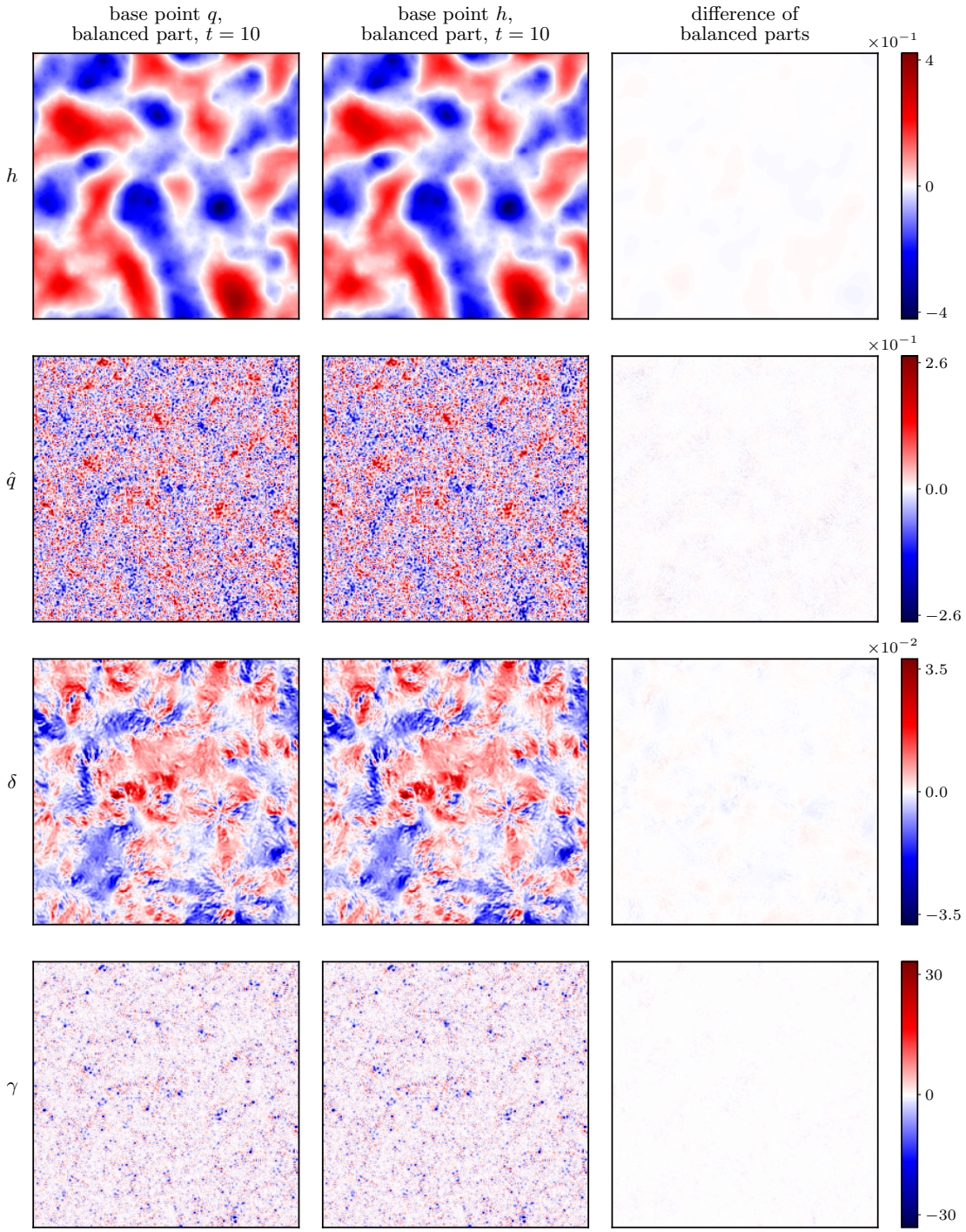

Figure A.6: The balanced parts extracted using base points q and h are nearly identical in the quasi-geostrophic regime, as can be seen in the relative difference (last column). The evolved fields in Figure A.5 are also balanced for base point h with the same ramp time, and the balanced part of both base points are displayed: that of base point q (left column) and that of base point h (middle column).

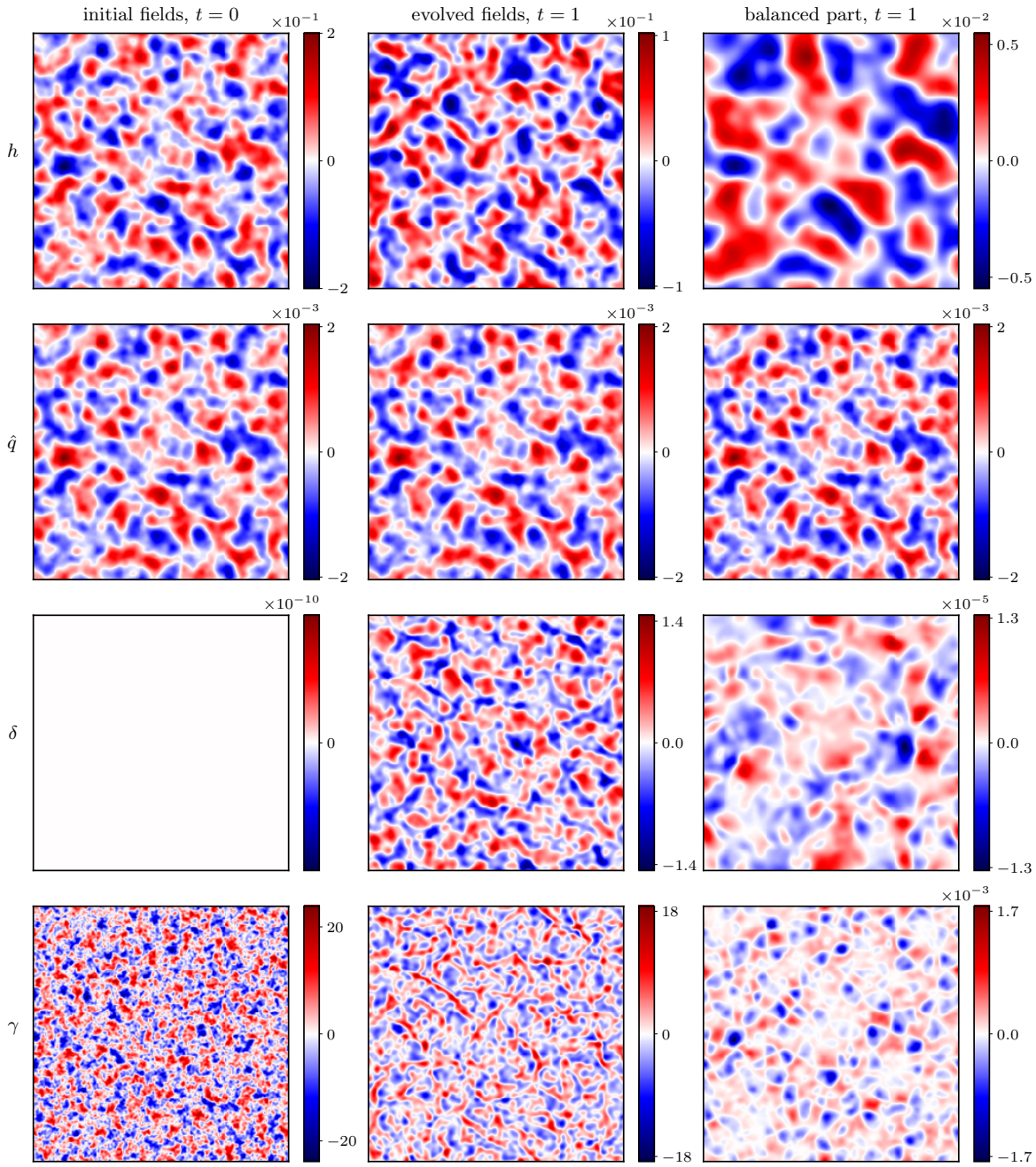

Figure A.7: Using zero velocity instead of holding geostrophic balance in the initial fields, the same test as in Figure A.5, is executed for a shorter physical-time length, $t = 1$, in the quasi-geostrophic regime. The change of the initial configuration produces ageostrophic δ-γ fields, which are dominated by imbalance, and optimal balance using base point q removes their significant part.

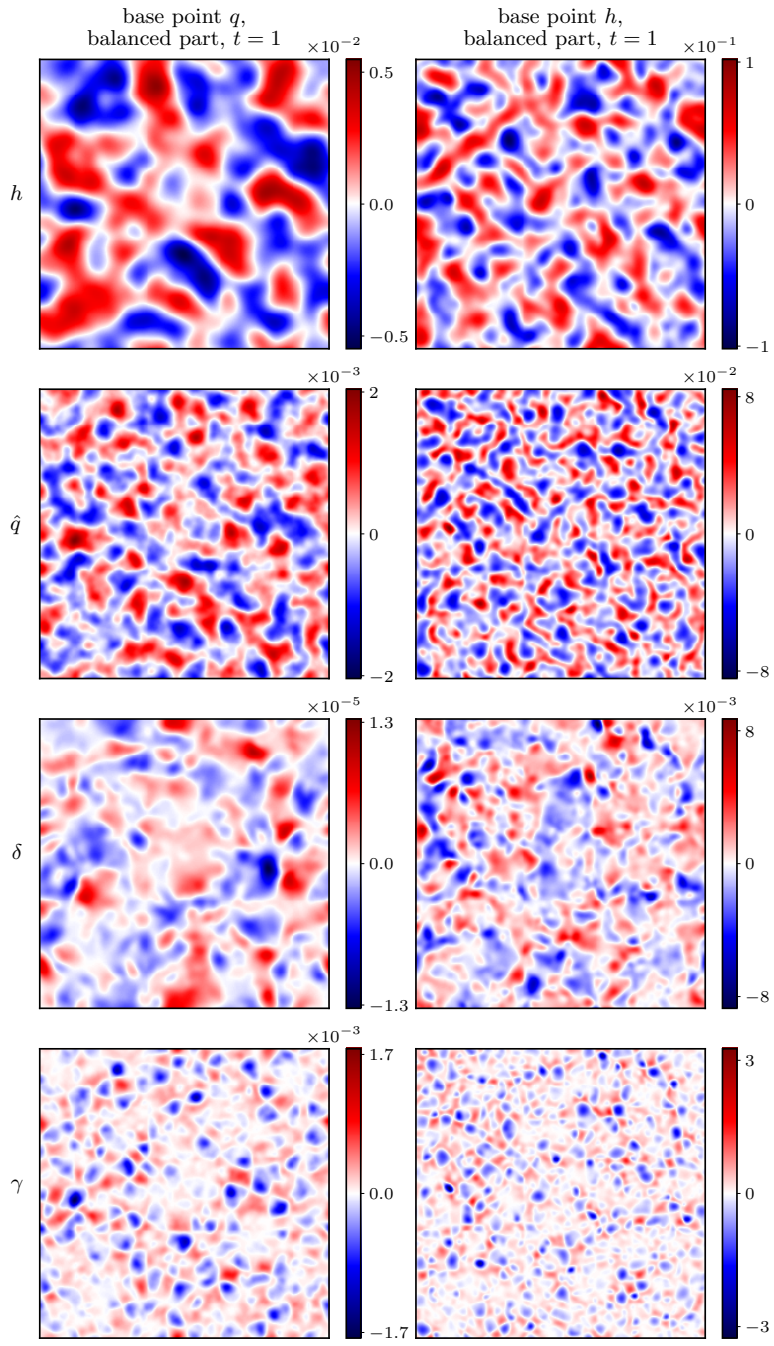

Figure A.8: Optimal balance with base point h is also applied on the evolved fields in Figure A.7, while the design settings are retained. The balanced parts of base points q and h show structural difference unlike those for the nearly-balanced initial condition in Figure A.6.

A.2 Viscosity

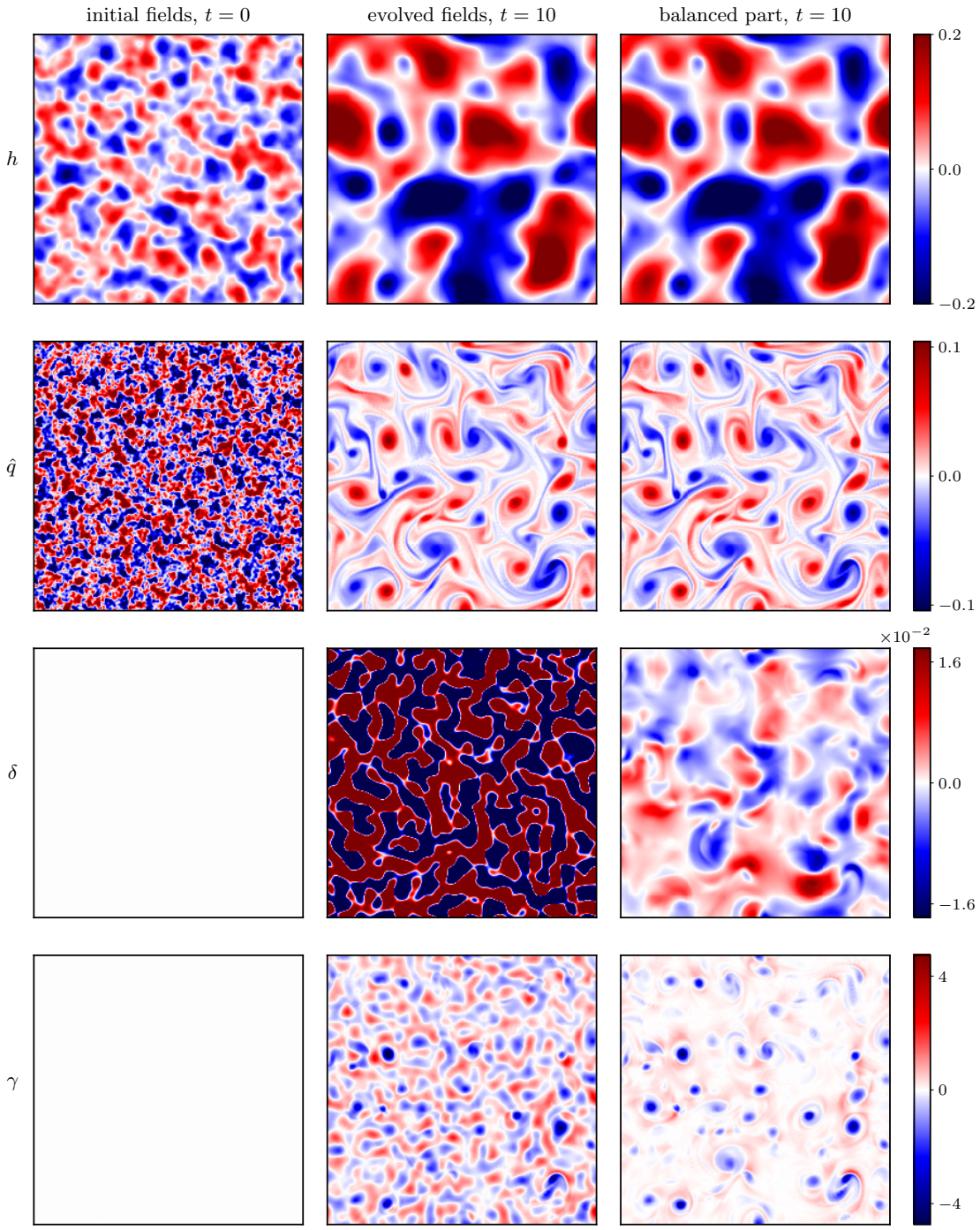

Figure A.9: The viscosity term dissipates the energy at small scales, so that flow fields have smooth pattern by being significantly void of small-scale structures. The figure displays the effect of the viscosity term for $\mathrm{Re} = 3 \times 10^3$ in the quasi geostrophic regime. The test case is the same as in Figure A.5 which is run without the viscosity term.

A.3 Convergence of the nudging scheme

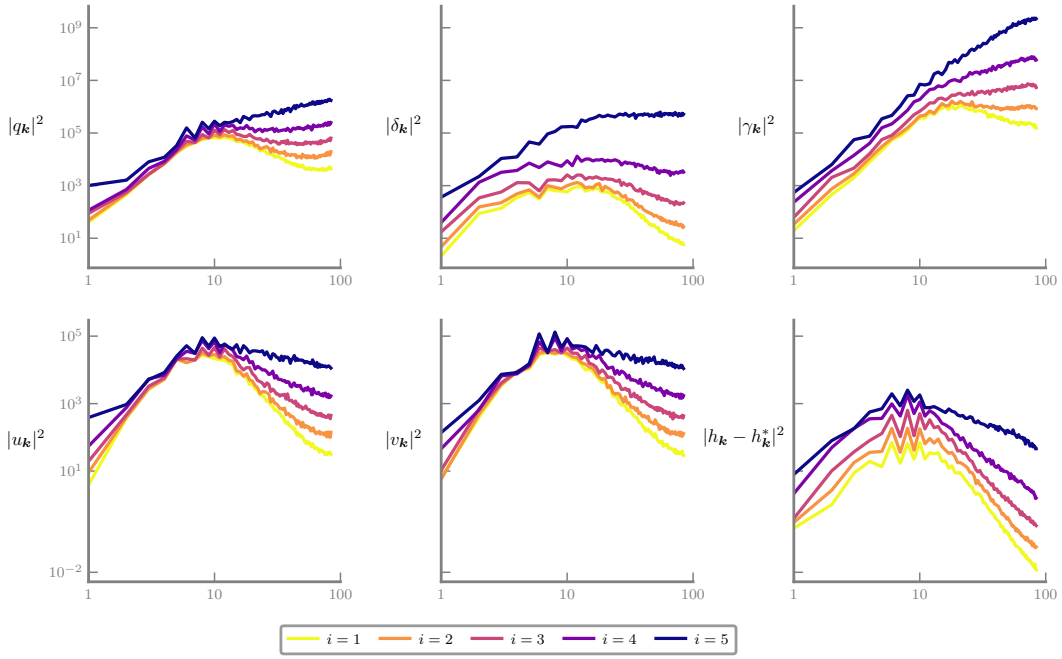

Figure A.10: The nudging iterates diverge for all fields, when h is preserved at the linear end in the semi-geostrophic regime. The iterative scheme using the h-preserving projector for $\varepsilon = 0.1$ is repeated with the same setting as in Figure 6.9, and energy spectra of all fields are separately presented.

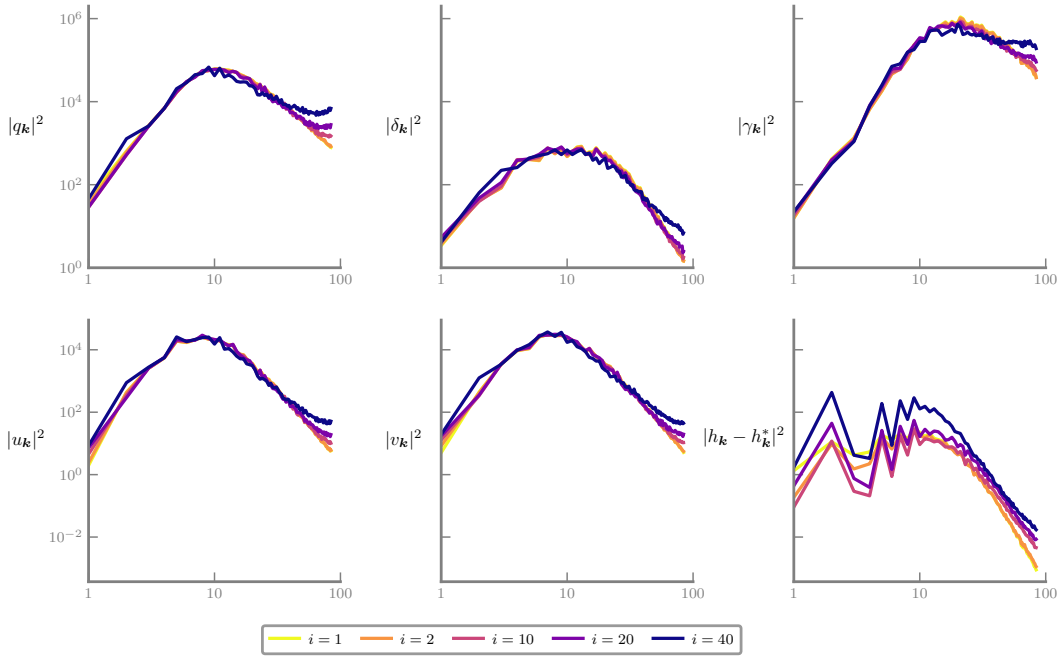

Figure A.11: The height field h diverges from base point h; however, energy accumulates at higher wave numbers for other fields, when ζ is preserved at the linear-end boundary in the semi-geostrophic regime. The other settings for the nudging scheme are maintained as in Figure 6.9.

A.4 Optimal integration time scales

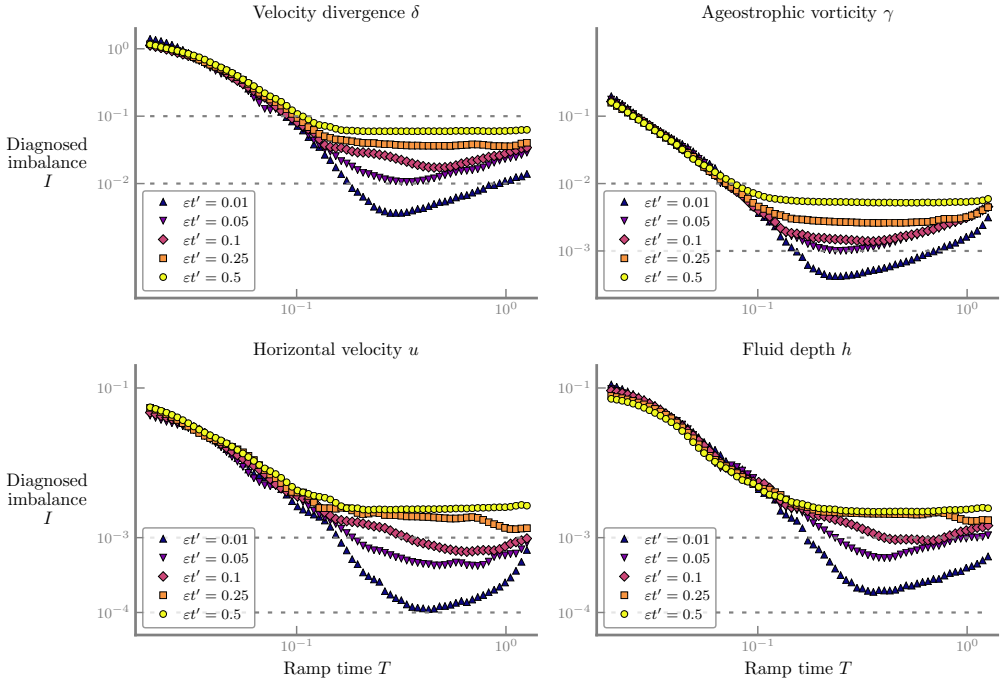

Figure A.12: Diagnosed imbalance I is analysed as a function of the ramp-time length T, $I(T)$, in the semi-geostrophic regime for $\varepsilon = 0.1$. The ramp-time length T horizon shows only T values, which are scaled by ε^{-1} in the computational algorithm. The base settings of optimal balance are preserved.

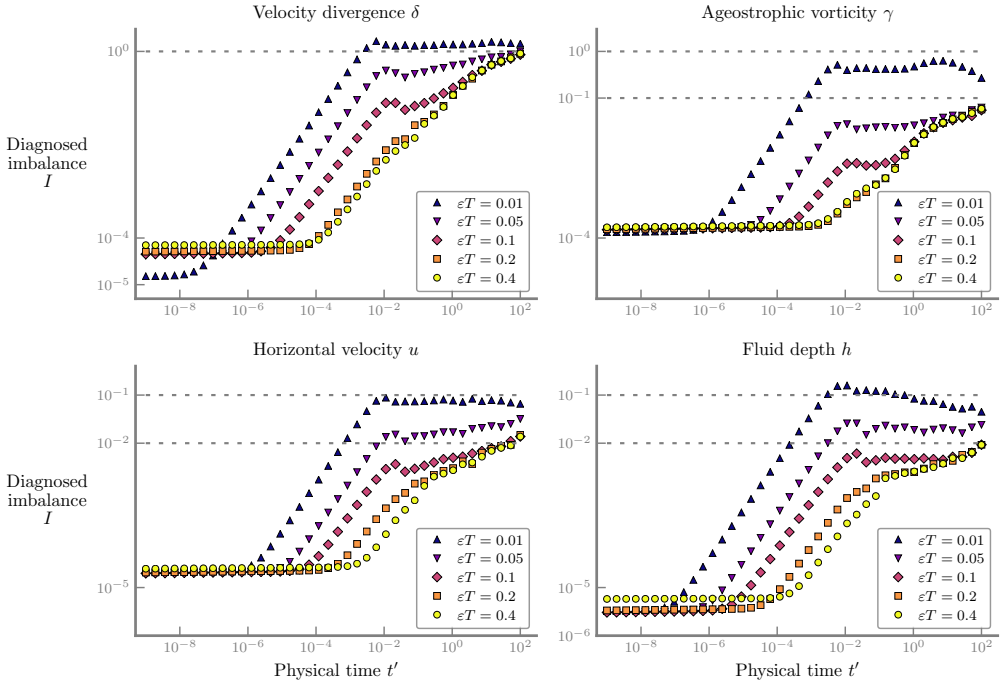

Figure A.13: For quantitative analysis, the physical time length t' is also needed. The diagnosed imbalance is studied as $I(t')$ employing the same setting as in Figure A.12. The t' horizon, here, is scaled by ε^{-1}.

A.5 Convergence tolerance

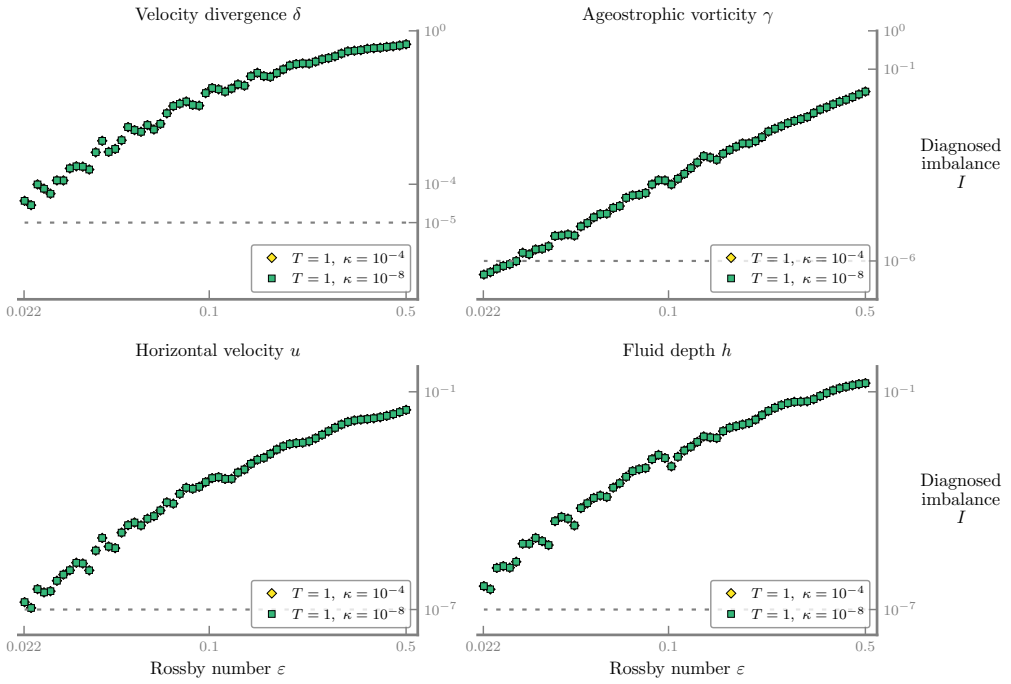

Figure A.14: Smaller convergence tolerance κ values used to terminate the nudging scheme provides comparable quality of balance in the quasi-geostrophic regime. The figure presents the diagnosed imbalance for $\kappa = 10^{-4}$, which takes place in the standard optimal balance setting, and the smaller tolerance $\kappa = 10^{-8}$ with $T = 1$ and $t' = 0.5$.

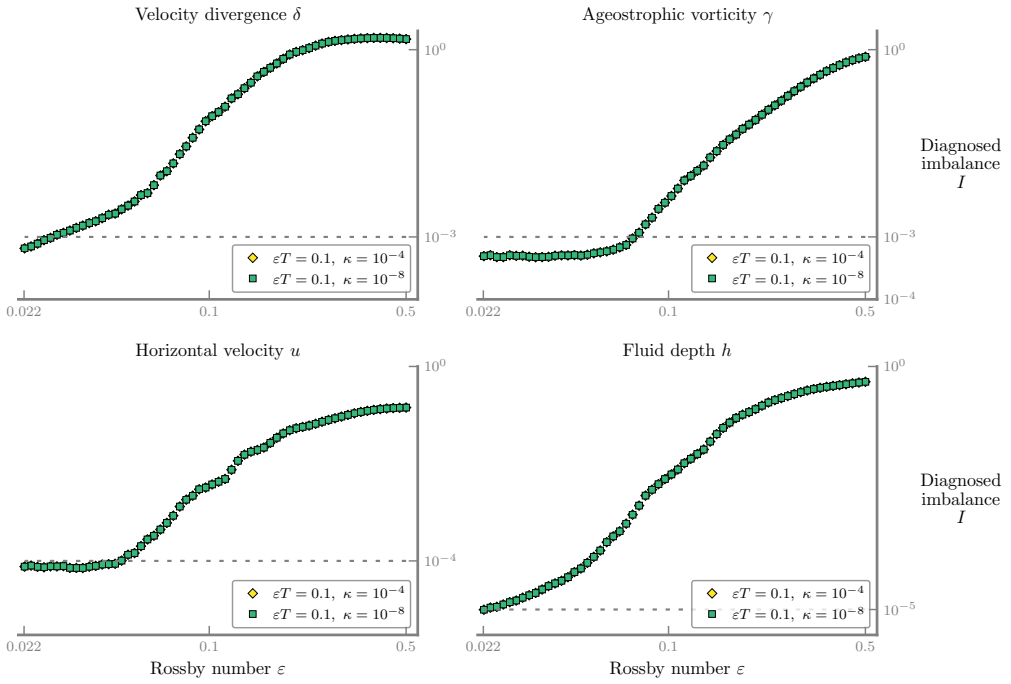

Figure A.15: The κ-sensitivity above is repeated for the semi-geostrophic regime using $T = 0.1/\varepsilon$ and $t' = 0.05/\varepsilon$. The diagnosed imbalance is κ-independent for base point q.

A.6 Linear-end voundary condition

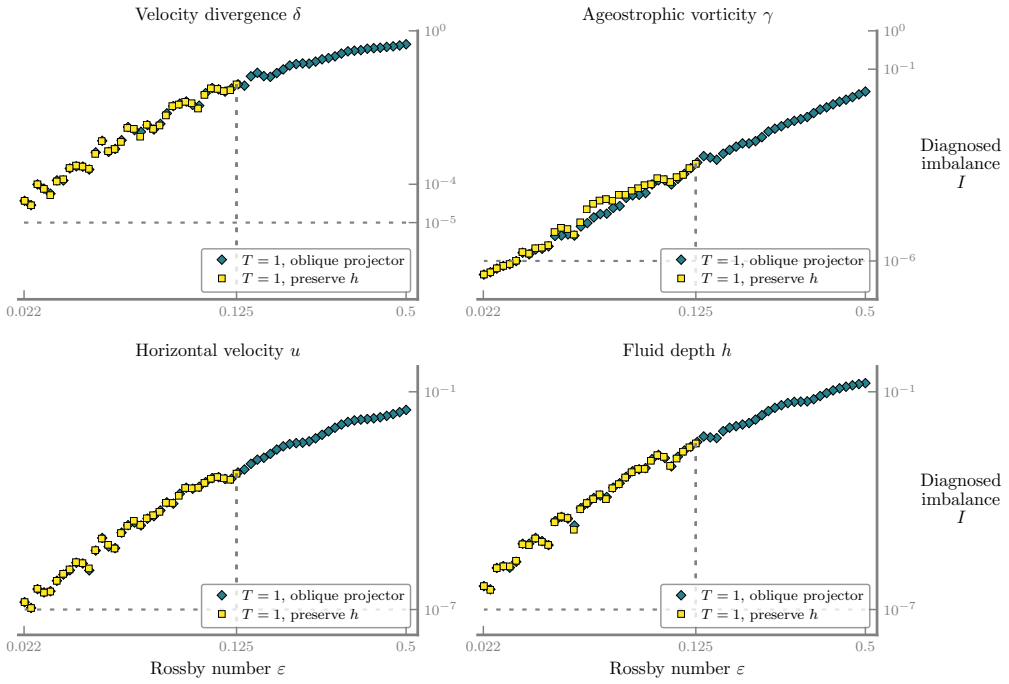

Figure A.16: Preserving h at the linear end provides convergence in the nudging scheme for smaller ε values, and the diagnosed imbalance has comparable quality to that of the oblique projector in the quasi-geostrophic regime. Using the h-preserving projector, the test is run for $T = 1$ and $t' = 0.5$.

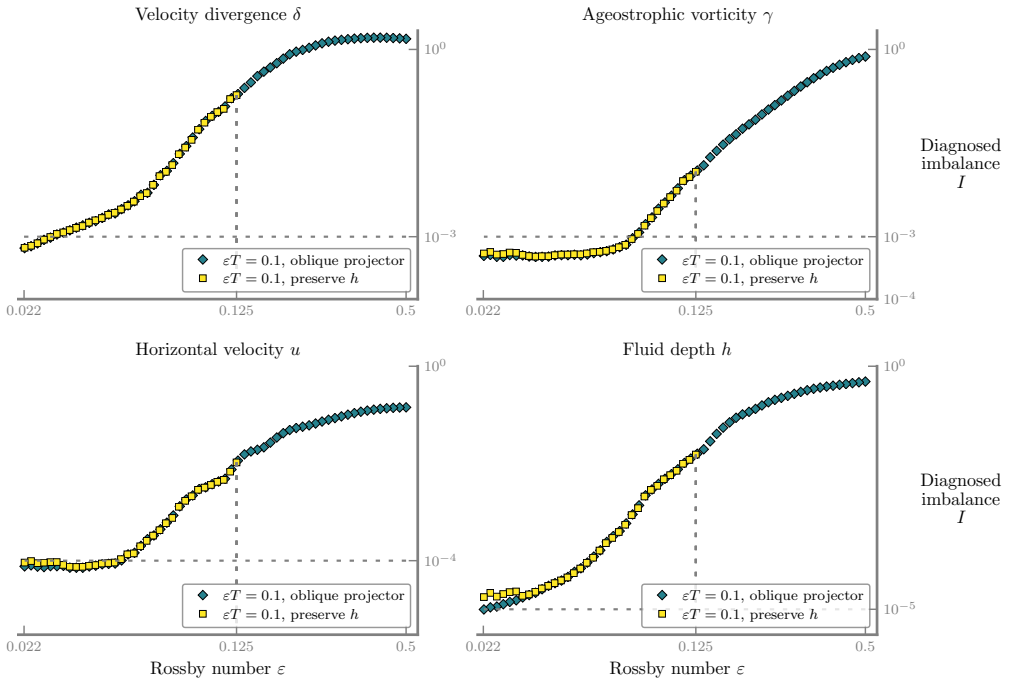

Figure A.17: The above test is executed in the semi-geostrophic regime with $T = 0.1/eps$ and $t' = 0.05/\varepsilon$. The diagnostics of the h-preserving projector have the same quality as other converging projectors over smaller ε-range.

A.7 Initial condition structure

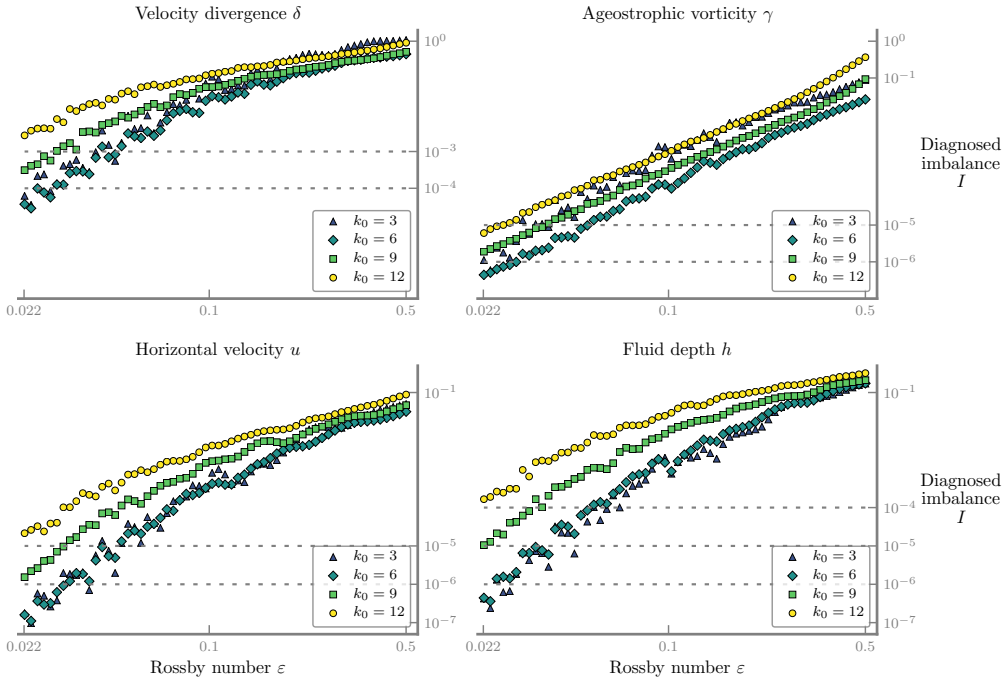

Figure A.18: Initial data holding spectral maximum k_0 at small wave numbers are balanced better for a fixed spectral decay $d = 6$ in the quasi-geostrophic regime. The potential vorticity q is preserved at both end points in the nudging scheme, when $T = 1$ and $t' = 0.5$.

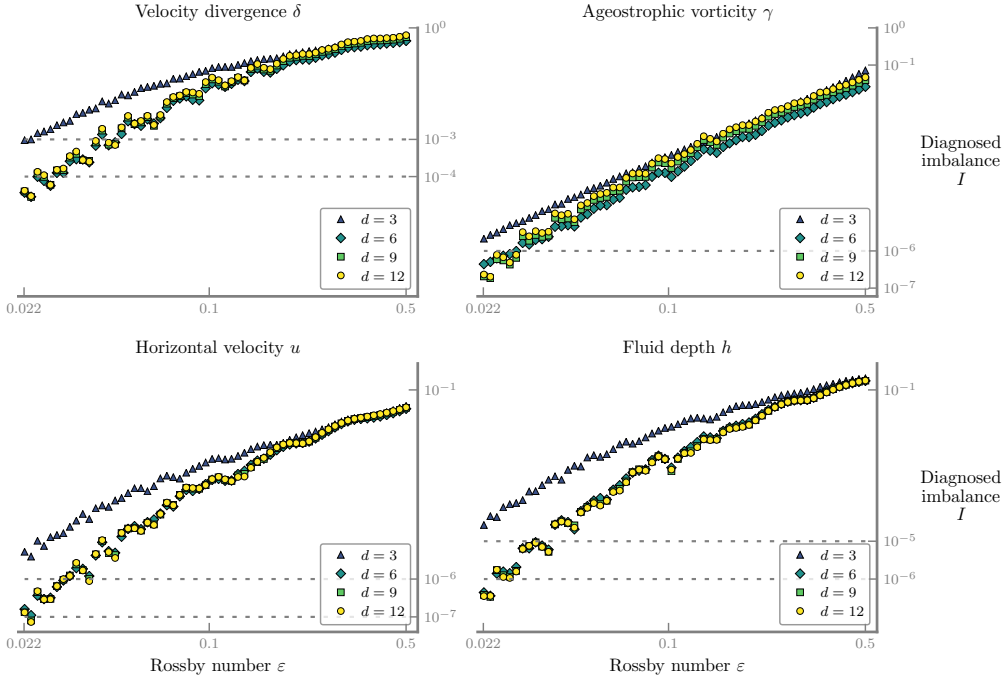

Figure A.19: Initial condition with fixed $k = 6$ and $d > 3$ are balanced with rather close quality in the quasi-geostrophic regime. Optimal balance uses the setting as in Figure A.18.

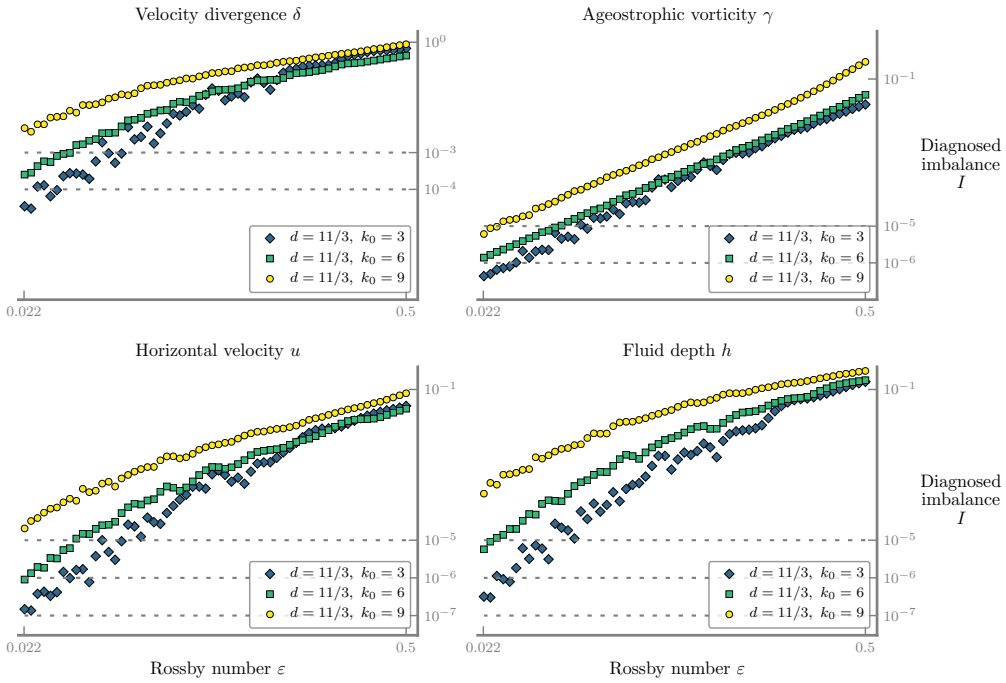

Figure A.20: Turbulent flows with the Kolmogorov $-5/3$ spectrum excites more imbalances, when the spectral maximum k_0 is at larger wave numbers in the quasi-geostrophic regime. The figure demonstrates the diagnosed imbalances of the turbulent flows in Figure A.22. The simulations have ramp time $T = 1$ and forward time $t' = 0.5$.

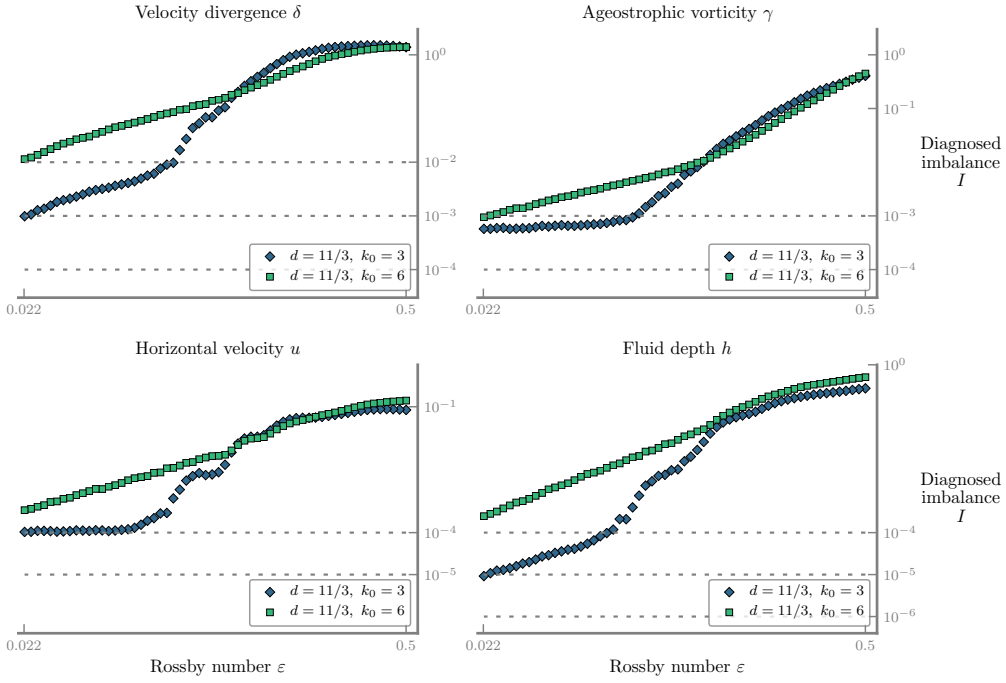

Figure A.21: The test case in Figure A.21 is applied in the semi-geostrophic regime with $T = 0.1/\varepsilon$ and $t' = 0.05/\varepsilon$. The small k_0 values, which gives the effect smaller resolution, increases the quality of balance.

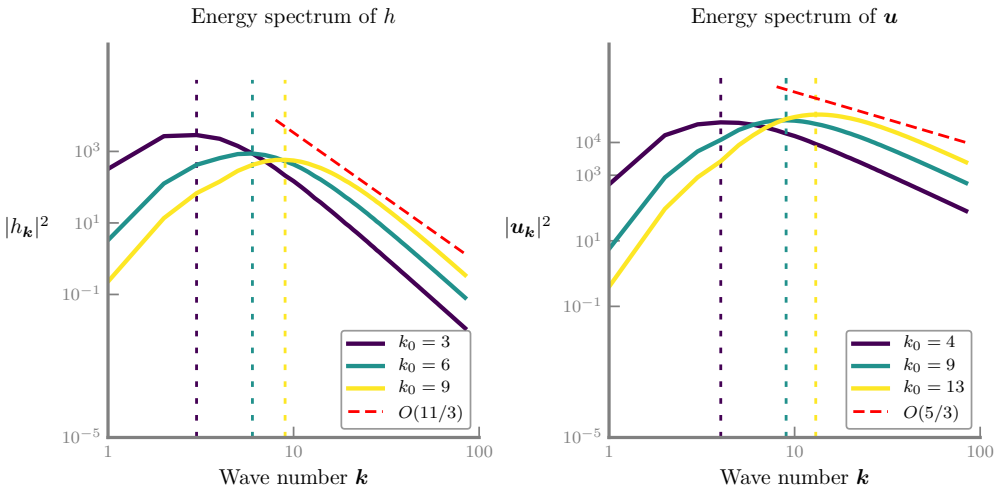

Figure A.22: The height h fields are randomly created when spectral decay is of order $d = 11/3$ and spectral maximum k_0 takes three values (left figure). The corresponding velocity $\boldsymbol{u}$ fields are found by geostrophic balance, and the spectral decay gives the "Kolmogorov -5/3 spectrum" (right figure).

Bibliography

Auroux, D. and Nodet, M. (2012). The back and forth nudging algorithm for data assimilation problems: theoretical results on transport equations. *ESAIM Control Optim. Calc. Var.*, 18(2):318–342.

Bokhove, O. and Shepherd, T. G. (1996). On Hamiltonian balanced dynamics and the slowest invariant manifold. *J. Atmospheric Sci.*, 53(2):276–297.

Calik, M., Oliver, M., and Vasylkevych, S. (2013). Global well-posedness for models of rotating shallow water in semigeostrophic scaling. *Arch. Ration. Mech. An.*, 207:969–990.

Camassa, R. (1995). On the geometry of an atmospheric slow manifold. *Phys. D*, 84(3-4):357–397.

Chandra, J. and Davis, P. W. (1976). Linear generalizations of gronwall's inequality. *Proceedings of the American Mathematical Society*, 60(1):157–160.

Cotter, C. (2013). Data assimilation on the exponentially accurate slow manifold. *Phil. Trans. R. Soc. A*, 371(1991):20120300.

Cotter, C. J. and Reich, S. (2006). Semigeostrophic particle motion and exponentially accurate normal forms. *Multiscale Model. Sim.*, 5(2):476–496.

Danioux, E., Vanneste, J., Klein, P., and Sasaki, H. (2012). Spontaneous inertia-gravity-wave generation by surface-intensified turbulence. *J. Fluid Mech.*, 699:153–173.

Dritschel, D. G. (1988). Contour surgery: A topological reconnection scheme for extended integrations using contour dynamics. *Journal of Computational Physics*, 77(1):240–266.

Dritschel, D. G. and Ambaum, M. H. (1997). A contour-advective semi-Lagrangian numerical algorithm for simulating fine-scale conservative dynamical fields. *Q. J. Roy. Meteor. Soc.*, 123(540):1097–1130.

Dritschel, D. G., Gottwald, G. A., and Oliver, M. (2017). Comparison of variational balance models for the rotating shallow water equations. *Journal of Fluid Mechanics*, 822:689–716.

Dritschel, D. G., Polvani, L. M., and Mohebalhojeh, A. R. (1999). The contour-advective semi-Lagrangian algorithm for the shallow water equations. *Mon. Weather Rev.*, 127(7):1551–1565.

Franzke, C. L. E., Oliver, M., Rademacher, J. D. M., and Badin, G. (2019). Multi-scale methods for geophysical flows. In Eden, C. and Iske, A., editors, *Energy Transfers in Atmosphere and Ocean*, pages 1–51. Springer, Cham.

Gottwald, G., Oliver, M., and Tecu, N. (2007). Long-time accuracy for approximate slow manifolds in a finite-dimensional model of balance. *J. Nonlinear Sci.*, 17(4):283–307.

Gottwald, G. A., Mohamad, H., and Oliver, M. (2017). Optimal balance via adiabatic invariance of approximate slow manifolds. *Multiscale Model. Simul.*, 15(4):1404–1422.

Gottwald, G. A. and Oliver, M. (2014). Slow dynamics via degenerate variational asymptotics. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 470(2170):20140460.

Hunter, J. K. and Nachtergaele, B. (2001). *Applied analysis*. World Scientific Publishing Company.

Kim, Y., Eckermann, S. D., and Chun, H. (2003). An overview of the past, present and future of gravity-wave drag parametrization for numerical climate and weather prediction models. *Atmos. Ocean*, 41(1):65–98.

Lorenz, E. N. (1980). Attractor sets and quasigeostrophic equilibrium. *J. Atmospheric Sci.*, 37(8):1685–1699.

Lorenz, E. N. and Krishnamurthy, V. (1987). On the nonexistence of a slow manifold. *Journal of Atmospheric Sciences*, 44(20):2940 – 2950.

Lynch, P. (2006). *The Emergence of Numerical Weather Prediction: Richardson's Dream*. Cambridge University Press.

MacKay, R. S. (2004). Slow manifolds. In Dauxois, T., Litvak-Hinenzon, A., MacKay, R. S., and Spanoudaki, A., editors, *Energy Localisation and Transfer*, pages 149–192. World Scientific, Singapore.

Masur, G. T. (2022). Optimal balance for rotating shallow-water model. GitHub repository, GitHub, https://github.com/gtmasur/Optimalbalance.

Masur, G. T., Mohamad, H., and Oliver, M. (2022). Quasi-convergence of an implementation of optimal balance by backward-forward nudging. Submitted for publication, arXiv:2206.13068.

Masur, G. T. and Oliver, M. (2020). Optimal balance for rotating shallow water in primitive variables. *Geophysical & Astrophysical Fluid Dynamics*, 114(4-5):429–452.

McIntyre, M. (2015). Dynamical meteorology – balanced flow. In Pyle, J. and Zhang, F., editors, *Encyclopedia of Atmospheric Sciences*, pages 298–303. Academic Press, Oxford, second edition.

Oliver, M. (2006). Variational asymptotics for rotating shallow water near geostrophy: a transformational approach. *Journal of Fluid Mechanics*, 551:197–234.

Poulin, F. J. (2016). PyRsw: Python rotating shallow water model. GitHub repository, GitHub, commit c504456, https://github.com/PyRsw/PyRsw.

Salmon, R. (1983). Practical use of hamilton's principle. *Journal of Fluid Mechanics*, 132:431–444.

Temam, R. and Wirosoetisno, D. (2007). Exponential approximations for the primitive equations of the ocean. *Discrete & Continuous Dynamical Systems - B*, 7(2):425–440.

Temam, R. and Wirosoetisno, D. (2010). Stability of the slow manifold in the primitive equations. *SIAM J. Math. Anal.*, 42:427–458.

Temam, R. and Wirosoetisno, D. (2011). Slow manifolds and invariant sets of the primitive equations. *Journal of the Atmospheric Sciences*, 68(3):675 – 682.

Vanneste, J. (2008). Exponential smallness of inertia-gravity wave generation at small rossby number. *Journal of the Atmospheric Sciences*, 65(5):1622 – 1637.

Vanneste, J. (2013). Balance and spontaneous wave generation in geophysical flows. *Ann. Rev. Fluid Mech.*, 45(1):147–172.

Vanneste, J. and Yavneh, I. (2004). Exponentially small inertia-gravity waves and the breakdown of quasigeostrophic balance. *Journal of the Atmospheric Sciences*, 61(2):211 – 223.

Verhulst, F. (1990). *Nonlinear differential equations and dynamical systems*. Universitext. Springer-Verlag, Berlin. Translated from the Dutch.

Viúdez, A. and Dritschel, D. G. (2004). Optimal potential vorticity balance of geophysical flows. *J. Fluid Mech.*, 521:343–352.

von Storch, J.-S., Badin, G., and Oliver, M. (2019). The interior energy pathway: inertial gravity wave emission by oceanic flows. In Eden, C. and Iske, A., editors, *Energy Transfers in Atmosphere and Ocean*, pages 53–85. Springer, Cham.

Warn, T., Bokhove, O., Shepherd, T. G., and Vallis, G. K. (1995). Rossby number expansions, slaving principles, and balance dynamics. *Quarterly Journal of the Royal Meteorological Society*, 121(523):723–739.

Wirosoetisno, D. (2004). Exponentially accurate balance dynamics. *Advances in Differential Equations*, 9(1-2):177 – 196.