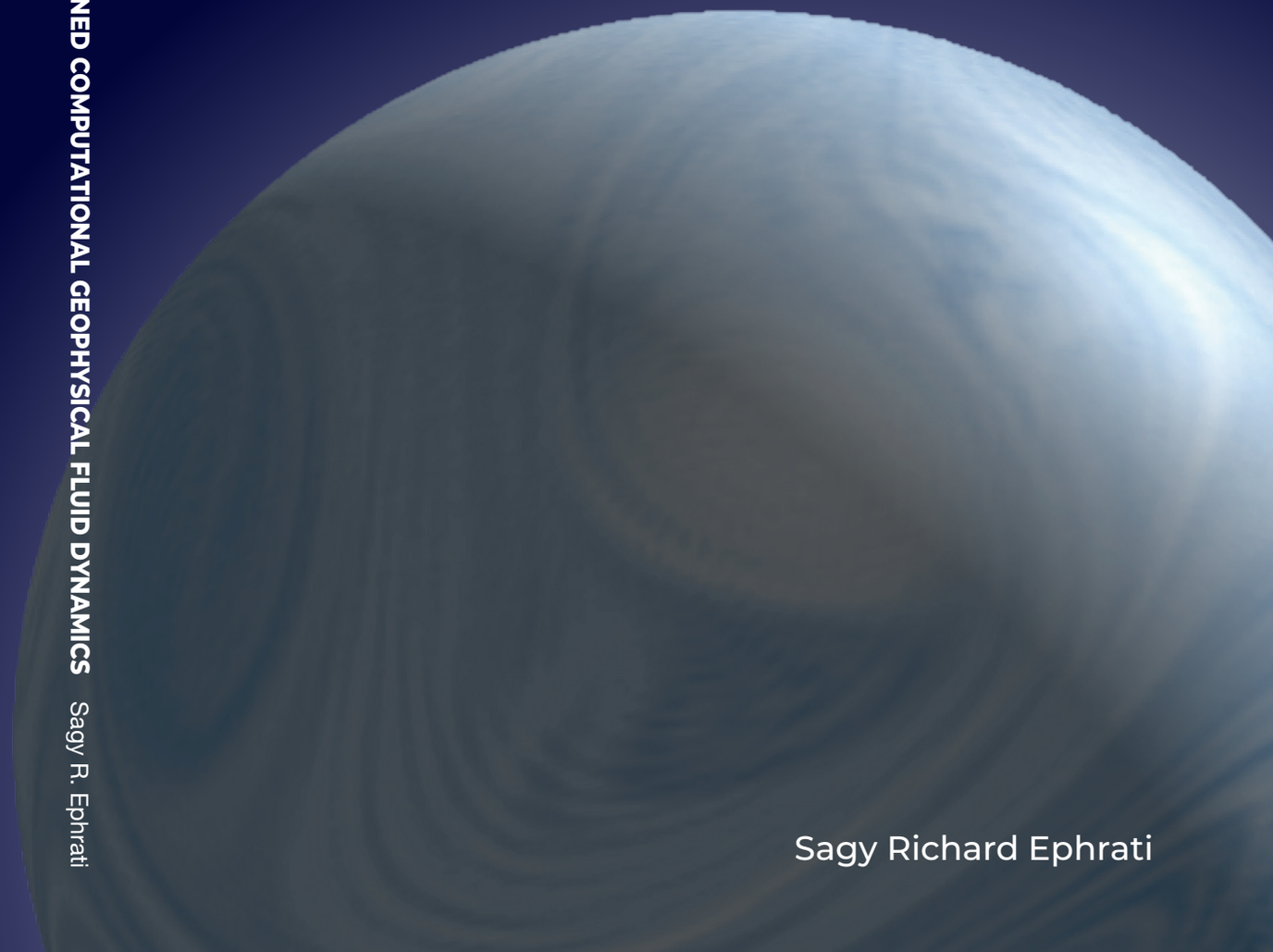
DATA-DRIVEN STOCHASTIC MODELING FOR COARSENED COMPUTATIONAL GEOPHYSICAL FLUID DYNAMICS Sagy R. Ephrati

DATA-DRIVEN STOCHASTIC MODELING FOR COARSENED COMPUTATIONAL GEOPHYSICAL FLUID DYNAMICS



Sagy Richard Ephrati

DATA-DRIVEN STOCHASTIC MODELING FOR COARSENED COMPUTATIONAL GEOPHYSICAL FLUID DYNAMICS

Sagy Ephrati

DATA-DRIVEN STOCHASTIC MODELING FOR COARSENED COMPUTATIONAL GEOPHYSICAL FLUID DYNAMICS

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof. dr. ir. A. Veldkamp,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op maandag 11 september 2023 om 12.45 uur

door

Sagy Richard Ephrati
geboren op 21 september 1996
te Ashkelon, Israël

Dit proefschrift is goedgekeurd door:
de promotor
prof. dr. ir. B.J. Geurts
de co-promotor
dr. ir. P. Cifani

Cover design by Gildeprint.
Printed by Gildeprint.
Lay-out by LaTeX Templates.
ISBN (print): 978-90-365-5702-3
ISBN (digital): 978-90-365-5703-0
URL: https://doi.org/10.3990/1.9789036557030

Promotiecommissie:

Voorzitter / secretaris:	prof. dr. J.N. Kok
Promotor:	prof. dr. ir. B.J. Geurts Universiteit Twente
Co-promotor:	dr. P. Cifani Universiteit Twente
Leden:	prof. dr. H.J. Zwart Universiteit Twente
	prof. dr. ing. B. Rosic Universiteit Twente
	prof. dr. C.J. Cotter Imperial College London
	prof. dr. ir. R.W.C.P. Verstappen Rijksuniversiteit Groningen

The research presented in this thesis was done at the group of Mathematics of Multiscale Modeling and Simulation of the Faculty of Electrical Engineering, Mathematics and Computer Science of the University of Twente in The Netherlands.

Acknowledgments

This thesis concludes my time as a PhD candidate at the University of Twente and an altogether longer period in Enschede. The work performed over the last few years has been made possible to a large extent by a number of people to whom I am grateful.

I want to thank my doctoral advisor Bernard for giving me the opportunity to do research in this challenging but exciting field of mathematics, and for providing me the freedom to pursue the research directions that I found interesting. Your open view toward new ideas is encouraging and your ability to ask precisely the right questions continues to amaze me. I also want to thank Paolo, for being invaluable in carrying out the research of the past years. Your precise comments and ideas as well as your coding wizardry have enabled us to perform better research. I want to thank Darryl as well, for being so enthusiastic about our work and always being open for discussion. A special thanks to my colleagues Arnout and Erwin are also in place. You have turned out to be good travel partners and fellow coffee drinkers, and luckily we share a passion for silly jokes.

I am grateful to the members of the graduation committee, Hans Zwart, Bojana Rosic, Colin Cotter, Roel Verstappen, Paolo Cifani, and Bernard Geurts, for investing time and effort to critically read and review this thesis.

Finally, I want to thank my family and my girlfriend for their unconditional support. Thanks for letting me vent about anything and always providing help when I needed it.

Contents

Chapter 1

Introduction

Geophysical fluid dynamics is a vast field of physics that deals with the study of fluids in the natural environment, of which the atmosphere and oceans are prime examples [173, 159]. The complexity and nonlinearity of the underlying physics make it challenging to develop accurate models that can capture the dynamic behavior of these fluids. Advances in data science and the growing availability of computational resources have opened up new avenues for developing data-driven models that can complement traditional physics-based models [124]. In this thesis, we focus on the development of data-driven models for geophysical fluid dynamics, with a particular emphasis on so-called global basis methods. These methods rely on the decomposition of spatiotemporal data into (fixed) spatial basis functions and corresponding time series, of which only the latter is modeled. This way, one can use techniques from time series analysis and knowledge of stochastic processes to derive fluid models. Our goal is to use this approach to develop models for use on coarse computational grids that can still capture the complex and nonlinear behavior of geophysical fluid flows while being computationally efficient and scalable.

The results presented in this thesis employ either proper orthogonal decomposition basis modes (POD modes, also referred to as empirical orthogonal functions, or EOFs) [13], Fourier basis functions, or spherical harmonic basis functions to decompose the spatio-temporal data into fixed spatial profiles and corresponding time series. The choice of basis depends on the problem under consideration: a periodic domain permits the use of a Fourier basis, the presence of boundaries confines us to the use of a POD basis, and problems on a spherical domain are better served by a spherical harmonic basis. These bases are used in deriving data-driven models for coarse numerical simulations of flows relevant to geophysical fluid dynamics: the shallow water equations, the two-dimensional Euler equations, and two-dimensional Rayleigh-Bénard convection.

The work in this thesis rests on three subjects: geophysical fluid dynamics,

coarse-grid data-driven modeling, and stochastic modeling. Namely, we consider flows relevant to geophysical fluid dynamics and simulate these on coarse computational grids to benefit from reduced computational costs. These simulations are augmented using data, which are incorporated via stochastic forcing with the aim of quantifying uncertainty or modeling subgrid-scale processes. In this section, we introduce these subjects and provide the outline of this thesis.

Geophysical fluid dynamics

Geophysical fluid dynamics (GFD) concerns large-scale flows occurring in natural situations. The most prominent examples of geophysical flows are oceanic and atmospheric flows, of which the study and numerical simulation play a crucial role in, for example, weather prediction. A characteristic feature of fluid flows is the presence of a large range of scales of motion [139]. In GFD, the largest scales of motion are typically of the order of hundreds of kilometers [173] whereas the smallest scales are of the order of millimeters [65]. This vast range of scales suggests that a good fluid model should resolve the largest scales down to the smallest dissipative scales. The presence of energy over decades of length scales implies that a fully resolved numerical prediction of a fluid model containing all scales of motion is computationally intractable. Workable models can subsequently be obtained by simplifying the underlying mathematical model, resulting in a less complete description of the dynamics, or by reducing the spatial and temporal resolution used in the numerical prediction, yielding a less accurate approximation of the dynamics.

Commonly used models in GFD, such as the primitive equations (PE), rotating shallow water equations (RSWE), Euler equations, and quasi-geostrophic equations (QGE) can be derived from the three-dimensional Navier-Stokes equations with an incompressibility condition after a sequence of simplifying assumptions. On the geophysical scale, a fluid is subject to planetary rotation and viscous dissipation effects in the fluid are negligible. This directly leads to the rotating three-dimensional Euler equations. An essential feature of the geophysical domain is its shallowness, meaning that the horizontal length scales are orders of magnitude larger than vertical length scales. As a result, the characteristic vertical velocity is significantly smaller than the characteristic horizontal velocity, allowing for the approximation of the vertical momentum equation as the hydrostatic equation. These assumptions lead to the PE and the RSWE. Subsequently, the QGE are derived from the RSWE by performing a leading-order expansion around geostrophic balance, leading to a simplified flow description in a parameter regime relevant for planetary flows. For a detailed derivation of these equations, we refer to [173, 87, 86, 111].

The RSWE, QGE, and two-dimensional Euler equations are models for two-dimensional geophysical fluid flows. In fact, these models belong to a larger class of fluid models that can be derived from a variational viewpoint, as demonstrated in [87, 86, 111]. As a result, these models contain an infinite number of conservation laws (Casimirs), defined as differentiable functions of the potential vorticity [113]. Specifically, the energy and enstrophy are conserved [155], which has led to the conjecture of Kraichnan of a double cascade in forced two-dimensional turbulence [99]. Here, energy is moved towards the large scales via the inverse energy cascade, and towards the small scales via the enstrophy cascade. This provides a fundamental difference with three-dimensional turbulence, in which energy is moved to successively smaller scales in the energy cascade until it is dissipated through molecular viscosity [139]. Numerical studies have provided evidence for the existence of the double cascade by simulating two-dimensional turbulence at unprecedented numerical resolution [18], or using Casimir-preserving methods [36, 128].

The double cascade is best explained as the triadic interaction of spectral modes of the vorticity. Due to the conservation of energy and enstrophy, energy moves from the median mode to the lower wavenumber mode and enstrophy moves from the median mode to the higher wavenumber mode [173]. Thus, the inverse energy cascade is directed towards large scales, whereas the enstrophy cascade is directed towards small scales. The enstrophy cascade results in the appearance of small scales of motion as thin vorticity filaments. The presence of small-scale features in the flow ensures that resolving all scales of motion in a numerical simulation becomes computationally demanding. On the other hand, the inverse energy cascade is observed as the merging of vortices, leading to large-scale structures only limited by the size of the computational domain and the Rossby deformation radius [102]. These large-scale structures often contain a significant amount of the total energy and dominate the qualitative behavior of the flow. This motivates the use of coarse computational grids, on which the larger motions of the flow are still well-resolved. While this effectively avoids the high computational costs required for resolving all present scales, it introduces the need to model the unresolved smaller scales of motion and their effect on the resolved scales.

Coarse-grid data-driven modeling

There is a significant interest in simplifying complex computational models to achieve predictions and simulations that are accurate enough for a given problem, while also requiring significantly less time and storage resources than the original detailed model. Deriving computational models of dimension orders of magnitude smaller than the full underlying description forms the objective

of reduced-order modeling (ROM) [27, 91]. For example, one might use flow snapshots to identify coherent structures using the proper orthogonal decomposition (POD) algorithm [114] and subsequently solve only the partial differential equations corresponding to the dominant modes. This idea has been successfully applied to many fluid dynamical problems. A non-exhaustive list of examples includes lid-driven cavity flow [30], oceanic boundary layer turbulence [150], and the QGE [140, 129]. Alternatively, one might choose to reduce the dimensionality of the computational model by carrying out numerical simulations on coarse computational grids and including a model term to represent unresolved dynamics. This is the strategy employed in large-eddy simulation (LES) and will be used throughout this thesis.

In LES, the large scales of the turbulent flow are resolved, while the smaller scales are modeled by a subgrid-scale model [147, 74, 92]. The underlying premise of LES is that the largest scales of turbulence are the most important in terms of their impact on the overall flow behavior, and therefore it is necessary to accurately capture their dynamics in order to obtain an accurate prediction of the flow field. Typically, the user imposes a filter width Δ which decomposes the prognostic variables into a large-scale filtered component and a small-scale residual component. This reduces the computational requirements since only the filtered components need to be fully resolved, but at the same time introduces a closure problem arising from the motions of unresolved residual components. Alternatively, the filtering may be defined implicitly as a result of applying discrete operators on coarse computational grids [11, 131]. In either case, the closure problem is often treated by including a subgrid-scale model designed to account for the interactions between the unresolved and resolved scales. A common approach is to use models that a priori prescribe a relation between the resolved and unresolved scales, based on physical or statistical arguments. For instance, the much-used eddy-viscosity models rely on the assumption that turbulent fluctuations on average dissipate energy and may therefore be modeled as a viscous term. Examples include the Smagorinsky model [151], the Vreman model [164], and the dynamic eddy-viscosity model [66].

Although no overall-best LES model has been developed, an abstract subgrid model has been derived that has been shown to be ideal [103]. Here, ideal means that the subgrid model exactly reproduces single-time, multi-point statistics and minimizes the error in instantaneous dynamics. The derived model is defined as the average evolution of the real subgrid force, which is the difference between the filtered evolution of the governing equations and the evolution of the filtered governing equations. In this model, the average is taken over the distribution of all unfiltered fields that correspond to the filtered field. This result implies that the model depends on the considered set

of equations as well as on the chosen filter and discretization. Additionally, this result suggests that one should look for a functional between the state of the system and its expected subgrid force and use this as a subgrid model [53]. With sufficient computational resources, such a relation can be found empirically through the use of data.

Recent advances in data-driven LES have shown the potential of using data to derive models for accurate coarse-grid fluid simulations. The growing availability of computational resources has made it possible to perform larger and more accurate simulations of fluid flows, making high-resolution data accessible from which fluid models may be acquired. Most research on data-driven turbulence modeling has been carried out using machine learning. Examples include its application to channel flow [148, 137], two-dimensional turbulence [119], and three-dimensional forced homogeneous isotropic turbulence [170, 101]. Generally, these methods are found to perform well in terms of kinetic energy spectra and flow statistics. However, a significant amount of data is usually required to determine a machine-learned model and verify its stability [12]. This may be remedied by employing techniques from data assimilation.

Data assimilation is a concept commonly used in the prediction of geophysical fluid dynamics, with the aim of combining numerical model predictions with observations, or data [104]. Data assimilation algorithms use statistical methods to blend the observations and model predictions in a way that optimally balances the uncertainties inherent to both of these aspects [59, 115]. This involves estimating the state of the system at a given time, as well as the uncertainty in that state estimate. For high-dimensional systems, these approaches suffer from the curse of dimensionality and may lead to inadequate results [41, 152]. To alleviate large computational costs, the full data assimilation formulation can be approximated using practical ad hoc approaches built on computational and physical considerations [17]. A conceptually straightforward method to incorporate data into a numerical prediction is through continuous data assimilation (CDA) [49], where real-time observational data is included in the computational model directly as it evolves over time. Particularly, the CDA algorithm proposed in [8, 7] is relevant to the work presented in this thesis.

The algorithm described by [8, 7] incorporates a feedback control term into the governing equations with the goal of nudging the predicted solution towards an observed state. Notably, convergence towards observations has been proven for the two-dimensional Navier-Stokes equations [67, 16] and for two-dimensional Rayleigh-Bénard convection [60, 4] for a range of nudging strengths. Convergence towards the reference has also been proven when the observations are subject to noise [15]. In a similar vein, the continuous-time limit of the 3DVAR data assimilation algorithm [45] has been shown to be a

stochastic differential equation with mean reversal towards a noisy observation [17]. These algorithms are designed for the convergence of the numerical solution towards an observed time-dependent solution. This is contrary to LES, where real-time observations are usually not available and instead, models are based on a priori knowledge of physics or flow statistics. One of the points of attention in this thesis is to derive self-contained models, as used in LES, which are based on a feedback term in the governing equations, as used in CDA. For this purpose, we use stochastic models based on measured flow statistics and investigate how these can be used to quantify uncertainty and act as subgrid-scale models.

Stochastic modeling

Stochasticity has been used in geophysical applications to model uncertainty for at least several decades [25, 134, 82, 105]. In the perspective on stochastic weather and climate models [136], it is argued that any numerical prediction about weather or climate will be subject to uncertainty stemming from two sources. Firstly, the initial conditions used in numerical simulations of geophysical systems are usually not known perfectly. Ensemble forecasts may be used to gauge the influence of the uncertainty, where the initial conditions per ensemble member are slightly varied [135, 90]. Secondly, uncertainty originates from model incompleteness. This is caused by an imperfect or lacking representation of physical processes and by a finite truncation of the model when it is numerically solved. We elaborate on these last two points. Even in the conceptual case where a mathematical model can be resolved without error, it is typically incomplete due to the complexity of real-world flows and involved physical processes. For instance, a perfect weather prediction model would not only need to include terms describing the flow of fluid and temperature but also involve the influence of clouds [52] and moisture [24] on the dynamics, to name but a few examples. These approximations and simplifications introduce model uncertainty and are often represented as stochastic parametrizations. A widely used approach to account for this uncertainty is the Stochastically Perturbed Parametrization Tendencies (SPPT) scheme [25, 33]. Here, the tendencies of parametrized physical processes are stochastically perturbed using spatially and temporally correlated random fields. On the other hand, even if a perfect geophysical fluid model were available, it would still suffer from insufficient computational resources to resolve all relevant dynamical scales. Thus, computationally feasible numerical simulations require the level of detail in the solution to be truncated at a certain length scale, thereby inducing uncertainty through unresolved dynamics. The work presented in this thesis deals with ideal fluid dynamics, i.e., the models used are assumed to be perfect.

Additionally, high-resolution numerical simulations are considered sufficiently accurate to be treated as the 'truth'. This eliminates some of the previously mentioned sources of uncertainty.

Stochastic parametrizations are often employed to represent the effect of unresolved dynamics on the resolved dynamics. This effect is regularly defined using estimates of subgrid processes in terms of resolvable processes [136]. The placement of the stochastic term in the governing equations is not set in stone and is, ultimately, a choice of the modeler. For example, one can distinguish between additive noise and multiplicative noise, both of which will reappear in this thesis. The former is obtained by adding a stochastic term to the governing equations, subject to scaling independent of the state of the system. This type of noise is used, for instance, in the data assimilation algorithms previously mentioned [15, 17] and has also been used in the context of stochastic climate models to model nonlinear interactions of unresolved flow components using Ornstein-Uhlenbeck (OU) processes [116]. Alternatively, one may choose to include a stochastic term with a magnitude multiplied by the value of a resolved component of the flow, which is referred to as multiplicative noise. The SPPT scheme is an example where multiplicative noise is employed. Additionally, it has been shown that multiplicative noise is generally required to maintain conservation laws under stochastic perturbations. The framework of stochastic advection by Lie transport (SALT) [85] may be used to derive stochastic partial differential equations that adhere to the same geometry as the underlying deterministic equations, thus preserving the Casimirs of the original system. Similarly, the method of location uncertainty (LU) [123] can be used to derive stochastic partial differential equations that preserve the kinetic energy of the original deterministic system. These methods are applicable to fluid dynamical systems as well as to finite-dimensional systems of differential equations.

Finite-dimensional systems often serve as a test bed for stochastic parametrizations eventually applied in geophysical fluid simulations. These systems are computationally cheap to evaluate but at the same time mimic certain qualitative behavior also often observed in geophysical flows. For example, the two-scale Lorenz '96 system [110] was designed as a simplified representation of the atmosphere, including variables evolving over different time scales. The model describes the motion of slow and fast variables, where the fast variables are often treated as unresolved scales [6, 167, 47] and parametrized as functions of the slow variables. Another example is the well-known Lorenz '63 system [109], which is obtained by projecting the governing equations of Rayleigh-Bénard convection on a Fourier basis and applying a finite truncation. This low-dimensional system exhibits a strong sensitivity to perturbations in the initial conditions and can be considered a model for chaotic behavior observed in geophysical flows. In particular, different types of multiplicative noise

may already lead to qualitatively different behavior in this system [71]. Alternatively, low-dimensional systems may contain conserved quantities and can therefore be used to study the effect of stochasticity on these quantities and to develop tailored numerical stochastic integrators that respect the conservation laws [112, 38, 32]. All mentioned low-dimensional dynamical systems have the benefit that the 'truth' is often still computationally cheap to obtain. In recent years, this has become increasingly feasible for discretized fluid systems as well.

Recently, high-resolution numerical simulations have been used to derive stochastic parametrizations for fluid simulations. Ideally, the subgrid terms determined from the high-fidelity data should be self-consistent [63], meaning that the obtained model is resolution-independent. As previously mentioned, the type of stochasticity used still remains a choice of the modeler. In simulations of two-dimensional fluid flows on coarse computational grids, there should exist a transfer of energy and enstrophy between the resolved and the unresolved scales [62, 63]. According to these sources, a stochastic subgrid-scale model should be used to represent the transfer of the above-mentioned quantities. This concept is in contrast to SALT and LU, which conserve the enstrophy and the energy, respectively. It is worth noting that the latter two approaches have been used predominantly for uncertainty quantification [40, 143, 144] and data assimilation [41, 145], rather than for subgrid-scale modeling. We will consider both uncertainty quantification and subgrid-scale modeling in this thesis.

Thesis outline

This thesis is structured as follows. In Chapter 2, we describe how to compute the perfect subgrid measurements on coarse computational grids, based on high-resolution data. This is demonstrated for the one-dimensional shallow water equations using two computational methods. The data is decomposed into spatial profiles, being the proper orthogonal decomposition modes (POD modes, also referred to as empirical orthogonal functions or EOFs), which illustrates how the subgrid measurements are dependent on the adopted discretization method at severe coarsening. We study how the high-fidelity result may be approximated using prescribed reduced-order corrections, and how the quality of the correction depends on the accuracy of the adopted discretization. Finally, we show the robustness of the correction under perturbations of the initial condition.

Chapter 3 concerns a study of SALT applied to the two-dimensional Euler equations on the unit square, extending the work of [40]. Here, we estimate the subgrid data as the differences between Lagrangian trajectories measured on a fine-grid solution and a filtered version thereof. Subsequently, the POD

algorithm is applied to obtain fixed spatial profiles and corresponding time series. The novelty of this study is the use of time series data to derive stochastic models. We compare three noise models: Gaussian noise, noise obtained from estimated probability density functions, and time-correlated noise. The latter two use more information from the measurements to mimic the statistical properties of the original time series and are found to yield a smaller spread of the stochastic ensemble, compared to Gaussian noise.

In Chapter 4, a generalization of the approach of Chapter 2 is presented. Instead of using exact pre-computed measurements to define a reduced-order correction, we use high-resolution simulation measurements to tune parameters for a dynamic forcing procedure. The resulting forcing can be applied deterministically or stochastically and enters the governing equations via a feedback term defined at the level of the basis coefficients. The forcing is used to augment coarse numerical simulations in order to accurately represent the statistically steady state from which the measurements were taken. This methodology is applied to the two-dimensional Euler equations on the sphere, employing a spherical harmonic basis to decompose the high-resolution signal into spatial basis functions and corresponding temporal components. The result is a data-driven deterministic or stochastic model defined independently of the adopted discretization or numerical resolution. We show that the proposed forcing leads to a coarse numerical simulation in which the kinetic energy spectrum of the reference solution is reproduced, at different coarse resolutions. Additionally, it is shown that the method leads to stable and accurate large-scale dynamics.

In Chapter 5, we apply the methodology presented in Chapter 4 to two-dimensional Rayleigh-Bénard convection. Effects of buoyancy and wall-bounded turbulence in the flow make this a challenging test case for computational models. We show that using a Fourier basis along the periodic direction of the domain allows for the use of the stochastic forcing in the same manner as in Chapter 4. Additionally, a constraint on the Fourier coefficients of the solution is introduced to ensure that a prescribed heat flux is approximated. As before, the model is shown to lead to stable numerical simulations and accurate kinetic energy spectra on coarse computational grids and is further assessed using flow statistics.

A summary of the results presented in this thesis and directions for future research are provided in Chapter 6.

Chapter 2

Computational modeling for high-fidelity coarsening of shallow water equations based on subgrid data

2.1 Introduction

The nonlinear nature of models in fluid dynamics causes small scale and large scale flow features to interact with each other. This implies that one would need to resolve the entire range of scales from the largest down to the smallest dynamically relevant Kolmogorov scale present in the particular problem, in order to have a good fluid-mechanical model. In geophysical fluid dynamics, typical largest length scales are in the order of hundreds of kilometres. This means that solving the entire range of scales down to the Kolmogorov length scale is by far too expensive for modern-day high performance computing. Any feasible approach will hence necessarily have to imply simplifications, either in the completeness of the mathematical model or in the spatial and temporal resolution with which the dynamics is approximated, or both. In this chapter we will work out an offline/online approach in which we use explicit knowledge of the smallest scale dynamics obtained from prior *offline* fully resolved simulations, in order to arrive at an *online* computational high-fidelity coarsening. This approach is illustrated for the shallow water equations in which we opt for an empirical orthogonal function (EOF) representation of the corresponding subgrid forcing. The accuracy and efficiency we find for this approach and

The material in this chapter was published in the journal *Multiscale Modeling and Simulation*, see [57].

the rate with which the EOF representation converges in selected cases, establishes the feasibility of this computational model reduction for shallow water models.

There is a strong interest into the coarsening of detailed computational models in order to reach predictions and simulations that are on the one hand of sufficient accuracy for a particular problem, while requiring considerably less effort in terms of time and storage compared to the underlying detailed description [74]. These problems are at the core of the field of 'Reduced Order Modeling' (ROM) [27]. A prominent example is so-called large-eddy simulation (LES) in which the spatially filtered Navier-Stokes equations form the point of departure for large-scale models that can handle turbulent flow at high Reynolds numbers [147]. The filtering of the nonlinear terms in the Navier-Stokes equations introduces a closure problem and additional high-pass smoothing associated with the spatial discretization method [72]. These aspects are typically addressed by the introduction of a subgrid scale model to represent the influence of the smaller scale dynamics on the retained resolved scales. The design of good subgrid parameterizations is challenging and LES models based on physical arguments are often based on a crude approximation of the actual subgrid dynamics. Moreover, artificial dissipation introduced by the truncation error of the coarse-grid PDE may be dominant, leading to an over-dissipative system.

In this chapter, we approach the problem of achieving accurate and effective coarsened flow models differently. Here, by introducing an explicit subgrid scale forcing extracted from a previously conducted direct numerical simulation (DNS) of the same problem, we account for the accumulated effects of the unresolved dynamics. Using high-resolution data to find subgrid parametrizations has been applied to, e.g., oceanic flows [19] and atmospheric processes [133]. By adding a corresponding correction term to the governing equations, an alternative representation of the small scale dynamics is obtained. This chapter is strongly motivated by the seminal work that led to the so called stochastic advection by Lie transport (SALT) approach and pursues the path of introducing tailored forcing to the equations in order to account for missing dynamics in the coarsened solution. In [85] a stochastic variational principle was introduced to derive equations in continuum mechanics in such a way that the geometric structure corresponding to these equations remains the same. The SALT method has important applications in geophysical fluid dynamics, for instance to address the fundamental problem of appropriately representing measurement error and uncertainty due to neglected physical effects, spatial and temporal coarsening of the dynamics, and incompleteness of the mathematical model. In [85] the subgrid dynamics are computed from the difference between fully resolved and filtered Lagrangian trajectories. Here we construct

a coarse-grid correction from the difference between the solution of the fine PDE and the coarse PDE at given time instances. The latter allows to take into account not only the effect of the subgrid scales but also the numerical error.

Analogously to [85], in this work we represent the coarse-grid correction by means of empirical orthogonal function (EOF) analysis [114]. The subgrid term structure is thus captured by the solution eigenvectors to the EOF problem, henceforth called $\boldsymbol{\xi}_i$. Differently from [85], no stochasticity is introduced here into the model and the effect of the coarsening is modelled as a deterministic forcing.

The technique of EOF analysis is well-known in atmospheric and oceanic dynamics, and is often called proper orthogonal decomposition (POD) in the context of fluid dynamics [156]. EOF analysis has been applied in atmospheric sciences since the 1950s, for instance in [79], [108], with the purpose of identifying coherent structures in the solution and reducing dimensionality of weather and climate systems. Examples of applications in fluid dynamics include the analysis of canonical problems in turbulence such as the lid-driven cavity [30], the turbulent jet [125] and channel flow [130]. Instead of using the EOF method to analyze flow structures, we apply it to construct a basis for the coarse-grid correction. We illustrate the method with shallow water flow under the influence of external agitation, complementing the earlier work on the Euler equations in periodic domains [40].

By construction the coarse-grid correction is dependent on the adopted numerical method. Hence, we will investigate two different methods for solving the shallow water equations and compare the type and size of EOF corrections needed to improve a coarse simulation. Moreover, the convergence of the corrections upon increasing the number of EOFs will be investigated. In the SALT approach, one investigates differences only in the velocity variables, since one introduces stochasticity in the vector fields that carry the flow properties. Results of [86] imply that for this situation, obtaining the $\boldsymbol{\xi}_i$ in one dimension and extending their domain to two dimensions corresponds to $\boldsymbol{\xi}_i$ obtained from the two-dimensional translation-invariant setting.

The following is an overview of the key results discussed in this chapter:

- A subgrid data measurement procedure is presented, applicable to any set of PDEs, here applied to the shallow water equations. These measurements are extracted from an *offline* computation of the fine and coarse PDE.

- Subgrid data is measured for two test cases which are both performed using a finite difference discretization and a finite element discretization. The test cases feature a submerged ridge as bathymetry and include

constant external forcing (first test case) and periodic external forcing (second test case). The subgrid data are decomposed into EOFs and their corresponding time series.

- The level of approximation of the original dataset when applying different numbers of EOFs is investigated for the test case with external forcing. A coarse numerical solution with zero error is obtained when the full set of EOFs is used. Truncating the reconstructed correction term to a subset of the EOFs significantly reduces the error on coarse computational grids, independent of the used numerical method.

- A significant error reduction is obtained when applying the developed reduced-order correction method to the same test case with different initial conditions. This indicates that the measured temporal coefficients tolerate some level of approximation without significant loss of accuracy.

The chapter is organised as follows. In section 2.2, we will introduce the governing equations as well as the discretisation methods that will be used to simulate the governing equations. Section 2.3 describes the measuring procedure and the reduced-order model. In section 2.4 we investigate the convergence of the EOF decomposition of the coarse-grid correction for two test cases: a steady flow (subsection 2.4.1) and a periodically forced flow (subsection 2.4.2) over a bathymetry represented by a Gaussian profile. In section 2.5 the developed reduced-order model is applied to the test cases of section 2.4. In particular, a range of grid resolutions is investigated as well as the behavior of the model for a varying number of EOFs. Finally, the measured corrections are applied to the same problem with perturbed initial conditions (subsection 2.5.3) and accuracy in the prediction of long-time averages is investigated (subsection 2.5.4). In section 2.6 we conclude the chapter and formulate future challenges in the outlook.

2.2 Governing equations and numerical methods

The model that is central to this work is the shallow water (SW) model. The SW equations, also called the Saint-Venant equations, describe the behaviour of a fluid in a shallow channel with a free surface and bottom topography. This model can be derived by vertically integrating the incompressible free surface Euler-Boussinesq equations over the shallow domain in the small aspect ratio limit, as is demonstrated in [86]. The SW model is nonlinear and consists of two coupled equations. The first equation describes the evolution of the velocity u and the second equation is the continuity equation that describes the evolution of the water depth η. The total depth is the difference between the free surface

elevation ζ and the bottom topography (or bathymetry) b, hence $\zeta = \eta - b$. Additionally, we will consider external forcing and damping of the velocity. In one spatial dimension the SW model with forcing and damping is given by

$$\begin{aligned} u_t + \frac{1}{2}(uu)_x + \frac{1}{\mathrm{Fr}^2}(\eta - b)_x &= a(t) - ru, \\ \eta_t + (u\eta)_x &= 0. \end{aligned} \tag{2.1}$$

The right-hand side of the momentum equation contains a time-dependent forcing term $a(t)$ and a damping coefficient r which induces damping proportional to the velocity. Here Fr is the Froude number, which is defined as the ratio between the typical velocity scale U and the fastest gravity wave $\sqrt{gH}$, where H is the typical depth and g is the gravitational acceleration. For the study of this chapter the one-dimensional model is a suitable formulation, combining low computational cost with a truthful representation of the underlying dynamics. In fact, this model is directly related to the two-dimensional rotating shallow water equations, which form a convenient model in geophysical fluid dynamics. It is known as the simplest model that incorporates the interaction between Rossby waves and gravity waves at geostrophic balance [173].

In the following we provide a description of the two numerical methods that are used in this study. The two corresponding methods are based on finite difference (FD) and a finite element (FE) discretization methods used for solving nonlinear PDEs and are employed here (i) to investigate convergence of the obtained numerical solutions and (ii) subgrid measurements, and (iii) to show the application of reduced-order corrections.

The main difference between the methods is that the FD method solves the momentum equation with first-order accuracy and the continuity equation with second-order accuracy, whereas the FE method solves these equations with second and first-order accuracy, respectively. The main benefit of the FD method is its simplicity and low computational cost, while the FE method is easily extendable to problems in more dimensions and on complex geometry. The approach demonstrated in this chapter is general and extendable to different numerical methods other than those analysed here.

The time integration is the same for both discretizations and is given by a fourth order Runge-Kutta method (RK4). The time-step is specified to satisfy numerical stability, which yields temporal discretization errors that are considerably smaller than the spatial discretization errors.

2.2.1 Collocated finite difference discretization (FD)

The finite difference discretization is based on a collocated arrangement of the discrete variables (u_i, η_i) approximating the exact solution $(u(x_i), \eta(x_i))$ at the

grid nodes x_i with i running from 0 to N, corresponding to an Arakawa A-grid [5]. The first-order upwind method has been employed for the discretization of the convection of momentum. This provides numerical stability of the resulting discrete hyperbolic partial differential equation. The pressure term and the continuity equation are discretized using second-order central differences. Conservation of mass is ensured by discretizing the conservative form of the continuity equation. The finite difference discretization is summarized as

$$\begin{aligned} \frac{1}{2}(uu)_x\big|_{x_i} = (uu_x)_{x_i} &\approx \begin{cases} u_i\left(u_i - u_{i-1}\right)/\Delta x & \text{if } u_i > 0, \\ u_i\left(u_{i+1} - u_i\right)/\Delta x & \text{if } u_i < 0, \end{cases} \\ (\eta - b)_x\big|_{x_i} &\approx \left(\eta_{i+1} - b_{i+1} - \eta_{i-1} + b_{i-1}\right)/\left(2\Delta x\right), \\ (u\eta)_x\big|_{x_i} &\approx \left(u_{i+1}\eta_{i+1} - u_{i-1}\eta_{i-1}\right)/\left(2\Delta x\right), \end{aligned} \tag{2.2}$$

with Δx the grid size. No modification of the numerical scheme (2.2) is required at the boundary, since periodic conditions are imposed. The discretized momentum equation has a formal order of accuracy of one, due to the chosen discretization of the convective term. The continuity equation is second-order accurate.

2.2.2 Compatible finite element discretization (FE)

The finite element discretization is given by a mixed compatible element method, which can be seen as a finite element version of a finite difference discretization based on an Arakawa C grid [5]. It has been proposed as a discretization method for numerical weather prediction in [44, 132], as it inherits the desirable properties of the C-grid – such as exact steady geostrophically balanced states for the linearized shallow water equations. A description of the method can be found in appendix A.

A pair of compatible spaces for u and η is given, e.g., by

$$\mathbb{V}_u = CG_k(\Omega), \qquad \mathbb{V}_\eta = DG_{k-1}(\Omega), \tag{2.3}$$

where $CG_k(\Omega)$ denotes the k^{th} polynomial order continuous Galerkin space and $DG_{k-1}(\Omega)$ the $(k-1)^{th}$ polynomial order discontinuous Galerkin space.

The governing shallow water equations (2.1) are discretized such that the divergence in the continuity equation is considered strongly, while the gradient

in the momentum equation is imposed weakly, leading to the mixed formulation

$$\langle w, u_t \rangle - \left\langle w_x, \frac{1}{2}u^2 + \frac{1}{\mathrm{Fr}^2}(\eta - b) \right\rangle = 0 \qquad \forall w \in \mathbb{V}_u, \tag{2.4}$$

$$\eta_t + F_x = 0, \tag{2.5}$$

where $\langle ., . \rangle$ denotes the L^2 inner product, and the flux F in (2.5) is given by the L^2-projection of ηu into the velocity space, i.e.,

$$\langle w, F - \eta u \rangle = 0 \qquad \forall w \in \mathbb{V}_u. \tag{2.6}$$

The above space discretization conserves mass locally as well as a discrete energy globally (for details, see e.g. [121]). In this chapter, we consider the lowest polynomial order $k = 1$ for this setup.

2.3 Data measurements and processing

This section describes the procedure of measuring the subgrid data and subsequently constructing a reduced-order correction based on these measurements. Given a truth u_{truth} and a coarse-grid result u_{sim}, we construct a function $f(x,t)$ via

$$u_{\text{truth}}(x,t) - u_{\text{sim}}(x,t) = f(x,t) = \bar{f}(x) + f'(x,t) \tag{2.7}$$

where the measurements are decomposed into a mean $\bar{f}(x)$, which will be referred to as $\boldsymbol{\xi}_0(x)$, and a fluctuating component $f'(x,t)$. The EOF decomposition is applied to the fluctuating component f', which is assumed to be stationary in the average or statistical sense. Specifically, on a numerical grid consisting of N cells, this algorithm yields N eigenmodes $\boldsymbol{\xi}_i(x)$ with corresponding temporal coefficients $\alpha_i(t)$:

$$f'(x,t) = \sum_{i=1}^{N} \alpha_i(t)\boldsymbol{\xi}_i(x). \tag{2.8}$$

The measuring procedure described below is such that it identifies the features missing from a (coarse) numerical solution. The constructed $f(x,t)$ can be introduced into coarse simulations as a forcing or correction term, thus correcting the numerical solution to match the reference truth. In the ideal setting, all data is available and the numerical solution can be corrected so that it perfectly recovers the truth on the coarse grid. However, this is typically not feasible in practice due to large data storage requirements. The EOF approach allows

for an optimal approximation of the entire data set using a finite number of modes.

This section presents this methodology as follows. The subgrid term measuring procedure is given in 2.3.1 and section 2.3.2 briefly summarizes the EOF algorithm. Subsequently, the reduced-order correction is detailed in section 2.3.3.

2.3.1 Subgrid term measurement procedure

A simulation, which will correspond to a dataset, runs from time $t = 0$ to $t = T$. The measuring intervals are indicated by Δt_M and are such that $N_M \Delta t_M = T$, where N_M denotes the number of measuring intervals. For consistency, the coarse-grid time step Δt is set to be equal to Δt_M. The measurements comprise of the difference of the evolution of the true velocity and free surface height $(u_{\text{truth}}, \eta_{\text{truth}})$ and their corresponding coarse-grid numerical solution $(u_{\text{sim}}, \eta_{\text{sim}})$, as in equation (2.7). The truth is calculated by performing a numerical simulation on a very fine grid. Throughout this study a grid consisting of 512 computational cells is considered sufficiently fine to accurately resolve all scales of motion. This has been verified by conducting a grid refinement study.

The numerical coarse grid solution $(u_{\text{sim}}, \eta_{\text{sim}})$ is the quantity that we wish to improve. Since the coarse grid solution and the truth are defined on different computational grids, comparing the two solutions is done by restricting $(u_{\text{truth}}, \eta_{\text{truth}})$ to the grid on which $(u_{\text{sim}}, \eta_{\text{sim}})$ is defined. This is carried out by introducing a restriction operator R, here chosen to be equal to the injection of fine-grid values onto coarse-grid values.

The subgrid term defined for the velocity and free surface height will be denoted by $\mathbf{f}(x,t) = (f_u(x,t), f_\eta(x,t))$. Let us assume u_{truth} at time t_0 to be known. The subgrid correction over a time-interval $[t_0, t_0 + \Delta t_M]$ is estimated by applying the following procedure.

1. Inject the truth to the coarse grid at $t = t_0$ and set $u_{\text{sim}}(x, t_0) = R u_{\text{truth}}(x, t_0)$ and $\eta_{\text{sim}}(x, t_0) = R \eta_{\text{truth}}(x, t_0)$, with R a coarse-graining operator.

2. Integrate the fine and coarse grid solution from $t = t_0$ to $t = t_0 + \Delta t_M$.

3. Evaluate

$$\begin{aligned} f_u(x, t_0 + \Delta t_M) &= R u_{\text{truth}}(x, t_0 + \Delta t_M) - u_{\text{sim}}(x, t_0 + \Delta t_M) \\ &= R \left(\int_{t_0}^{t_0 + \Delta t_M} u_{t,\text{truth}} \, \mathrm{d}t \right) - \int_{t_0}^{t_0 + \Delta t_M} u_{t,\text{sim}} \, \mathrm{d}t, \end{aligned} \tag{2.9}$$

and analogously for $f_\eta(x, t_0 + \Delta t_M)$. These measurements are done offline. In the next subsections we describe how the measurements are processed and subsequently applied *online* as a correction term in coarse numerical simulations.

2.3.2 Empirical Orthogonal Function Analysis

The measurements $\mathbf{f}$ are stored in a matrix $\mathbf{V}^N \in \mathbb{R}^{M\times N}$, where M is the number of coarse grid points and N is the number of measurements. The entry $(\mathbf{V}^N)_{ij}$ corresponds to the subgrid difference at grid point x_i at the j^{th} measuring instant. The time-mean from M time series is subtracted from the matrix $(\mathbf{V}^N) \in \mathbb{R}^{M\times N}$ to form the anomaly matrix $\mathbf{A}$, whose rows have zero mean. The time-mean is the spatial profile previously introduced as $\boldsymbol{\xi}_0$. One would then compute the covariance matrix $\mathbf{R} = \mathbf{A}\mathbf{A}^T$ and solve the eigenvalue problem

$$\mathbf{R}\mathbf{C} = \mathbf{C}\boldsymbol{\Lambda}, \tag{2.10}$$

where the columns of $\mathbf{C}$ are the eigenvectors $\boldsymbol{\xi}_i$ (EOFs) and the eigenvalues (EOF variances) are on the diagonal of $\boldsymbol{\Lambda}$. A drawback of this method is that computing the covariance matrix becomes very numerically expensive as the amount of stored data rapidly increases with the number of snapshots. This can be dealt with by computing the SVD of the anomaly matrix [50, 158]. Subtituting $\mathbf{A} = U\Sigma V^T$ into the the definition of the covariance matrix yields

$$\mathbf{R} = \mathbf{A}\mathbf{A}^T = U\Sigma\Sigma^T U^T. \tag{2.11}$$

Comparing equations (2.10) and (2.11), it is observed that $\mathbf{C} = U$ and $\boldsymbol{\Lambda} = \Sigma\Sigma^T$.

An insufficient number of measurements leads to statistical error in the computation of the covariance matrix. In this study, it is assumed a sufficient number of measurements is available for the EOF algorithm.

2.3.3 Defining a reduced-order correction for the SWE

Having introduced the measurement procedure and the EOF algorithm, we can now define a correction term based on the decomposed measurements. This term is included in the numerical simulation such that, if all available data is used, the corrected coarse solution would equal the truth on the coarse grid. We denote the EOFs for the velocity and free surface height by $\xi_{i,u}$ and $\xi_{i,\eta}$, respectively, with corresponding time series $\alpha_{i,u}$ and $\alpha_{i,\eta}$. The correction based

on n EOFs is denoted by $(f_{n,u}(x,t), f_{n,\eta}(x,t))$ for u and η individually, where

$$\begin{aligned} f_{n,u}(x,t) &= \xi_{0,u}(x,t) + \sum_{i=1}^{n} \alpha_{i,u}(t)\xi_{i,u}(x), \\ f_{n,\eta}(x,t) &= \xi_{0,\eta}(x,t) + \sum_{i=1}^{n} \alpha_{i,\eta}(t)\xi_{i,\eta}(x). \end{aligned} \tag{2.12}$$

For an explicit Euler scheme, the reduced-order model is formulated as follows:

$$\begin{aligned} u^{k+1} &= u^k + \Delta t \mathcal{L}(u^k, \eta^k) + f_{n,u}^{k+1}, \\ \eta^{k+1} &= \eta^k + \Delta t \mathcal{D}(u^k, \eta^k) + f_{n,\eta}^{k+1}, \end{aligned} \tag{2.13}$$

where k is the time level, $\mathcal{L}$ is the discrete differential operator of $-\frac{1}{2}(uu)_x - \frac{1}{\mathrm{Fr}^2}(\eta - b)_x + a(t) - ru$, $\mathcal{D}$ is the discrete divergence $(u\eta)_x$ and f_n^{k+1} is the correction measured at time $k+1$ over an interval Δt and decomposed into n EOFs. Extension of (2.13) to RK4 is straightforward.

Finally, the temporal coefficients are obtained by projecting the governing equations on the spatial structures. Given an inner product $\langle \cdot, \cdot \rangle$, $\alpha_i(t)$ can be determined from $\langle f'(x,t), \xi_i(x) \rangle$ when the decomposition (2.8) is used. In matrix notation, this is given by

$$\boldsymbol{\alpha} = \mathbf{AC}. \tag{2.14}$$

The algorithm for computing and applying the subgrid corrections is summarized as follows:

1. The difference between the reference (fine-grid) evolution and coarse-grid evolution are measured, as per (2.9).
2. The measurements are stored in a matrix which serves as input for the EOF algorithm.
3. An n^{th}-order correction term is constructed by considering the time-mean and the first n EOFs, by means of (2.12).
4. The corrections are applied to the coarse numerical solution after completing a time step, as in (2.13).

2.4 Convergence analysis of EOFs of subgrid data

In this section, we present the results of simulations using the two numerical methods for the shallow water equations introduced in Section 2.2. A comparison is performed for two test cases for which the subgrid corrections on several

coarse grids are determined. The bathymetry for both test cases is defined by $b(x) = 1 - A \exp\left(\frac{-(x-0.5L_x)^2}{B^2}\right)$. The latter describes a submerged ridge of height A and width B. The values for A and B are 0.01 and 0.15, respectively. The initial conditions are $u(x,0) = 0$ and $\eta(x,0) = b(x)$.

We force the flow differently in both tests. The first case uses a constant forcing, modeling a fixed 'tilting' of the entire domain. Damping is added to keep the flow bounded. For the second case a time-periodic external forcing is applied to emulate tidal behaviour or 'sloshing'. The Froude number for each test case is fixed at $\mathrm{Fr} = 0.75$ to steer away from the possibility of shocks occurring in the solution. The latter behaviour is not the focus of this chapter.

In the analysis of the results, all $\boldsymbol{\xi}_i$ are multiplied by the square root of the corresponding eigenvalues and convergence of the $\boldsymbol{\xi}_i$ is quantified by comparing the infinity norm of the eigenfunctions on various grids.

The reference solution is defined as the numerical solution on a grid of 512 computational cells. The corresponding coarse simulations range from 256 down to 8 grid cells. The ratio between the coarse and fine time step sizes is fixed at 4. For all simulated coarse grids one could choose a different Δt on each grid to ensure stability. Since the method used here is general and applies for any value of Δt, for convenience and without loss of generality we have adopted the same time step size for all grids.

2.4.1 Steady flow over a periodic ridge

The steady flow over a periodic ridge can be computed reliably at a range of spatial resolutions, using both simulation methods. Here we analyze the profiles of the eigenvectors $\boldsymbol{\xi}_i$ and the energy associated to them for different grid coarsenings.

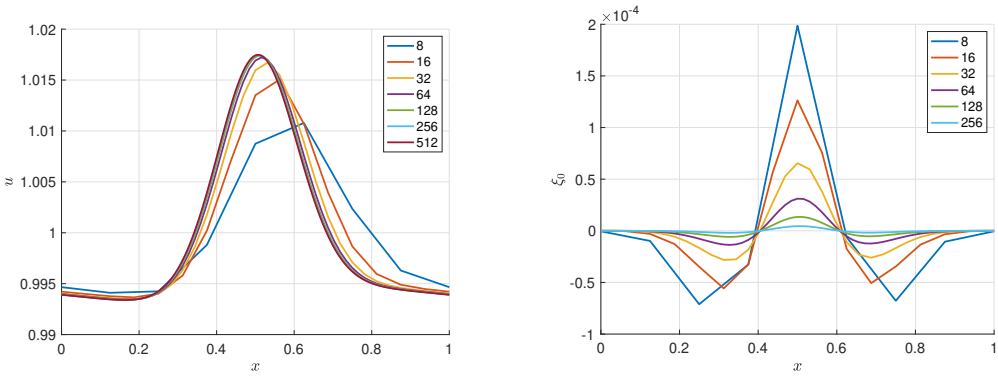
By introducing forcing and a counterbalance damping, which emulates tilting of the domain, the model reaches a nontrivial stationary state. In a practical setting, the damping can be thought of as a necessary term to control the discharge rate of the fluid. From this point the measurements of the coarse-grid correction are gathered. For a value of the forcing and damping rates (a and r in equation (2.1)) equal to 0.5, an approximately steady state is reached at $t = 30$. Measurements are then collected for one time unit, a time interval deemed sufficiently long to generate enough data for the EOF algorithm.

Since the flow is at steady state, the time mean $\boldsymbol{\xi}_0$ in equation (2.12) captures virtually all of the coarse-grid difference that should be added, at each time-step, to maintain the fine solution on the coarse grid. Ideally only the coarse-grid correction after one time-step is needed to recover a steady solution. However, given the fact that the fine grid solution is still varying slowly, we accumulate measurements over one time unit.

The velocities at $t = 30$ for various grid sizes are shown in figure 2.1a for the finite difference discretization and in figure 2.2a for the finite element discretization. The corresponding profiles of $\boldsymbol{\xi}_0$ are reported in figures 2.1b and 2.2b. For the FD method the dominant error is due to artificial dissipation, associated with the first-order upwind scheme. This error is expected to increase for grid coarsening, as is clearly visible in figure 2.1b. Additionally, $\boldsymbol{\xi}_0$ does not undergo a qualitative change as the grid is coarsened, only increasing in magnitude is observed to attain its largest value where the second derivative of the true velocity is at its highest, indicating that $\boldsymbol{\xi}_0$ captures the effect of energy dissipation.

The measured errors for the FE method are illustrated in figure 2.2b and show a neat difference compared to the FD results. The FE error is several orders of magnitude smaller than the FD error and is growing in the direction of the flow, which suggests a dispersive-type error.

The orders of convergence of the amplitude of $\boldsymbol{\xi}_0$ are found to reflect the expected order of accuracy of the methods. This is shown in figure 2.3. Using the FD method, $\boldsymbol{\xi}_0$ shows first-order convergence, second-order convergence is observed for the FE method.

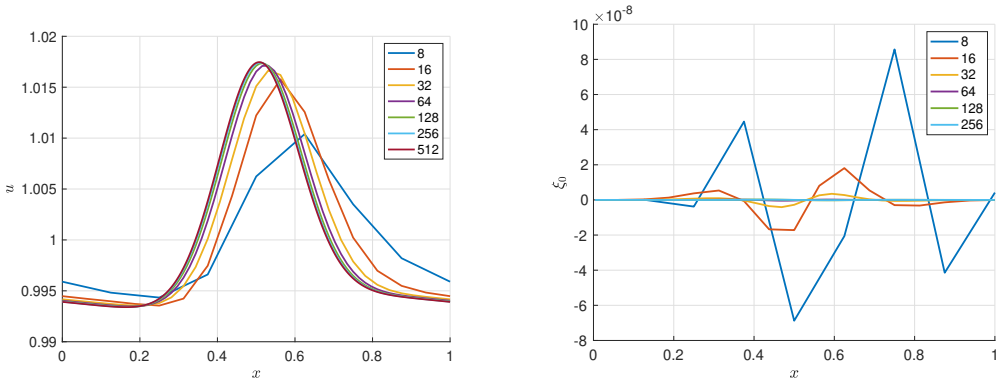


(A) (B)

FIGURE 2.1: Left: Steady state of the velocity u for various spatial resolutions using the FD discretization. Right: Time-independent profile $\boldsymbol{\xi}_0$ as obtained from the EOF algorithm for various spatial resolutions using the FD discretization.

2.4.2 Periodic sloshing over a periodic ridge

In the second test case time-periodic forcing is applied. This ensures that the velocity does not reach a steady state making this case suitable for analysing the eigenfunctions $\boldsymbol{\xi}_i$ and their corresponding temporal coefficients $\alpha_i(t)$. The



(A) (B)

FIGURE 2.2: Left: Steady state of the velocity u for various spatial resolutions, using the FE discretization. Right: Time-independent profile $\boldsymbol{\xi}_0$ as obtained from the EOF algorithm for various spatial resolutions using the FE discretization.

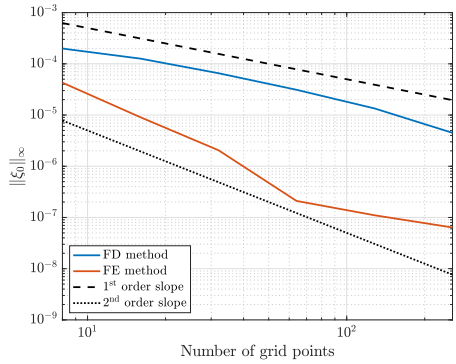


FIGURE 2.3: Infinity norm of $\boldsymbol{\xi}_0$ for various spatial resolutions, for the FD method and the FE method. The dashed and dotted lines depict the slopes for first-order and second-order convergence, respectively.

forcing consists is defined as follows:

$$a(t) = C \sum_{j=1}^{l} n_j \cos\left(\frac{2\pi t}{n_j}\right). \tag{2.15}$$

Here n_j denotes the j^{th} mode, with corresponding period n_j and l denotes the number of used modes. The parameter C can be chosen freely and affects the forcing amplitude. We have chosen a value of $C = 1/15$ along with the

low-frequency forcing term using $n = 10$ and high-frequency forcing terms using $n = 2$ and $n = 1$, respectively. The low-frequency component affects the solution on a long time scale and is the dominant forcing term. The high-frequency components are small disturbances affecting the solution on shorter time scales. The dominance of low-frequency components is incorporated by relating the amplitude of the forcing with the frequency of the forcing.

A spin-up time and a measuring time of five low-frequency forcing periods are adopted. It has been verified by comparing different measuring spin-up times and interval lengths that the spin-up time and measurement interval are sufficiently long to ensure a reliable measurement acquisition. The data that are obtained from these measurements can be regarded as a training data set.

The eigenvalues corresponding to $\boldsymbol{\xi}_i$ represent the fraction of energy related to the mode i. Of particular interest is the relation between the cumulative energy and the fraction of the available EOFs on various grids. The cumulative energy of n EOFs is given by

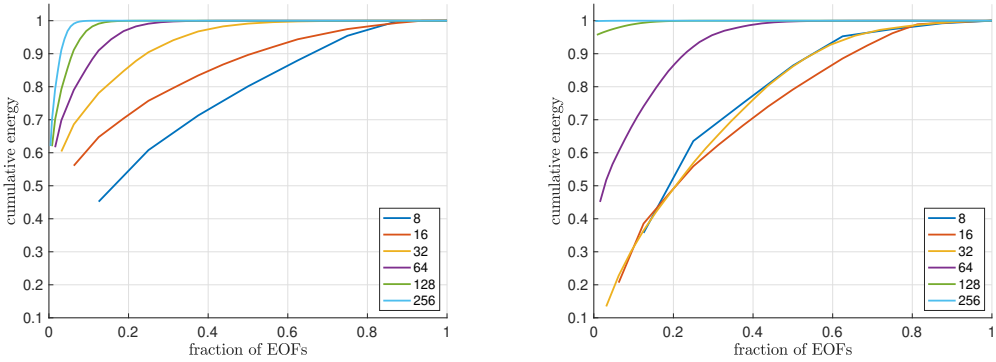
$$Q(n) = \frac{\sum_{i=1}^{n} \lambda_i}{\sum_{k=1}^{N} \lambda_k}, \tag{2.16}$$

where N denotes the total number of EOFs available from the simulation and λ_i the eigenvalues. Figures 2.4a and 2.4b Q as a function of the available EOFs for the FD method and the FE method, respectively. The difference between the truth and the coarse grid simulations decreases as the coarse grids are refined. Correspondingly, the correction toward the truth simulation can be reduced and less of the available data is required to capture the solution's variability.

Apart from the coarsest grid, the FD method requires the same number of EOF modes to capture nearly all energy of the correction, i.e., with 16 EOFs 99 percent of the variability is captured on all grids with 32 or more grid cells. The cumulative energy for the FE method show a markedly different convergence. Almost all variability of the correction on the finest grid is contained within the first EOF, indicating that the coarse-grid solution follows the truth very closely on each of the coarse grids selected. The coarsest solutions each require the same fraction of available EOFs to fully represent the reference solution on the respective grids. A similar result was observed for the cumulative energy of the free surface height.

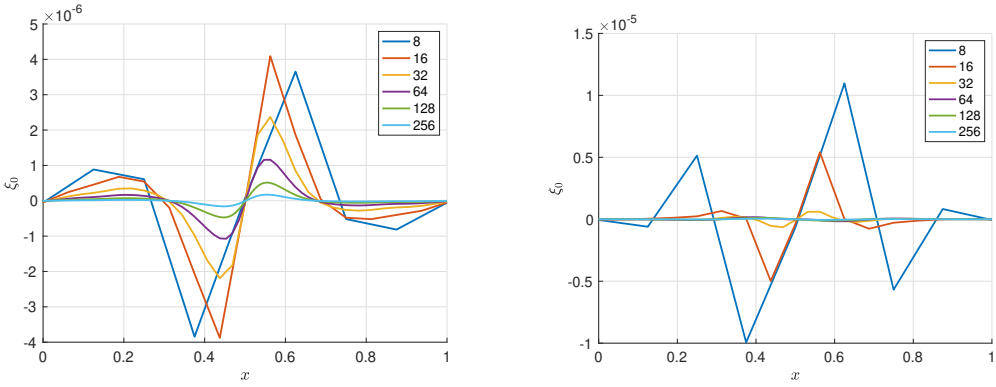
In figures 2.6 the first EOF mode for both considered methods is shown. Comparing the different methods, the modes display qualitative differences. The strong difference between the coarsest grid and the finer grids indicates that 8 grid points are too few to resolve the solution of the sloshing problem with the FD method, and hence the captured correction differs strongly from

the other computational grids. Convergence of the infinity norm for grid refinement is shown in figure 2.7. The FD method displays first-order convergence and the FE method exhibits second-order convergence. In a similar manner, the EOFs for the free surface height were found to exhibit faster convergence due to the second-order discretization of the continuity equation. First-order convergence was observed for the EOFs for the free surface height.



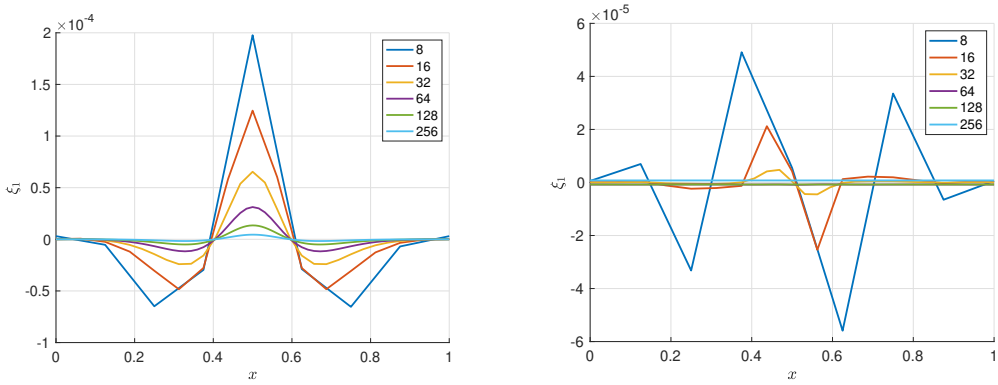
(A) (B)

FIGURE 2.4: Cumulative energy of the subgrid velocity measurements as a function of the number of EOFs for various spatial resolutions, obtained using the FD method (a) and the FE method (b).



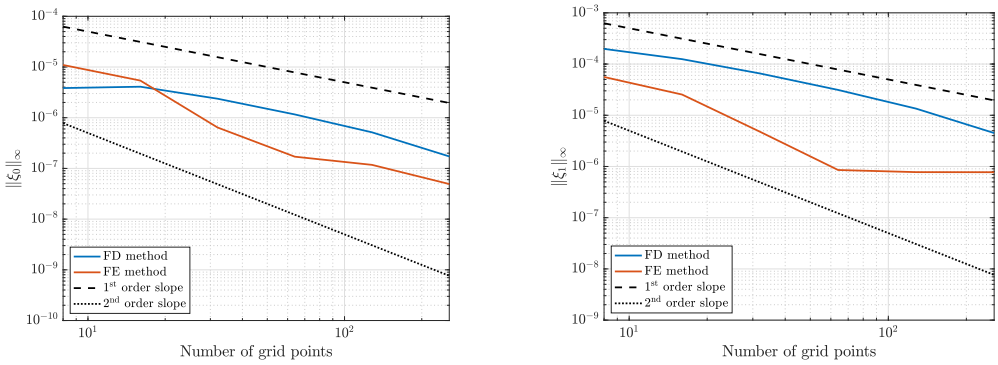
(A) (B)

FIGURE 2.5: Time-independent profile $\boldsymbol{\xi}_0$ for the velocity measurements at different grid resolutions using the FD method (a) and the FE method (b).



(A) (B)

FIGURE 2.6: First EOF mode $\boldsymbol{\xi}_1$ of the velocity measurements, for various spatial resolutions obtained using the FD method (a) and the FE method (b). The normalized EOF modes have been multiplied by the square root of the corresponding eigenvalues.



(A) (B)

FIGURE 2.7: Infinity norm of $\boldsymbol{\xi}_0$ (a) and $\boldsymbol{\xi}_1$ (b) for various spatial resolutions, for the FD method and the FE method. The dashed and dotted lines depict the slopes for first-order and second-order convergence respectively.

2.5 Reduced-order corrections based on EOF data

In this section, we apply the reduced-order model developed in section 2.3.3 to the coarse solutions of the previously presented test case with periodic forcing. In sections 2.5.1 and 2.5.2, we investigate how the ability of the measured terms to reconstruct the original data set. In these cases, we analyze two grid coarsenings: 32 and 8 grid cells. The former resolution allows comparing the

FD and FE method for the situation in which they show comparable accuracy, as it was verified numerically. The latter resolution represents a challenging case given the extreme coarsening. To disentangle the effect of the coarse-grid correction on u and η, we present the results first for the case in which the reduced-order model is applied to both state variables and then applied to them separately.

In sections 2.5.3 and 2.5.4 we investigate the robustness of the model for different initial conditions in a periodic regime. Here it is shown that general use of such models requires estimation of the temporal coefficients of the EOFs and how mean quantities might be improved in periodic regimes.

The L^2-norm of the pointwise velocity difference with the reference solution is adopted as the error measure, where the reference solution is injected on the coarse grid. Both the FD and the FE discretization use nested grids for the velocity and thus injection is performed trivially. As a measure for the error between the fine and coarse grid solutions we define

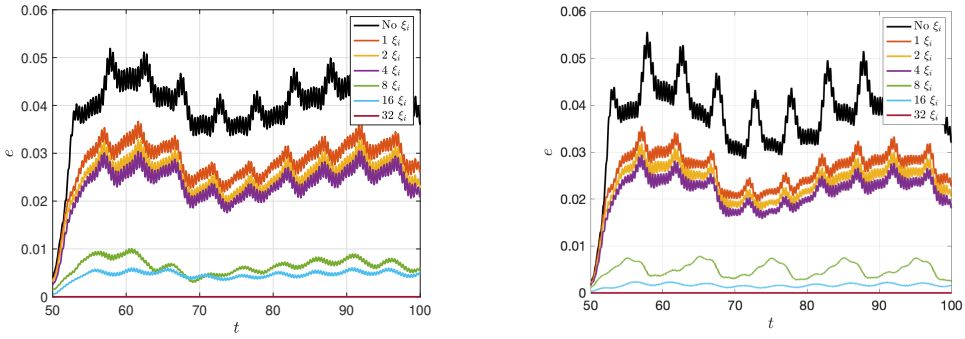
$$e(t) = \frac{1}{K+1} \sum_{k=0}^{K} \left[\sum_{i=1}^{N} \left(u_{\text{truth}}(x_i, t + k\Delta t) - u(x_i, t + k\Delta t) \right)^2 \right]^{\frac{1}{2}}, \tag{2.17}$$

where N and x_i denote the number and positions of the coarse grid points, respectively. Time averaging of the error is performed over a period of $K\Delta t$. This time interval is chosen to be one time unit so that the contribution of the high-frequency forcing component to the error remains visible.

2.5.1 Error reduction when correcting all state variables

Coarse-grid corrections are applied to both u and η. Figures 2.8a and 2.8b show the error reduction over time using an increasing number of EOFs. The mean error values over the time interval $[60, 100]$ and the percentage of reduction compared to the coarse solution without correction are given in table 2.1. Including one EOF in the correction already reduces the error by over 30 percent for both methods. Using a quarter of the available data, 8 out of 32 EOFs, reduces the error by over 80 percent for this test case. The computational cost for an increasing number of EOFs for the FD method is given in table 2.2 and is measured as the CPU time on a local computing cluster. Generating the DNS data takes is the most time consuming part of the algorithm, followed by the computation of the EOFs. No increase in computational cost is observed when including the EOFs in the coarse-grid simulations.

Figures 2.9a and 2.9b illustrate the error reduction for both methods performed on a grid with 8 computational cells. The method of correction follows from the same principle as shown for 32 cells, but very coarse grids do not



(A) (B)

FIGURE 2.8: Error (2.17) on a grid consisting of 32 cells for an increasing number of EOFs using the FD method (a) and the FE method (b). Note that using 32 EOFs recovers the reference solution and zero error is measured.

TABLE 2.1: Average values of (2.17) on a grid consisting of 32 cells over the time interval $[60, 100]$ as an increasing number of EOFs in included in the coarse-grid correction. The error reduction percentage is calculated with respect to the situation where no correction is applied.

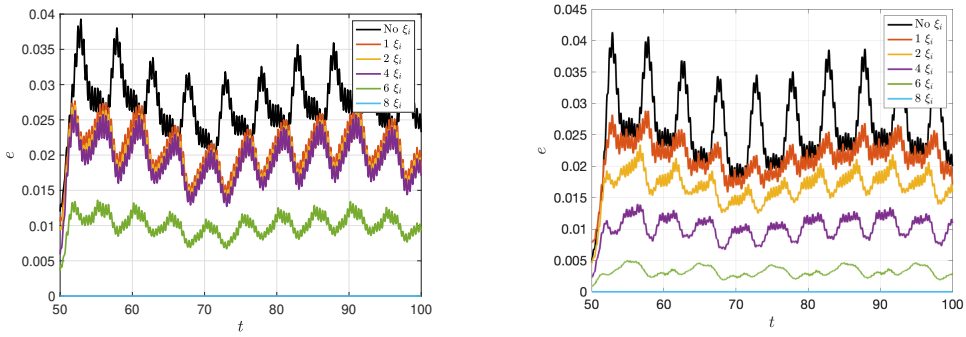
	FD		FE	
	Mean error	Reduction	Mean error	Reduction
No correction	4.156×10^{-2}		3.908×10^{-2}	
1 $\boldsymbol{\xi}_i$	2.898×10^{-2}	30.2%	2.617×10^{-2}	33.0%
2 $\boldsymbol{\xi}_i$	2.579×10^{-2}	37.9%	2.363×10^{-2}	39.5%
4 $\boldsymbol{\xi}_i$	2.407×10^{-2}	42.1%	2.162×10^{-2}	44.7%
8 $\boldsymbol{\xi}_i$	6.343×10^{-3}	84.7%	5.006×10^{-3}	87.2%
16 $\boldsymbol{\xi}_i$	4.730×10^{-3}	88.6%	1.662×10^{-3}	95.7%
32 $\boldsymbol{\xi}_i$	2.005×10^{-13}	100%	2.617×10^{-13}	100%

allow for an accurate resolution of bathymetry and hence the dynamics of the numerical solution can be vastly different than that of the DNS. The best obtainable result is then achieved by accurately representing the largest scales of the solution and doing so with low computational cost is valuable.

The mean values of the error are provided in table 2.3. It can be observed from this table that significant error reduction is possible on this grid even when not using all EOFs. For example, using 6 out of 8 available EOFs yields an error reduction of over 60 percent and 80 percent for the FD method and the FE method, respectively.

TABLE 2.2: Computational cost in seconds for performing the DNS, applying the EOF algorithm and performing coarse-grid simulations with an increasing number of EOFs.

	Cost
DNS	7.579
EOF algorithm	2.652
Coarse grid, no correction	0.1599
Coarse grid, 1 $\boldsymbol{\xi}_i$	0.1522
Coarse grid, 2 $\boldsymbol{\xi}_i$	0.1535
Coarse grid, 4 $\boldsymbol{\xi}_i$	0.1521
Coarse grid, 8 $\boldsymbol{\xi}_i$	0.1547
Coarse grid, 16 $\boldsymbol{\xi}_i$	0.1519
Coarse grid, 32 $\boldsymbol{\xi}_i$	0.1519



(A) (B)

FIGURE 2.9: Error (2.17) on a grid consisting of 8 cells for an increasing number of EOFs using the FD method (a) and the FE method (b). Note that using 8 EOFs recovers the reference solution and zero error is measured.

2.5.2 Error reduction when correcting one of the two the state variables

Figures 2.10a and 2.10b show the error reduction when only one of the variables is corrected, using the FD method and the FE method, respectively. The coarse grid consists of 32 computational cells for this comparison and coarse-grid corrections are implemented using the full set of computed EOFs for the considered state variable.

The reduced error in figure 2.10a shows a considerable improvement if the u correction is analyzed. This is in agreement with the fact that the first-order

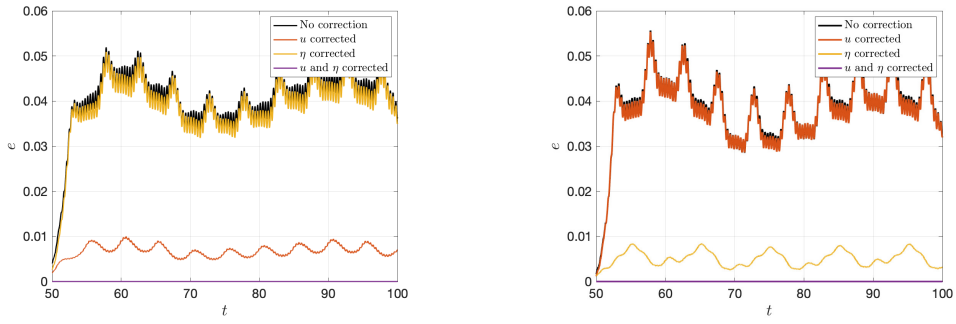
TABLE 2.3: Average values of (2.17) on a grid consisting of 8 cells over the time interval $[60, 100]$ as an increasing number of EOFs in included in the coarse-grid correction. The error reduction percentage is calculated with respect to the situation where no correction is applied.

	FD		FE	
	Mean error	Reduction	Mean error	Reduction
No correction	2.675×10^{-2}		2.611×10^{-2}	
1 $\boldsymbol{\xi}_i$	2.053×10^{-2}	23.3%	2.126×10^{-2}	18.6%
2 $\boldsymbol{\xi}_i$	1.953×10^{-2}	27.0%	1.669×10^{-2}	36.1%
4 $\boldsymbol{\xi}_i$	1.901×10^{-2}	28.9%	1.015×10^{-2}	61.1%
6 $\boldsymbol{\xi}_i$	9.922×10^{-3}	62.9%	3.202×10^{-3}	87.7%
8 $\boldsymbol{\xi}_i$	1.025×10^{-13}	100%	2.104×10^{-13}	100%

upwind scheme used for convection introduces the dominant source of error. Applying a correction to the free surface height does not yield significant improvement, since the error in the momentum equation dominates. Conversely, for the FE method the correction of the momentum equation does not lead to any significant improvement, as the FE method employed here shows high accuracy by itself. As mentioned in section 2.2, the FE method adopts first and zeroth order polynomials in the discretization of the momentum equation and continuity equation, respectively. Thus, it is reasonable to expect that correcting the free surface height strongly reduces the overall error since this is the dominant source of error. This is observed in figure 2.10b.

TABLE 2.4: Average values of (2.17) on a grid consisting of 32 cells over the time interval $[60, 100]$ as either the velocity or the free surface height are fully corrected. The error reduction percentage is calculated with respect to the situation where no correction is applied.

	FD		FE	
	Mean error	Reduction	Mean error	Reduction
No correction	4.156×10^{-2}		3.908×10^{-2}	
u corrected	6.888×10^{-3}	83.4%	3.872×10^{-2}	0.870%
η corrected	4.014×10^{-2}	3.27%	5.089×10^{-3}	87.0%
u and η corrected	2.005×10^{-13}	100%	2.617×10^{-13}	100%



(A) (B)

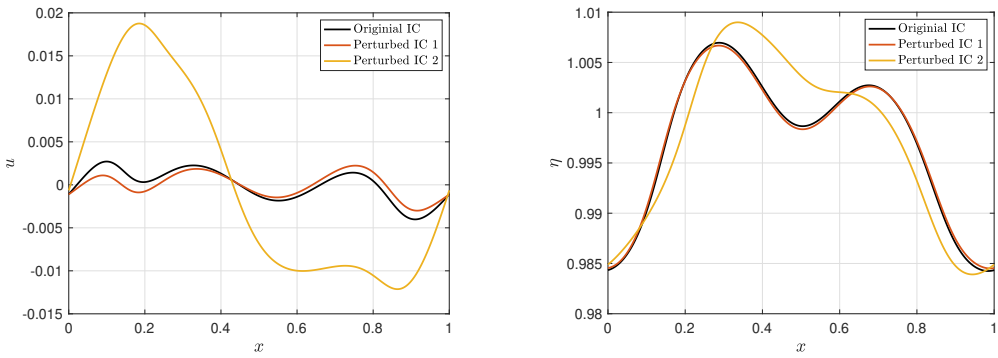
FIGURE 2.10: Error (2.17) on a grid consisting of 32 cells as either the velocity or the free surface height is fully corrected, using the full set of EOFs for the FD method (a) and the FE method (b). Note that correcting both the velocity and the free surface height produces zero error.

2.5.3 Sensitivity to initial conditions

In this subsection, we investigate the accuracy of predictions under perturbations of the initial conditions. The aim is to probe the robustness of the model in actual predictions, where the initial condition is in general different from that used in the dataset the model was trained on. By changing the initial conditions, the evolution of the flow is changed and thus the measured time series and EOFs constitute a correction term that no longer coincides with the exact subgrid data. The results are presented for the finite difference method using a reference grid of 512 computational cells and a coarse grid of 32 computational cells.

The perturbed initial conditions and the initial conditions used to generate the reduced-order corrections are found in figure 2.11. The perturbed ICs are obtained by sampling the DNS result at times $t = 100$ and $t = 10$ and are referred to as perturbed IC 1 and perturbed IC 2, respectively. It can be observed that the former slightly deviates from the original initial condition, while the latter deviates significantly.

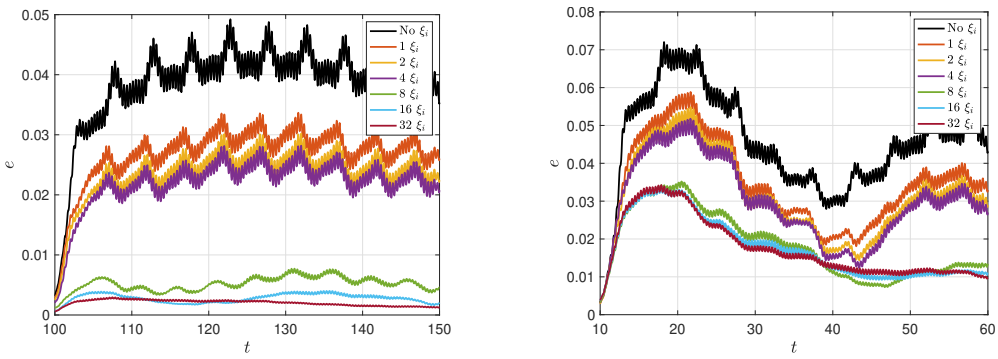
The measured errors for these initial conditions are given in figure 2.12. Application of the correction term leads to a decrease of the error, which becomes especially apparent when applying the correction to perturbed IC 1 while less so for perturbed IC 2. This behavior is to be expected, since the correction term was designed for one specific initial condition. However, the results presented in figure 2.12 indicate that the measured temporal coefficients tolerate some level of approximation without a significant loss in error reduction. We



(A) (B)

FIGURE 2.11: Initial conditions used to establish the sensitivity of the reduced-order correction term. Both the initial velocities (a) and the initial free surface height profiles (b) are obtained by sampling the numerical solution at specified times. The initial conditions for the original data set is given in black, the red line and yellow lines denote the perturbed initial conditions.

note that a further reduction of the error may be achieved by constructing an estimation of the temporal coefficients (2.14) for the $\boldsymbol{\xi}_i$ and would further extend the applicability of the reduced-order correction terms. Examples of such methods have been suggested in literature, such as regarding the temporal coefficients as a stochastic process [40] or state-dependent subgrid forcing [6].



(A) (B)

FIGURE 2.12: Error (2.17) on a grid consisting of 32 cells for an increasing number of EOFs using the FD method for perturbed IC 1 (a) and perturbed IC 2 (b).

2.5.4 Approximation of long-time averages

Often, in practical situations, one does not wish to reproduce the instantaneous fields but rather long-time averages or statistics of the underlying fields. In this subsection, we study this problem by obtaining the EOFs and corresponding time series from a particular data set and subsequently applying the obtained forcing to the same flow with a different initial condition.

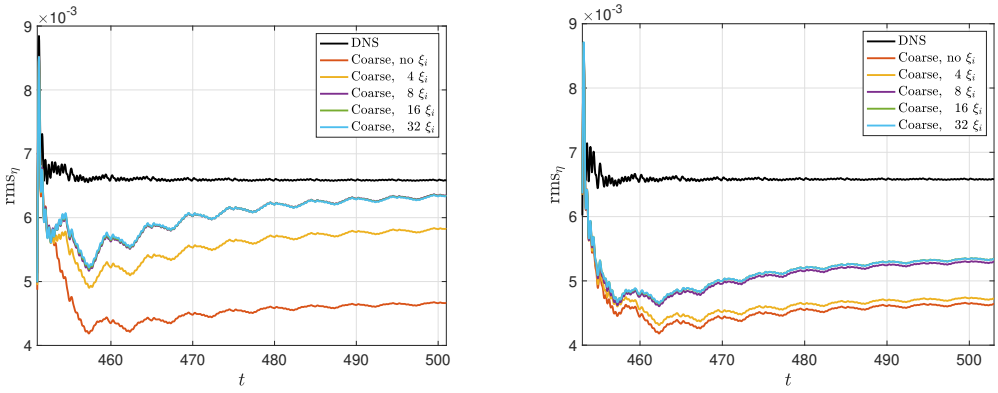
The EOFs and corresponding time series are computed for the second test case after the flow has reached a periodic regime due to the periodic forcing. The EOFs are measured for 10 time units, one period of the forcing, from the periodic state. In this regime, it has been verified that the EOFs are the same for each periodic measuring interval, as expected. Therefore, the change in the initial condition only affects the temporal coefficients coefficients.

To study the ability of the measured corrections to approximate mean quantities of the flow, we compare the root mean square (rms) variation of the free surface height,

$$\mathrm{rms}_\eta(t_k) = \left(\frac{1}{N} \sum_{j=1}^{N} \left[\eta(x_j, t_k) - \frac{1}{N} \sum_{i=1}^{N} \eta(x_i, t_k) \right]^2 \right)^{1/2}. \tag{2.18}$$

We consider two initial conditions in the periodic regime. Compared to the measuring time, the first initial condition is phase-shifted by one time unit and the second initial condition is phase-shifted by three time units. Applying the measured corrections to these situations yields the rms of η shown in figure 2.13. It can be observed that including the correction terms leads to an improvement in the prediction for both cases, but the level of improvement depends on the chosen initial condition. This behavior is to be expected, since the correction terms are tailored for one specific situation.

Analogous to what was shown in section 2.5.3, a further reduction of the error is expected to take place when a model able to also account for the current state of the solution is applied to the temporal coefficients.



(A) (B)

FIGURE 2.13: Moving time-mean of the rms of η for a different number of EOFs and for two initial conditions in the stationary regime. The initial conditions are phase-shifted by one time unit (a) and by three time units (b).

2.6 Concluding remarks

In this chapter, we have compared subgrid measurements of the difference between a highly resolved truth and a corresponding coarse representation obtained with a finite difference and a finite element method for the one-dimensional shallow water equations. This difference was used to obtain a reduced-order correction on coarse grids. Special attention was given to the definition of these measurements, such that subgrid features caused by numerical error could be account for. This error draws contributions from both an incomplete representation of the spatial derivatives as well as from inaccuracies with which details in the bathymetry are included. The measurements of coarse-grid correction were decomposed into empirical orthogonal functions (EOFs) subsequently used to define a high-fidelity reduced-order model.

The EOFs were found to reflect the associated error of the particular discretization. While the reduced order correction can be constructed such that any numerical errors can in principle be fully eliminated for any discretization, the actual characteristics of the corrections are highly specific to the adopted discretization approach. Convergence of the subgrid corrections towards zero was observed for both discretization methods and for each eigenfunction with grid refinement. Going from coarser to finer grids, less of the available data is required to capture a certain fraction of the variability of the subgrid measurements. This procedure was applied to a steady case and a periodically forced case, for a given bathymetry.

The developed reduced-order correction has been defined such that the DNS representation on coarser grids could be reconstructed exactly. Here, this implies that the fine-grid solution on the coarse grid locations is captured fully. Even using only a fraction of the available EOFs for each state variable yields a significant improvement over the coarse grid solution. This procedure also identifies the weakest point in each discretization, by showing where one can improve the most upon using more EOFs.

The reduced-order corrections were applied to several situations that differ from the original data set that the model was trained on. It was observed that predictions of mean quantities were improved when including the correction term. The level of improvement depends on the number of EOFs used in the model and on the distance from the initial condition of the data set. In addition, sensitivity to initial conditions was further explored and it was found that the corrections tolerate some level of approximation. This result makes it clear that accurately predicting the time series of each of the EOFs in the correction term will lead to further error reduction.

The results presented in this chapter may be used in future work regarding coarse-grid predictions of geophysical fluid flows. Of particular interest is the application of the reduced-order correction for complex models such as the (thermal) rotating shallow water equations which are characterized by a richer dynamics than that of the sloshing case for the shallow water equations. Here we have shown that using a subgrid model constructed by a suitable subset of the EOFs of the actual subgrid term yields effective error reductions of coarse-grid predictions. This held true also in the situation for which initial conditions were not too far from those used for generating the dataset. The latter observation hints at the relevance of the modeling of the temporal coefficients once provided with an EOF basis from data. Additionally, the numerical experiments presented in this chapter can be performed using different numerical methods to gain better understanding of the behaviour of the EOFs on coarse grids for different numerical methods. This can in turn lead to better predictions of the behavior of EOFs when DNS is not available, or when different flow conditions are considered.

Chapter 3

Data-driven stochastic Lie transport modeling of the 2D Euler equations

3.1 Introduction

A major challenge in geophysical and observational sciences is the representation and quantification of uncertainty in numerical predictions. The uncertainty stems from various sources, most relevantly from incomplete inclusion of all relevant physical mechanisms in the models and uncertainty in the initial and boundary conditions [135]. Important models for geophysical fluid dynamics, such as the two-dimensional Euler equations, quasi-geostrophic equations or rotating shallow water equations are derived from the three-dimensional Navier-Stokes equations. A sequence of simplifying assumptions is applied in order to reduce the complexity of the model to a more manageable level, while retaining main flow physics [173]. Stochastic extensions to these models have also been derived [86]. These approximate models are nevertheless rich in dynamics and contain a wide range of spatial and temporal scales. Numerically resolving the entire spectrum of scales is often not computationally feasible, meaning that either the complexity of the model should be reduced even further such that the resulting model is simple enough to be solvable numerically, or the complex model is represented on a coarse computational grid and unresolved scales are replaced by a sub-grid model. The latter option may be combined with stochastic forcing, which provides an effective way to represent

The material in this chapter was published in the *Journal of Advances in Modeling Earth Systems*, see [56].

unresolved scales in numerical simulations [25], [61], [116]. The use of stochasticity as a means to represent the unresolved scales serves to restore some of the missing small-scale dynamics and at the same time probes an ensemble of solutions and hence also investigates uncertainty. In this chapter, we embrace these ideas and develop and assess stochastic data-driven models for the two-dimensional Euler equations on the unit square.

Data-driven stochastic models in dynamical systems have been studied actively in recent years. For weather and climate models, stochasticity was used as a tool to represent uncertainty in initial conditions and in the model, as shown by [136]. A commonly used example to illustrate the data-driven stochastic approach is the two-scale Lorenz '96 (L96) system, introduced by [110] and originally proposed as a simplified model of the atmosphere that incorporates interactions between slow and fast scales. Data of the fast scales, interpreted as unresolved sub-grid scales, may serve to construct a data-informed stochastic model. Examples are given in [6] where sub-grid features are modeled using different types of noise including additive, multiplicative and state-dependent noise. This study established that stochastic parametrizations could accurately account for modelling error, with a considerably improved forecasting skill when temporal correlation was included in the noise. The correlated noise was modeled as a one-step autoregressive model with parameters fitted from data. Alternative ideas such as stochastic parametrization based on Markov chains inferred from data are presented by [47], where unresolved processes are represented as stochastic processes dependent on the state of the resolved variables and an assumed probability density. Using this approach, good agreement was found for the probability density functions and autocorrelation functions of resolved state variables.

Data-driven machine learning has also been adopted to represent small-scale dynamics for a large range of parameters [64]. It was found that several configurations of machine learning accurately reconstructed spatio-temporal correlations of the original system. These methods are not limited to simplified models such as the L96 system, but have also been successfully applied to more complete geophysical models. Examples include oceanic flows as considered in [19] and atmospheric processes as investigated in [133]. Both studies obtain a parametrization using machine learning based on off-line computed high-resolution model output. This machine learning approach could accurately predict the relation between resolved and unresolved turbulent processes, although a reliable generalization is principally not guaranteed. Here, we follow another data-driven 'offline/online' route and express the differences between a fully resolved model and a coarsened model in terms of a converging series of empirical orthogonal functions (EOFs) and introduce explicit forcing to update the coarsened model to high accuracy. This direct forcing strategy can also be

extended to structure-preserving stochastic models as will be clarified below. In a seminal work [85], stochastic partial differential equations are derived for fluid dynamics by means of a variational principle. As a result, the solution of the SPDE is compliant with the geometry of the underlying equations. This means that conservation laws are maintained under the inclusion of stochastic perturbations. This approach goes by the name of stochastic advection by Lie transport (SALT). In a similar approach, stochastic fluid models can be derived following the framework of location uncertainty (LU) [123], in which the kinetic energy is conserved. In these approaches, spatial correlations of observational data can be used to model the unresolved scales in a numerical simulation. The spatial correlations can be decomposed into EOFs [114, 79]. These are coupled to noise generated from stochastic processes in a separate modeling step. Together, these terms constitute a stochastic forcing term for the coarse PDE, which models unresolved scales. The conservation properties in the framework of SALT require a calculus in which the chain rule coincides with those of deterministic calculus. When the stochastic integration is of the Stratonovich type, it has been shown that integration with respect to semimartingales preserves the conservation properties [154]. For processes with unbounded variance one should resort to pathwise approaches. Recently, conservation properties for SALT and LU have been established for geometric rough paths [46].

The SALT approach finds meaningful applications within geophysical fluid dynamics, since these models are directly based on a variational point of view. To illustrate the SALT approach, [40] apply it to the two-dimensional Euler equations. In this study a fine-grid simulation is performed from which the Lagrangian trajectories are extracted and compared to the corresponding trajectories acquired after filtering the velocity field. The difference between these trajectories is a measure of the unresolved scales to which the EOF decomposition is applied to form an optimal basis for this term. Subsequently, a coarse SPDE is constructed according to SALT where the amplitude of the EOF basis is modeled as a decorrelated stochastic Gaussian process. It is shown that an ensemble of stochastically forced flows captures the mean values of the true solution over considerable time intervals. In a follow-up study [41], a particle filter was added to the SALT two-dimensional Euler equations and data assimilation was motivated this way. It was demonstrated that significant model reduction is possible, reducing the number of degrees of freedom by two orders of magnitude without losing reliability of the results. Similar studies on the quasi-geostrophic equations have been done [39], with a focus on data assimilation [42].

Stochastic forcing allows for the use of data-driven models outside of the dataset from which the EOFs are obtained and the parametrization of the

stochastic forcing ultimately remains a modeling choice. Global basis methods adopting e.g. EOFs, Fourier modes or spherical harmonics can be motivated by the nonlocality of turbulence. Such approaches for stochastic forcing based on DNS data have been applied to, for example, barotropic flow on the sphere [62] and to three-dimensional atmospheric flows [95]. A review of parametrizations for atmospheric flows using stochastic models based on DNS data is provided in [63]. In [145] a data-driven parametrization was compared to a self-similar parametrization, using SALT in the quasi-geostrophic equations. It was found that both parametrizations accurately predict numerical errors and possess good uncertainty skills. The work by [2] adopts EOFs and compares several dependent stochastic models and found that models that include the dynamics and time-delay effects perform well.

In this chapter, we extend the work presented by [40] of stochastic forcing for the two-dimensional Euler equations. The extension presented in this work consists of the inclusion of additional information in the data-driven approach. This information is readily available from the EOF procedure and is used to define two additional types of stochastic processes. Providing a space-time array of measurements to the algorithm yields the EOFs, which are spatial profiles, and the amplitudes of the EOFs in order to reconstruct the input measurements. The amplitudes of each EOF are a time series and provide the data that are used in this chapter to calibrate stochastic processes for each of the EOFs. In order to mimic the measurements, we generate signals that have the same probability distribution function as the measured time series or have similar temporal correlation. By retaining these statistical quantities in the modeled time series, the forcing stays true to the characteristic features of the measurements.

The following numerical experiments and findings are reported in this chapter. We perform a direct numerical simulation (DNS) of the two-dimensional Euler equations on the unit square, subject to impenetrable boundary conditions. We measure the difference between trajectories of particles advected by the fully-resolved velocity field and the corresponding filtered velocity field. The EOFs and time series that represent the amplitudes of the EOFs are obtained from this data. Stochastic ensembles are generated using three stochastic processes: Gaussian noise, noise based on the underlying pdf of the EOF time series, and noise with a temporal correlation similar to that of the EOF time series. The process of developing the time series into stochastic processes is explained in detail in a subsequent section of the chapter. The results presented in this chapter show that using the developed stochastic processes leads to a reduction of the ensemble mean error and ensemble spread, compared to using Gaussian noise. This is further explored by performing statistical tests for ensemble solutions. The latter is done for time scales on which data may

be assimilated, where the numerical SPDE results may serve as input [41].
The chapter is structured as follows. In section 3.2.1 we introduce the deterministic and stochastic governing equations and describe the numerical experiment in detail. This is followed by a description of the data acquisition procedure in section 3.2.2. The method used for generating random signals as a model for the measured data is described in section 3.2.3. The results of the numerical experiments are presented in section 3.3. In section 3.3.1 a maximal prediction horizon is established and in section 3.3.2 an adapted reference solution defined. These results aid the uncertainty quantification of ensemble predictions, presented in 3.3.3. Predictions on much shorter timescales are further assessed in section 3.3.4, comparing additional ensemble statistics. In section 3.3.5 we assess the forecast quality for different lengthscales in the flow by analyzing the results in spectral space. We conclude the chapter in section 3.4 and specify future challenges.

3.2 SPDE formulation and stochastic models

This section presents the formulation of the stochastic Euler equations using the SALT approach (Subsection 3.2.1), the data acquisition procedure (Subsection 3.2.2) and the derivation of the stochastic models (Subsection 3.2.3).

3.2.1 Governing equations and flow conditions

The two-dimensional Euler equations are central to this work. These equations are determined fully by the evolution of the vorticity dynamics [173]. The behaviour of the vorticity ω in terms of the velocity $\mathbf{u}$ and streamfunction ψ is given by

$$\partial_t \omega + (\mathbf{u} \cdot \nabla)\omega = Q - r\omega, \tag{3.1}$$

$$\mathbf{u} = \nabla^\perp \psi, \tag{3.2}$$

$$\Delta \psi = \omega, \tag{3.3}$$

which are solved on the unit square, denoted by $\mathcal{D}$. The perpendicular gradient $\nabla^\perp$ is defined as $(-\partial_y, \partial_x)$. A forcing and a damping term are added to the equations in order to drive the flow to a nontrivial statistically steady state. In particular, $Q(x, y) = 0.1 \sin(8\pi x)$ and $r = 0.01$, which enforce eight spatial gyres that are constant in time. An impenetrable boundary condition is applied via

$$\psi|_{\partial\mathcal{D}} = 0 \tag{3.4}$$

along the boundary $\partial\mathcal{D}$ of $\mathcal{D}$. For this system a characteristic time scale is the large eddy turnover time, here estimated to be 2.5 time units [40].
The stochastic equations associated with the Euler equations follow from the principle of stochastic advection by Lie transport (SALT) for ideal fluid dynamics [85]. In this approach, SPDEs are derived from a variational principle. In fact, a stochastically constrained functional is minimised to obtain an SPDE which retains the geometric properties equivalent to the corresponding PDE. The result is that quantities that are advected along an infinitesimal vector field $\mathbf{u}\mathrm{d}t$ in the deterministic setting are advected along an infinitesimal vector field $\bar{\mathbf{u}}\mathrm{d}t + \sum_i \boldsymbol{\xi}_i \circ \mathrm{d}B_t^i$ in the stochastic setting. In this chapter, $\bar{\cdot}$ denotes a filtered field representative of scales that can be resolved accurately on a coarse numerical grid. As a rough rule of thumb, the resolved scales would comprise of structures for which $\Delta \gtrsim kh$ where h denotes the uniform grid spacing and k is a factor that quantifies the desired accuracy requirements. Typically, one may think of $k \gtrsim 4$ for second order accurate methods [69]. The velocity fields $\boldsymbol{\xi}_i$ are defined as the eigenvectors of the velocity-velocity correlation tensor [85], B_t^i is a Wiener process. The symbol $\circ$ implies that the stochastic integral should be understood in the Stratonovich sense. This means that the integral is approximated by Riemann sum defined on the midpoints of the subintervals. For a good introduction to this material [96] and [84] can be consulted.
Since the velocity field $\mathbf{u}$ is divergence-free, each velocity field $\boldsymbol{\xi}_i$ is divergence-free [40] and can be expressed by a potential function ζ_i via $\boldsymbol{\xi}_i = \nabla^\perp \zeta_i$. The advection velocity can then be written in terms of the potential as

$$\bar{\mathbf{u}}(t)\mathrm{d}t + \sum_i \boldsymbol{\xi}_i \circ \mathrm{d}B_t^i = \nabla^\perp \bar{\psi}(t)\mathrm{d}t + \sum_i \nabla^\perp \zeta_i \circ \mathrm{d}B_t^i. \tag{3.5}$$

Numerically, the velocity fields are projected to divergence-free fields to guarantee non-divergence. In this equation the filtered variables are used since the aim of the stochastic model is to represent the components of the fine-grid solution that are not resolvable on the coarse grid. The resulting SPDE then reads [40]

$$\mathrm{d}\bar{\omega} + \nabla^\perp \left(\bar{\psi}\mathrm{d}t + \sum_i \zeta_i \circ \mathrm{d}B_t^i \right) \cdot \nabla\bar{\omega} = (Q - r\bar{\omega})\mathrm{d}t, \tag{3.6}$$

$$\Delta\bar{\psi} = \bar{\omega}. \tag{3.7}$$

3.2.2 Data acquisition

The numerical method for the solution of (3.6)-(3.7) and the flow parameters are the same as those used in earlier studies [40, 41]. A full description of the numerical implementation can be found in the former references. Here, for

completeness, we illustrate the key aspects. A finite element method is employed to solve the system of equations (3.6) and (3.7). The Poisson equation for the streamfunction is discretized using a continuous Galerkin scheme. The vorticity equation (3.1), including the stochastic terms, is discretized using a discontinuous Galerkin scheme. The space of discontinuous test functions guarantees numerical conservation of energy in the absence of source terms [14]. Numerical time integration is performed by applying a third-order strong stability preserving Runge-Kutta (SSPRK3) method [149]. Writing the stochastic advection equation (3.6) in the general Stratonovich SPDE form

$$\mathrm{d}\bar{\omega} = L(\bar{\omega})\mathrm{d}t + \sum_{i=1}^{m} G^i(\bar{\omega}) \circ \mathrm{d}B_t^i, \tag{3.8}$$

where

$$\begin{aligned} L(\bar{\omega}) &= -\nabla^\perp \bar{\psi} \cdot \nabla \bar{\omega} + (Q - r\bar{\omega}), \\ G^i(\bar{\omega}) &= -\nabla^\perp \zeta_i \cdot \nabla \bar{\omega}, \end{aligned} \tag{3.9}$$

the SPDE (3.8) is integrated in time via

$$\begin{aligned} \bar{\omega}_{(1)} &= \bar{\omega}_n + \Delta t L(\bar{\omega}_n) + \sum_{i=1}^{m} G^i(\bar{\omega}_n)\Delta B_n^i, \\ \bar{\omega}_{(2)} &= \frac{3}{4}\bar{\omega}_n + \frac{1}{4}\left[\bar{\omega}_{(1)} + \Delta t L\left(\bar{\omega}_{(1)}\right) + \sum_{i=1}^{m} G^i\left(\bar{\omega}_{(1)}\right)\Delta B_n^i\right], \\ \bar{\omega}_{n+1} &= \frac{1}{3}\bar{\omega}_n + \frac{2}{3}\left[\bar{\omega}_{(2)} + \Delta t L\left(\bar{\omega}_{(2)}\right) + \sum_{i=1}^{m} G^i\left(\bar{\omega}_{(2)}\right)\Delta B_n^i\right]. \end{aligned} \tag{3.10}$$

The subscript n denotes the n^{th} numerical time step. The stages of the Runge-Kutta algorithm are denoted by the subscripts (1) and (2). The time step size is given by Δt and is chosen such that the CFL number does not exceed 1/3. Here ΔB_n^i denote random samples drawn from an assumed probability distribution with variance Δt. For deterministic systems, the functions G^i equal zero.
The term $\nabla^\perp(\sum_i \zeta_i \circ \mathrm{d}B_t^i)$ in (3.6) is unknown in the coarsened description and needs to be modelled. The latter is approximated as follows:

$$\mathbf{f}(x,t)\sqrt{\Delta t} = (\mathbf{u} - \bar{\mathbf{u}})\Delta t \approx \sum_i \boldsymbol{\xi}_i(x)\Delta B_n^i. \tag{3.11}$$

The forcing $\mathbf{f}$ in (3.11) is computed as the difference of the Lagrangian trajectories originating by the velocity fields $\mathbf{u}$ and $\bar{\mathbf{u}}$ projected onto the coarse grid. As such, the forcing is a large-scale correction to $\bar{\mathbf{u}}$ which measures the part

of the velocity fluctuation resolved by the coarse grid. The right hand side incorporates these fluctuations as a stochastic forcing. The measurements are approximated by coarse-grid resolved fields. This approximation may become inaccurate in the case of severe coarsening, in which case a large number of terms must be introduced in the approximation of $\mathbf{f}$ to properly capture the effects of the small scales.
The process of measuring $\mathbf{f}(x,t)$ is as follows. A grid with 512^2 computational cells is adopted for the DNS and all subsequent stochastic results are obtained on a coarse grid of 64^2 computational cells. The filtered fields are derived from the fine-grid DNS results and are obtained by applying a Helmholtz operator to the streamfunction. Given a streamfunction ψ, the filtered streamfunction $\bar{\psi}$ is obtained by solving

$$(I - c\nabla^2)\bar{\psi} = \psi, \tag{3.12}$$

where $c = 1/64^2$ to filter out length scales smaller than the coarse grid size. The numerical resolutions and the filter width coincide with those adopted in [40]. The filtered vorticity $\bar{\omega}$ and filtered velocity $\bar{\mathbf{u}}$ are recovered from applying the relations (3.2) and (3.3) to $\bar{\psi}$. Over the entire simulation interval, this filter was found to remove approximately 12% of the kinetic energy.
The initial vorticity is prescribed, as

$$\begin{aligned}\omega_0 =& \sin(8\pi x)\sin(8\pi y) + 0.4\cos(6\pi x)\cos(6\pi y) \\ &+ 0.3\cos(10\pi x)\cos(4\pi y) + 0.02\sin(2\pi y) + 0.02\sin(2\pi x),\end{aligned} \tag{3.13}$$

from which the system will be spun-up during an interval of 100 time units so that a statistical equilibrium is reached. The time at which this is reached is denoted by $t = 0$. The initial fields and corresponding filtered fields at the end of the spin-up interval are found in Fig. 3.1.

The method to estimate the eddy velocity is the same as presented by [40] and is based on measuring the trajectories of fluid parcels. A decomposition of the true trajectories into a drift and a stochastic perturbation is assumed, which is equivalent to the SALT equations [43]. Here, the drift is computed as the filtered velocity field. Thus, a space-time sequence of measurements for determining $\mathbf{f}$ from (3.11) is obtained by computing the difference of Lagrangian trajectories of particles advected by the velocity field $\mathbf{u}$ and those advected by the filtered velocity field $\bar{\mathbf{u}}$. The difference is measured over a single coarse-grid time step. The particles are released on the coarse grid points and thus a difference in traveled distance can be related to each grid point. A correction field is subsequently obtained by dividing the difference in trajectories by the square root of the coarse-grid time step, in a manner analogous to particle image velocimetry measurement techniques in experimental fluid flow analysis [1]. The measurements are done at each coarse-grid time step, from $t = 0$ (the point at

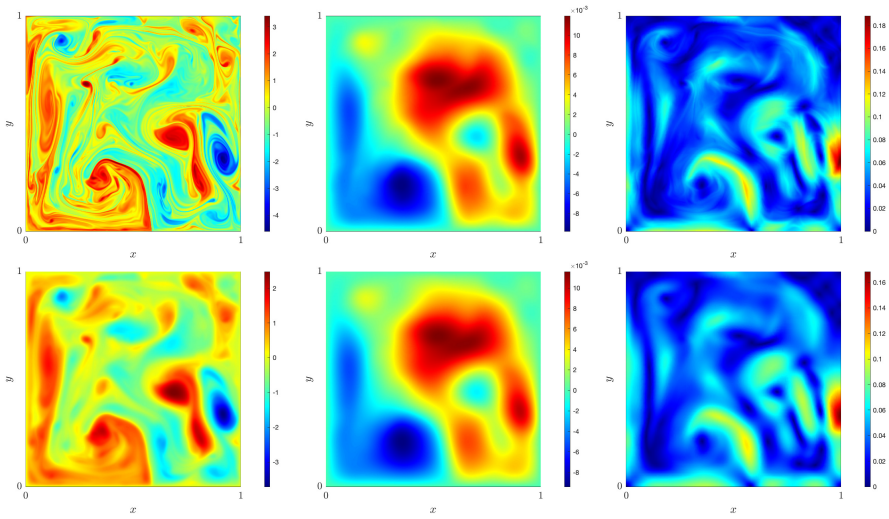


FIGURE 3.1: Fine-grid fields of the vorticity (left), streamfunction (middle) and velocity magnitude (right) after the spin-up interval. The top row shows the unfiltered fields, the bottom row shows the corresponding fields after applying the filter (3.12).

which the simulation is spun-up) until $t = 365$. By doing so, an array $\mathbf{f}(x,t)$ of fields is constructed. This space-time array of measurements is decomposed into empirical orthogonal functions (EOFs or EOF modes) [114, 79]. Here, a total of 4096 EOFs are available (64^2 degrees of freedom), of which the first 225 are used. These EOFs account for 90% of the energy of the measurements. The fact that only a small portion of EOFs is needed for an accurate reconstruction of the forcing is attributed to the Helmholtz filter being a graded filter. The latter, by construction, filters out not only small-scale dynamics but also part of the large-scale field, which can be captured by the first 5.5% of EOF modes. A potential concern in this general framework is that errors might become too large in case of further coarsening.
Application of the EOF algorithm to a flow that has a definite statistically steady state yields

$$\mathbf{f}(x,t) = \boldsymbol{\xi}_0(x) + \sum_{i=1}^{N} a_i(t)\boldsymbol{\xi}_i(x), \tag{3.14}$$

where $\boldsymbol{\xi}_0(x)$ is the time-mean of the measurements, $\boldsymbol{\xi}_i(x)$ are normalized spatial EOF modes, also referred to as 'topos', and $a_i(t)$ are the corresponding coefficients with reference to the measurements, also referred to as 'chronos'. These are recorded as time series and may be written instead as $\sqrt{\lambda_i}\bar{a}_i(t)$.

Here, $\sqrt{\lambda_i}$ is the standard deviation of $a_i(t)$ and carries the same dimension as the measurements. The time series $\bar{a}_i(t)$ has unit variance and is dimensionless. The EOF modes are orthonormal with respect to the inner product, thus $\left(\boldsymbol{\xi}_i, \boldsymbol{\xi}_j\right) = \delta_{ij}$, where

$$(\boldsymbol{f}, \boldsymbol{g}) = \int_\Omega \boldsymbol{f}(x)\boldsymbol{g}(x)\,\mathrm{d}x \tag{3.15}$$

with Ω the flow domain. Due to the orthonormality, the coefficients $a_i(t)$ are readily obtained by projecting the measured velocity fields onto the basis of EOFs by

$$a_i(t) = (\mathbf{f}(x,t) - \boldsymbol{\xi}_0(x), \boldsymbol{\xi}_i(x))\,. \tag{3.16}$$

In order to have a self-contained model which allows to obtain predictions, e.g., beyond the time span of the dataset, or as surrogate statistical sample of the flow, we choose to model the time traces $a_i(t)$ as independent stochastic processes. This will be described in the next section, where also the possible connection to the available data will be elaborated.

3.2.3 Generating random signals

We will now introduce the models for the time traces (3.16) and subsequently describe how random signals are generated using these models. By comparing (3.14) with (3.11) it is clear that modelling $B_t^i(t)$ amounts to modelling $a_i(t)$. The following models are employed:

1. The stochastic process B_t^i in (3.11) is modelled by Gaussian noise. For its discrete increments ΔB^i in (3.10) we use $\Delta B^i = \sqrt{\lambda_i}\sqrt{\Delta t} r_i$, where $r_i \sim \mathcal{N}(0,1)$. [84].

2. The probability density function (pdf) of $\bar{a}_i(t)$ in is estimated from the measured signals (3.16) and is subsequently used to draw uncorrelated samples. Thus, the increments ΔB^i in (3.10) are computed as $\Delta B^i = \sqrt{\lambda_i}\sqrt{\Delta t} r_i$, where r_i is randomly drawn from the estimated pdf.

3. The time series $a_i(t)$ in (3.14) is approximated by an Ornstein-Uhlenbeck (OU) process, using the correlation time obtained from the measurements (3.16). The constructed OU process is then used to compute ΔB^i in (3.10).

The probability distributions of model 2 are estimated by fitting a histogram to the values of the corresponding time series, yielding a separate distribution for each EOF. The histograms are fully determined by the smallest and largest measurements and the number of measurements. The number of bins is chosen as the smallest integer larger than $\sqrt[3]{2N_M}$, where N_M denotes the number of

measurements, i.e., the length of the time series. This choice minimizes the asymptotic mean squared error of the histogram as an estimator of the underlying pdf [168]. Moreover, the measurements are finite due to the spatial continuity of the numerical solution and the finite time step size, resulting in histograms with compact support. This guarantees bounded quadratic variation and finite moments, at the discrete level. Uncorrelated samples from these distributions are drawn using inverse transform sampling. In the latter a random number x is drawn from a uniform distribution between 0 and 1, which can intuitively be thought of as a probability of an event happening, and subsequently the largest value X is found such that $P(X \leq x)$ holds for the estimated distribution [51]. It is expected that the results obtained from model 2 will converge to those obtained from model 1 when the time step size is decreased, due to the central limit theorem.
In model 3, the noise generated using the OU process mimics the temporal correlation of the measured time series. Denoting by B_t^i the approximation of the time series $a_i(t)$, the OU process is defined as [139]

$$\mathrm{d}B_t^i = -B_t^i \frac{\mathrm{d}t}{T_i} + \left(\frac{2\mathrm{d}t}{T_i}\right)^{1/2} \sqrt{\lambda_i}\sqrt{\mathrm{d}t} r_i, \tag{3.17}$$

where $r_i \sim \mathcal{N}(0,1)$. We set T_i to be the correlation time of the measured time series. These variables are determined for each EOF separately. Here, the correlation time is defined as the smallest time at which the autocorrelation function of the time series is smaller than the computed 95% confidence bound.
A consistent choice for a fourth model is one that incorporates the measured temporal correlation, whilst retaining the estimated probability distribution of measurements. However, for this approach no tractable algorithm to generate the stochastic processes was found.
The conservation properties of SALT hold for the proposed stochastic processes, since these are semimartingales. Conservation of advected quantities then follows from the results of [154]. Convergence of model 2 for decreasing time step sizes is guaranteed because the histograms have finite moments. Convergence of model 3 is established in [96] since the processes are semimartingales.
In the next section, we assess the proposed stochastic models by comparing simulations on the SPDE models to findings from deterministic reference solutions.

3.3 Assessment of forecast ensembles

In this section, we provide results of forecast ensembles using the aforementioned methods to generate stochastic signals that serve to force the coarsened dynamics. We first identify a maximal prediction horizon for assessing the forecast ensembles. An adapted reference solution is defined based on the measurements, incorporating on the coarse numerical grid the measured effects of small-scale motions. Subsequently, we show results of forecast ensembles. Statistics are computed and compared to the filtered DNS and the adapted reference solution to quantitatively compare the different stochastic forcing methods. Finally, the results are compared in terms of EOF coefficients, to distinguish between the forecast quality for different lengthscales present in the flow.

3.3.1 Establishing a maximal prediction horizon

In order to define the maximal prediction horizon until which stochastically forced coarse numerical solutions can reasonably be compared to the DNS results, we set up the following numerical experiment. Starting from an initial condition on the fine grid, we generate a set of perturbed initial conditions of which we then follow the evolution over time. The perturbations are applied in Fourier space by shifting the phase of the Fourier coefficients, while keeping the amplitudes the same. The phase shift is applied only to modes of wave lengths smaller than the smallest scale resolved by the corresponding coarse grid, leaving the resolved modes unaltered. Specifically, a value l is chosen and all Fourier modes with wave numbers $|k| = (k_x^2 + k_y^2)^{1/2} \in [l, l+1)$ are affected by the additional phase shift. Here k_x and k_y denote the wave numbers in the $x-$ and $y-$direction, respectively, and l is chosen as $64, 128$ and 256. The phase shift is set to π to satisfy the boundary conditions.
As time evolves, the initial perturbation increasingly affects the resolved scales, up to the point where the instantaneous resolved fields will be entirely different from each other. We define this point of no longer truthfully following the unperturbed solution as the maximal prediction horizon $T_{\max}$, after which no model can be expected to consistently give accurate point-wise predictions owing to the sensitivity of the evolving solution to the initial conditions. The value of $T_{\max}$ is expected to depend on the choice of perturbed modes and choice of simulation parameters. However, in this numerical experiment it serves to provide an estimate of the maximal predication horizon.
The observed behaviour following the small-scale phase-shift perturbations is illustrated in Fig. 3.2, together with the results obtained from the unperturbed solution. The evolution of the vorticity for the various initial conditions

has been measured on four illustrative points in the domain, at $(0.25, 0.25)$, $(0.25, 0.75)$, $(0.75, 0.25)$ and $(0.75, 0.75)$, of which two points are shown in the figure. It can be seen that the evolution of the vorticity values at the measured points in the domain is initially indistinguishable. At $t = 10$ slight differences are visible and at $t = 20$ the measured values are markedly different. The latter result is especially clear at the point $(0.25, 0.25)$, in the left figure. Thus, we conclude that subsequent stochastic realizations can not be reasonably assessed after $t = 20$, which we set as the value for the maximal prediction time $T_{\max}$.

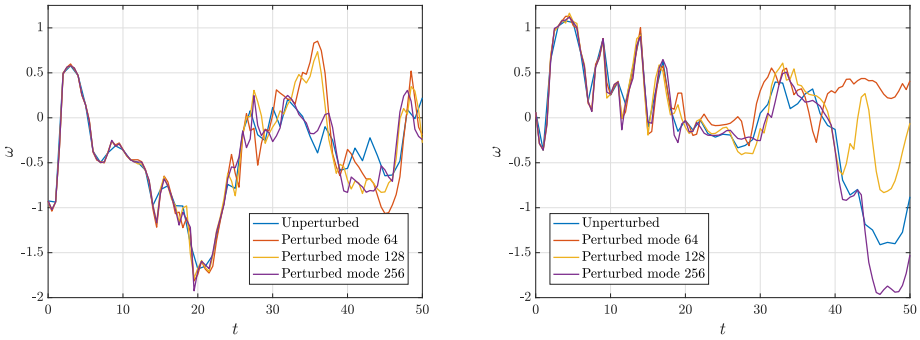


FIGURE 3.2: Development of the vorticity in two points of the domain, obtained by DNS of perturbed fine-grid initial conditions. On the left, the vorticity is shown at $(0.25, 0.25)$ and on the right at $(0.75, 0.75)$. The results for the unperturbed and three perturbed initial conditions are shown. The perturbations are defined by phase-shifting small-scale modes of the streamfunction field. In the results shown here, the Fourier modes with wave numbers $|k| = (k_x^2 + k_y^2)^{1/2} \in [l, l + 1)$, $l = 64, 128, 256$ are phase-shifted by π.

3.3.2 Defining the reference solution

In order to compare the different stochastic models one has to define a reference solution. The choice of the latter is not unique. In this work we define two reference solutions that are employed to measure performance of a given forcing model. The first one is the filtered fine-grid solution, employing the filter (3.12), and is indicative of flow scales that can be resolved on the coarse grid. Next to the filtered fine-grid solution, we define a reference solution as the numerical solution of (3.6)-(3.7) where the reconstructed signal (3.14), (3.16) is used in (3.11) instead of the stochastic forcing. This provides a prescribed deterministic forcing for the coarse numerical simulation. We call this the adapted reference solution. We note that the structure of the closure term

(3.11) does not account for discretization error and is itself not an exact closure since the noise is introduced only in the advection velocity. The inclusion of discretization error is what sets the filtered DNS and the adapted reference solution apart. Therefore, by comparing the stochastic ensembles against the adapted reference solution, one is able to distinguish between modeling error from the proposed stochastic models and the discretization error.
The adapted reference solution at $t = 0, 10, 20, 30$ is shown in the top row of Fig. 3.3. At the same points in time, a single realization of each of the stochastically forced solutions is shown. The second row shows a realization using Gaussian noise, the third row using estimated pdfs and the bottom row using OU processes. The various realizations show no qualitative difference, suggesting that a more detailed, quantitative comparison of the methods is required. This is provided in the following subsections.

3.3.3 Uncertainty quantification of ensemble predictions

The evolution of the vorticity and streamfunction is used for uncertainty quantification. First, the ensemble predictions are compared globally to the reference solution. In this subsection, the ensembles are compared quantitatively only to the adapted reference solution so that accumulation of discretization error in the coarse numerical solutions is not included in the comparison. Subsequently, similar to [40] four points in the domain are picked for pointwise uncertainty quantification. For each point one ensemble standard deviation around the ensemble mean solution is shown and compared to the reference solution at the same point. In these tests, the ensemble is initialized from a single initial condition in order to isolate the effects of the stochastic processes on the uncertainty of the numerical solution. The initial condition is obtained by injecting the DNS vorticity field onto the coarse grid. Each SPDE is simulated up to $T_{\max} = 20$, and every ensemble is composed of 200 realizations of the SPDE. Our interest here lies in comparing the errors and spreads for the different types of stochastic processes used in the forcing (3.11). Different error measures will be monitored as outlined next.
For global comparison to the reference solution, we define the pattern correlation

$$\frac{(\omega, \omega_{\text{ref}})}{\sqrt{(\omega, \omega)(\omega_{\text{ref}}, \omega_{\text{ref}})}} = \frac{\int \omega \omega_{\text{ref}} \, \mathrm{d}x}{\sqrt{\int \omega \omega \, \mathrm{d}x \times \int \omega_{\text{ref}} \omega_{\text{ref}} \, \mathrm{d}x}}, \tag{3.18}$$

which can be considered a global measure of likeness between the vorticity ω obtained from the stochastically forced numerical solution and the vorticity ω_{ref} obtained from the reference solution. The same quantity is computed for the streamfunction. The pointwise comparisons are acquired by measuring the instantaneous vorticity and streamfunction at several grid points.

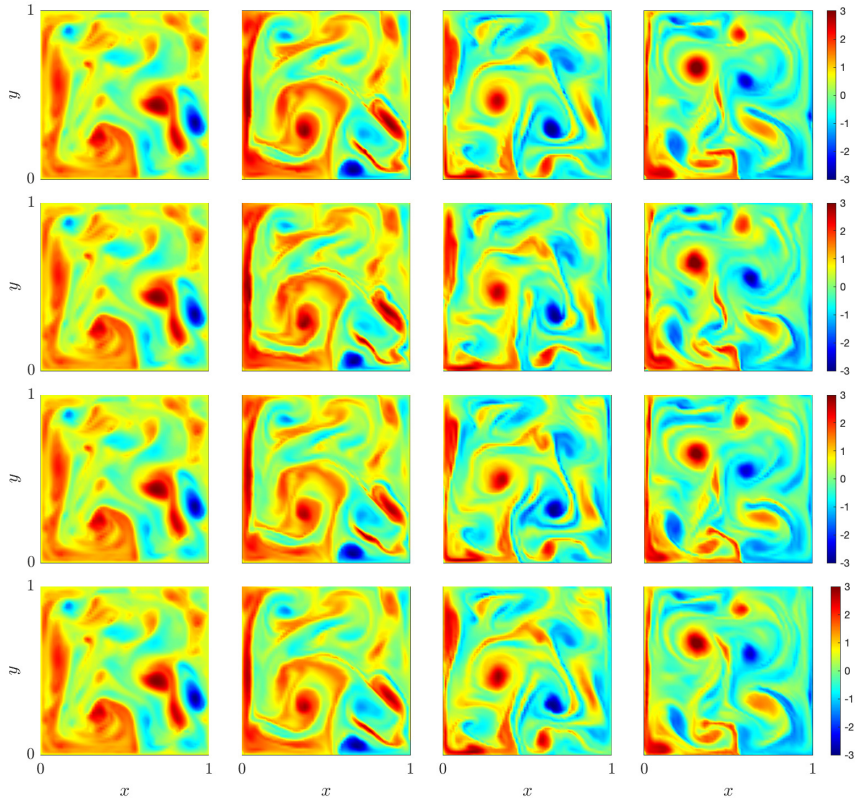


FIGURE 3.3: Coarse-grid fields of the vorticity at various points in time. The top row shows, from left to right, the adapted reference solution at $t = 0, t = 10, t = 20$ and $t = 30$. The other rows show realizations of stochastically forced numerical solutions at the aforementioned times. The second row uses Gaussian noise, the third row uses random samples from estimated distributions and the bottom row uses OU processes.

The stochastic ensembles are assessed using the ensemble mean, ensemble standard deviation and ensemble mean error. Here, we denote an ensemble of N stochastic realizations by $\{X_{i,j}\}$, where $i = 1, \ldots, N$ denotes the realization and $j = 0, \ldots, T$ denotes the time index. Then, the ensemble mean at time instance j is defined as

$$\langle X_j \rangle = \frac{1}{N} \sum_{i=1}^{N} X_{i,j}, \tag{3.19}$$

and the standard deviation, here referred to as spread, is defined as

$$\text{Spread}(X_{i,j}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_{i,j} - \langle X_j \rangle)^2}. \tag{3.20}$$

A small spread indicates a sharp ensemble forecast and a large spread suggests an increased uncertainty in the forecast. The reference solution $Y_j, j = 0, \ldots, T$ is computed at the same time instances as $\{X_{i,j}\}$. The ensemble mean error of $\{X_{i,j}\}$ is then defined as

$$\text{ME}(X_{i,j}, Y_j) = |\langle X_j \rangle - Y_j|. \tag{3.21}$$

A small ensemble mean error indicates that the ensemble closely follows the reference solution, whereas a large value implies that the ensemble and the reference solution have deviated considerably from each other.
The correlation measure (3.18) is shown in Fig. 3.4 for the vorticity and the streamfunction. Using estimated pdfs or OU processes show favourable results when compared to using Gaussian noise, for both quantities. A clear difference between the methods can be observed for the vorticity on the time scale of $T_{\max}$. At this point, using estimated pdfs or OU processes yields a smaller spread than using Gaussian noise, and the results of the latter show a smaller correlation with the adapted reference solution. A significant increase in the correlation can also be observed for the streamfunction. The results using estimated pdfs or OU processes, as opposed to using Gaussian noise, exhibit both a larger likeness with the reference solution as well as a smaller spread. Compared to the ensemble obtained using Gaussian noise, at $t = 20$ the ensemble standard deviation of the pattern correlation of the vorticity was found to be reduced by 24% and 42% when using estimated pdfs and OU processes, respectively. For the streamfunction, these values were correspondingly observed to be 67% and 84%. Moreover, the results for the estimated pdfs and the OU processes are nearly indistinguishable before $t = 5$.
The evolution of the vorticity in four points of the domain is shown in Fig. 3.5. The locations considered are $(0.25, 0.25), (0.25, 0.75)$, $(0.75, 0.25)$ and $(0.75, 0.75)$. In each of these plots, the colored bands present are the ensemble standard deviations around the corresponding ensemble mean. In all measured points, forcing based on Gaussian noise produces the largest spread. It is clearly visible that using the OU process yields the smallest ensemble spread and using the estimated pdfs only slightly increases the spread compared to using the OU process.
The ensemble mean error and the ensemble standard deviation are shown in Fig. 3.6, where the ensemble mean error (3.21) is taken with respect to the

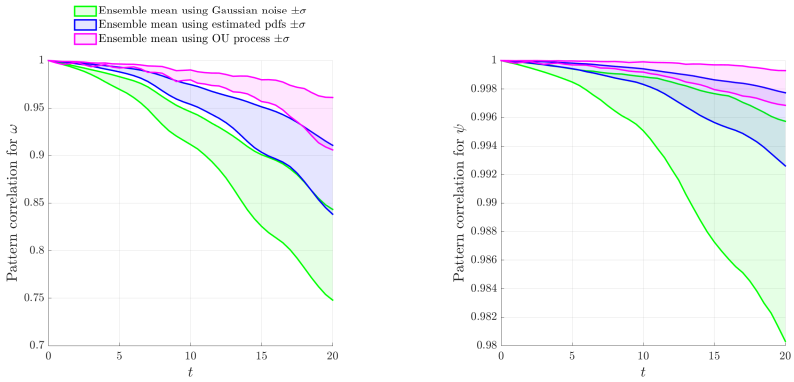


FIGURE 3.4: Pattern correlation (3.18) between the forecast ensembles and the adapted reference solution for the vorticity (left) and the streamfunction (right). Each band is defined as one ensemble standard deviation around the ensemble mean. The green band is generated using Gaussian noise, the blue band uses the estimated pdfs and the purple band uses OU processes. The results for each method are generated for an ensemble of 200 realizations.

adapted reference solution. It becomes evident that the mean error develops similarly for each ensemble. The mean errors for ensembles using the estimated pdfs and the OU process are nearly indistinguishable until $t = 10$, after which some smaller differences can be observed. In contrast, using Gaussian noise results in a much larger spread.
Fig. 3.7 shows the development of the streamfunction in the aforementioned points of the domain. The streamfunction is a smoother function than the vorticity, which is reflected in the smooth evolution of the former. In this figure it can also be observed that all ensembles accurately capture the adapted reference solution, with the OU model performing slightly better. The plots in Fig. 3.8 show the ensemble mean error and the ensemble standard deviation for the same points in the domain. Analogously to the vorticity, we find that the ensembles using the OU process and the estimated pdfs result in a smaller spread than the ensemble using Gaussian noise. Furthermore, it is observed that the ensemble mean error does not exceed the ensemble standard deviation before $t = 10$ and only does so occasionally after this point in time, indicating the reference solution is captured well by the ensembles.

In this subsection we have shown that the three considered stochastic processes accurately follow the adapted reference solution for multiple characteristic time units. Compared to Gaussian noise, using estimated pdfs or OU

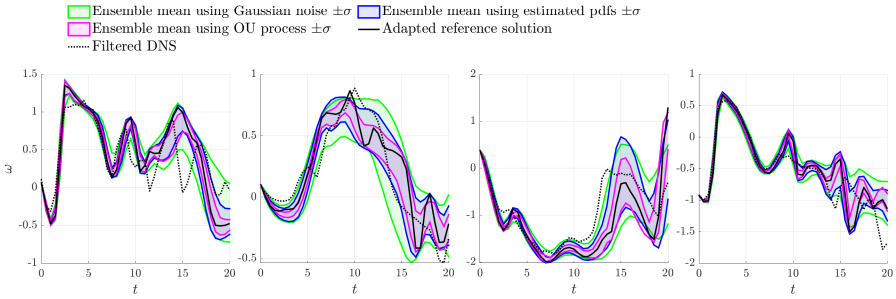


FIGURE 3.5: Vorticity measured on four points in the domain. From left to right, $(0.25, 0.25), (0.25, 0.75), (0.75, 0.25), (0.75, 0.75)$. The solid and dotted black lines show the development of adapted reference solution and filtered DNS, respectively. The green band is generated using Gaussian noise, the blue band uses the estimated pdfs and the purple band uses OU processes. The results for each method are generated from an ensemble of 200 realizations.

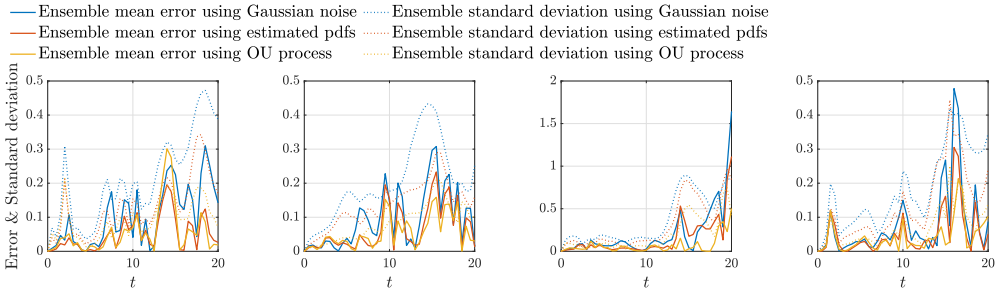


FIGURE 3.6: Ensemble mean error with respect to the adapted reference solution and ensemble standard deviation for the vorticity on four points in the domain. From left to right, $(0.25, 0.25), (0.25, 0.75), (0.75, 0.25), (0.75, 0.75)$. The ensemble mean error is depicted by the solid lines, the ensemble standard deviation by the dotted lines.

processes to define the stochastic forcing yielded a smaller spread of the ensemble forecast. Using a global measure, it is found that the latter two types of forcing yield ensembles that better resemble the adapted reference solution. In the next subsection, we perform additional statistical tests to assess short-time predictions.

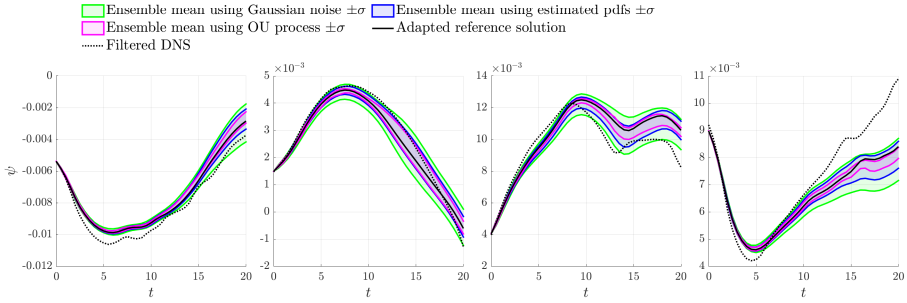


FIGURE 3.7: Streamfunction measured on four points in the domain. From left to right, $(0.25, 0.25), (0.25, 0.75), (0.75, 0.25), (0.75, 0.75)$. The solid and dotted black lines show the development of the adapted reference solution and filtered DNS, respectively. The green band is generated using Gaussian noise, the blue band uses the estimated pdfs and the purple band uses OU processes. The results for each method are generated from an ensemble of 200 realizations.

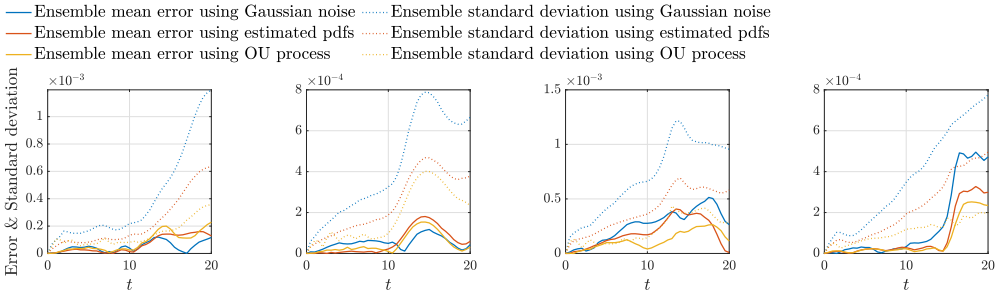


FIGURE 3.8: Ensemble mean error with respect to the adapted reference solution and ensemble standard deviation for the streamfunction on four points in the domain. From left to right, $(0.25, 0.25), (0.25, 0.75), (0.75, 0.25), (0.75, 0.75)$. The ensemble mean error is depicted by the solid lines, the ensemble standard deviation by the dotted lines.

3.3.4 Statistical tests for ensemble forecasts

Additional ensemble statistics are collected in order to further assess the numerical results of the SPDEs. In particular, forecast ensembles are generated for short lead times.

Two sets of initial conditions are generated to assess the stochastic models by sampling from two reference solutions: the filtered DNS and the adapted reference solution as presented in section 3.3.1. The filtered DNS does not contain

discretization and modelling error, whereas the adapted reference solution does. Therefore the use of both reference methods provides insight into the effects of these errors on the statistical quantities. Two distinct sets of initial conditions are acquired by sampling the reference solutions at $t = 0, 5, 10, \ldots, 350$, measured after the spin-up time. An ensemble forecast consisting of one hundred stochastic realizations is computed for each initial condition. Every stochastic realization is run for two time units and stored every 0.04 time units in order to study the results for short lead times. This time interval is similar to time intervals at which data may be assimilated [41]. Subsequently, the statistics are computed by comparing the ensembles to the corresponding reference solution. The statistics are provided below for both sets of initial conditions separately. As a first quantity we compute the root mean square error (RMSE). Recall that $\{X_{i,j}\}, i = 1, \ldots, N, j = 0, \ldots, T$ denotes an ensemble of N realizations measured at $T+1$ times. The RMSE between the ensemble mean of the SPDE and the reference solution is computed from

$$\mathrm{RMSE}(X_{i,j}, Y_j) = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(\langle X_j \rangle - Y_j)^2}. \tag{3.22}$$

This provides a measure for the average error of the ensemble [107]. The plots in Fig. 3.9 show the development of the RMSE and the spread (3.20) for increasing lead time for the different stochastic processes. In the left figure the stochastic ensembles are compared to the filtered DNS, in the right figure the ensembles are compared to the adapted reference solution. The RMSE values in the left graph of Fig. 3.9 show rapid growth, indicating that the ensemble mean deviates quickly from the filtered DNS. In contrast, the RMSE values obtained using the adapted reference solution show a significant error reduction. This suggests that the rapid error growth in the left figure is due to the fact that the gap between the coarse-grid SPDE and the filtered DNS contains not only the modelling error but also the discretization error. In addition, the right plot in Fig. 3.9 shows that using the estimated pdfs and the OU process yield similar values of the RMSE and the spreads develop comparably as well.

The second statistical quantity that we compute are rank histograms, which are a tool for measuring the reliability of an ensemble of forecasts [78]. A rank histogram is obtained by plotting the number of occurrences of particular outcomes of the rank function. Here, the rank function R keeps track of where the reference solution appears in the list of sorted ensemble members. That is, given a reference value Y_j and a list of N sorted ensemble members $\{X_{i,j}\}$, R is equal to the integer r that identifies the position of Y_j in the sorted list. It

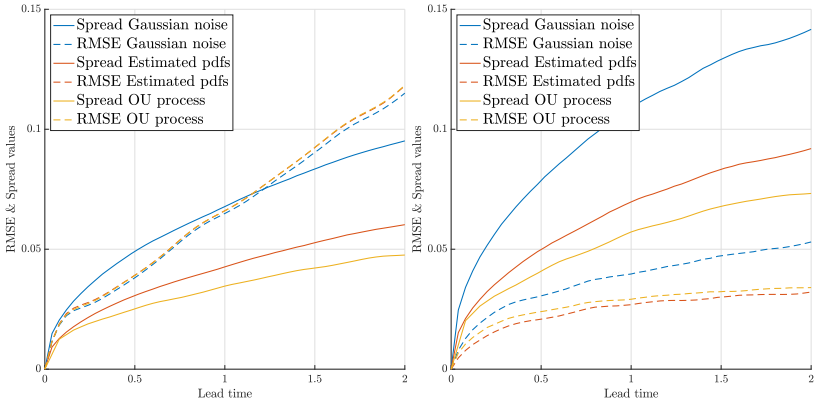


FIGURE 3.9: RMSE and spread as a function of time when comparing the stochastic ensembles to two different reference solutions. On the left, the filtered DNS is regarded as the reference solution and on the right the coarse simulation including the measured $\boldsymbol{\xi}_i$ is used. The data for each figure consists of 71 ensembles of 100 stochastic realizations each.

is defined as follows:

$$R\left(Y_j, \{X_{i,j}\}\right) = \begin{cases} r & \text{if } Y_j \geq X_{r,j}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.23}$$

If the forecast is reliable, then the reference value and the stochastic realizations are indistinguishable. This means that the underlying distributions of the reference value and the stochastic realizations are the same, which implies that the reference value is equally likely to be larger than any number of ensemble members. Thus, the rank function is equally likely to take on any value between 1 and N for reliable forecast ensembles and should therefore produce a rank histogram which approximates a uniform distribution.
Figures 3.10 and 3.11 show the rank histograms when using the filtered DNS and the adapted reference solution, respectively, as reference. The measurements at the points $(0.25, 0.25), (0.25, 0.75), (0.75, 0.25)$ and $(0.75, 0.75)$ at a lead time of 0.2 time units are used to generate the histograms. For each ensemble forecast the point values are compared to the reference solutions, leading to 284 ensemble outcomes that are compared to reference values. Only the rank histograms at this particular lead time are shown here, rank histograms at different lead times displayed similar results.
The rank histograms using the filtered DNS (Fig. 3.10) show clear peaks at the edges, caused by all ensemble members either overestimating or underestimating the truth. This effect is least pronounced when applying Gaussian

noise, due to the larger spread in the ensemble. The rank histograms obtained when comparing the ensembles to the adapted reference solution (Fig. 3.11) show peaks around the center. This is indicates that the reference solution ranks within the middle range of the ensembles. This is a direct result of the small mean error. The peaks at the edges are significantly reduced when using the adapted reference solution. This is especially clear when using estimated pdfs, which indicates that these ensembles, while showing a small spread, more accurately capture the reference solution. Overall, the differences between the rank histograms of the different methods are small. This indicates that reliability of the ensembles does not seem to depend on the choice of stochastic forcing.

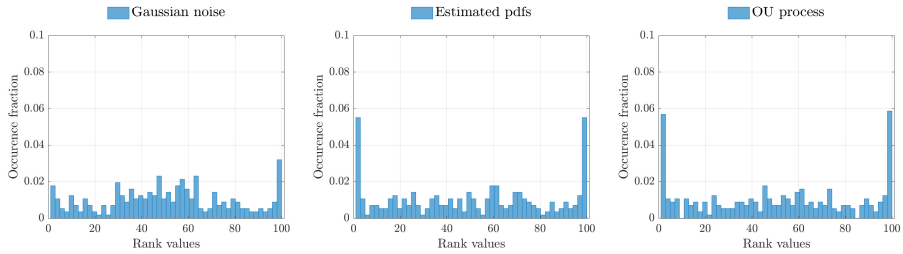


FIGURE 3.10: Rank histograms using measurements at the points $(0.25, 0.25), (0.25, 0.75), (0.75, 0.25)$ and $(0.75, 0.75)$ at a lead time of $t = 0.2$. A total of 71 ensembles are computed and measured at the specified points, each consisting of 100 stochastic realizations and compared to the filtered DNS at the corresponding time.

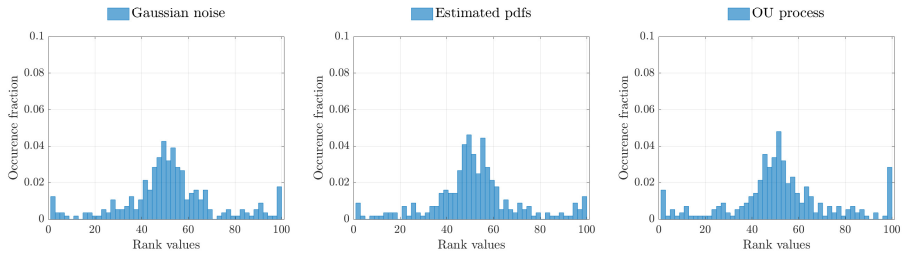


FIGURE 3.11: Rank histograms using measurements at the points $(0.25, 0.25), (0.25, 0.75), (0.75, 0.25)$ and $(0.75, 0.75)$ at a lead time of $t = 0.2$. A total of 71 ensembles are computed and measured at the specified points, each consisting of 100 stochastic realizations and compared to the adapted reference solution at the corresponding time.

The third statistical quantity that is presented here is the evolution of the vorticity over different time spans, conditioned on the vorticity value at a reference time. That is, the conditional probability distribution

$$P\left[\omega(t+\tau) - \omega(t) | \omega(t) = \omega_{\text{ref}}\right] \tag{3.24}$$

is estimated for different values of τ. This quantity describes the statistical evolution of the vorticity over a time interval of length τ, given a fixed initial configuration.

The conditional distributions are shown in Fig. 3.12, at lead time $\tau = 0.04$, and in Fig. 3.13, at lead time $\tau = 1$ to illustrate both short-time and long-time evolution. In both figures, the conditional distributions obtained from the reference solutions are shown in the left panel. For comparison, contour lines of these distributions have been overlaid in the conditional distributions obtained from the stochastic models. The filtered DNS provides the reference for the top row of distributions, the adapted reference solution is used in the bottom row. In particular, the distributions of the stochastic models have been computed from a set of initial conditions sampled along the filtered DNS and the adapted reference solution, respectively. In these figures, a large spread in the vertical direction indicates large uncertainty. This becomes especially clear for the shortest lead times considered. On such short timescales, the stochastic forcing adds considerable variance to the numerical solution. Applying Gaussian noise yields the largest spread, whereas using the estimated pdfs and the OU produce a smaller spread, in accordance with previously presented results. At lead time $\tau = 1$ (Fig. 3.13), the stochastic conditional distributions do not show significant differences. To better judge the agreement between the stochastic conditional distributions and the reference distributions, we compute the Hellinger distance. This measure allows for a quantitative comparison between the different distributions. Given two discrete probability distributions $p = (p_1, \ldots, p_K)$ and $q = (q_1, \ldots, q_K)$, we compute the Hellinger distance [83]

$$H^2(p, q) = \frac{1}{2} \sum_{i=1}^{K} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2 . \tag{3.25}$$

The distance $H^2(p, q)$ of (3.24) is shown in Fig. 3.14 for the filtered DNS (left figure) and for the adapted solution (right figure). The initial conditions of the stochastic ensemble and the reference solutions are the same, therefore the Hellinger distance at $\tau = 0$ is zero. As τ increases, $\omega(t)$ deviates from its reference value and accumulation of error leads to larger values of $H^2(p, q)$. Using the filtered DNS as reference solution yields a comparable Hellinger distance for each method. In contrast, the comparison of the stochastic ensembles to

the adapted reference solution clearly favours the models obtained using the estimated pdfs and OU processes over those where Gaussian noise is employed. Despite the quantitative difference in the Hellinger distance, the qualitative behaviour is the same for each of the stochastic models.
An overall smaller rate of increase is observed when comparing to the adapted reference solution with respect to the filtered DNS. The latter findings underpin once more the benefits of using the adapted reference solution when assessing the quality of different stochastic models.

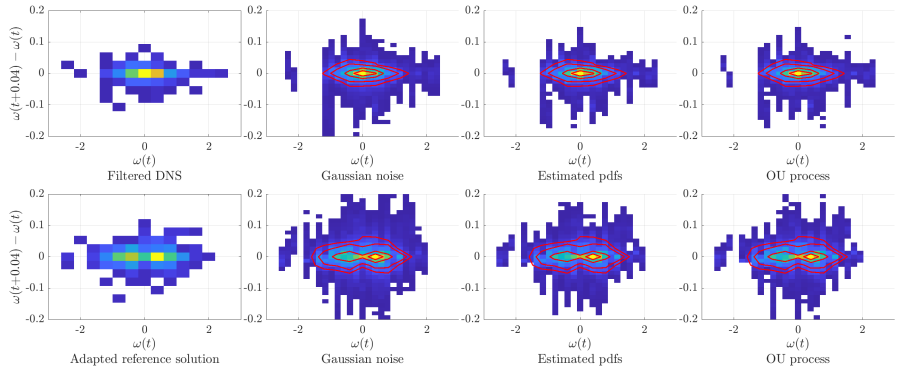


FIGURE 3.12: Conditional probability (3.24) for lead time $\tau = 0.04$. The top row shows the distributions using the filtered DNS as a reference, the bottom row uses the adapted reference solution. The contour lines of the reference conditional distributions are overlaid on the distributions obtained from the stochastic ensembles for easier qualitative comparison.

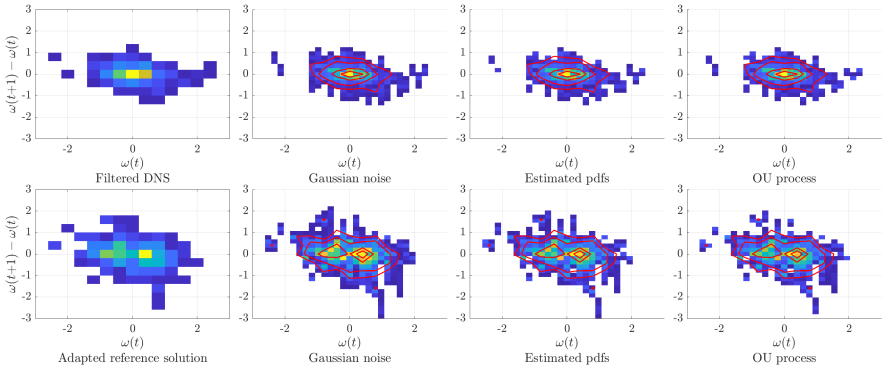


FIGURE 3.13: Conditional probability (3.24) for lead time $\tau = 1$. The top row shows the distributions using the filtered DNS as a reference, the bottom row uses the adapted reference solution. The contour lines of the reference conditional distributions are overlaid on the distributions obtained from the stochastic ensembles for easier qualitative comparison.

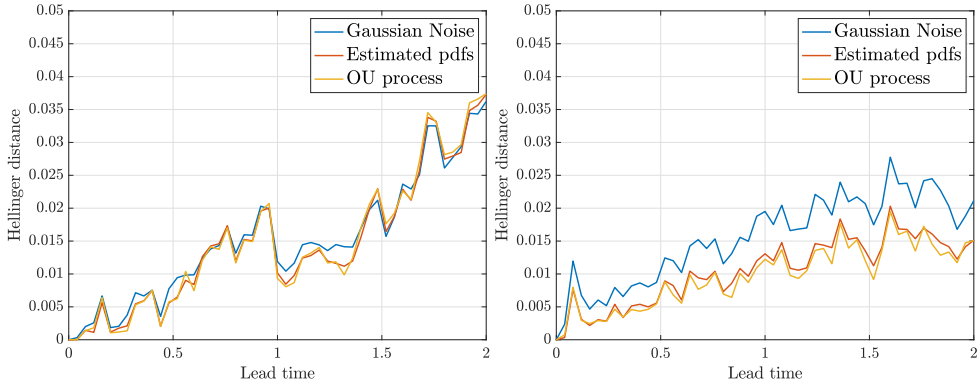


FIGURE 3.14: Hellinger distances as a function of time between the reference solution and the stochastic ensembles of distribution (3.24). On the left, the filtered DNS is used as a reference solution, on the right, the adapted reference solution provides the reference.

3.3.5 Quantitative assessment of results in spectral space

To distinguish the quality of the proposed models across different lengthscales we assess the outcomes of the models in spectral space. The EOF modes with a large energy content and a small energy content are representative of large lengthscales and small lengthscales, respectively. Therefore, one might discriminate between the lengthscales that are present in the solution by projecting

the latter onto the basis of EOF modes. This translates into applying the projection established in (3.16) for the reference solutions and the stochastic realizations and subsequently examining the obtained temporal coefficients. We perform uncertainty quantification at different lengthscales by comparing the EOF coefficients of the stochastic realizations to those of the reference solutions, computed for specified modes. The evolution of the coefficients of four modes, $i = 1, 10, 50, 150$, representative of large, intermediate and small scales, is shown in Fig. 3.15. It is found that the stochastic models accurately follow the adapted reference solution, but deviate somewhat from the filtered DNS result, independently of their lengthscale. Similar to the results in previous subsections, using the OU processes yields the smallest spread, followed by the estimated pdfs and the Gaussian noise.

To quantify the forecast quality as a function of time, for each EOF mode

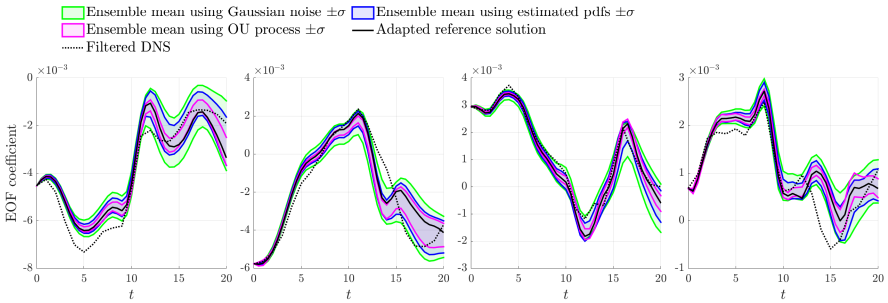


FIGURE 3.15: From left to right, EOF coefficients for modes 1, 10, 50, 150. The solid and dotted black lines show the development of the adapted reference solution and filtered DNS, respectively. The green band is generated using Gaussian noise, the blue band uses the estimated pdfs and the purple band uses OU processes. The results for each method are generated from an ensemble of 200 realizations.

separately, we define the root integrated mean-squared error (RIMSE):

$$\mathrm{RIMSE(t)} = \frac{1}{t} \frac{\left(\int_0^t \frac{1}{N} \sum_{i=1}^{N} \left(a_i - a_{\mathrm{ref}} \right)^2 \, \mathrm{d}\tau \right)^{1/2}}{\left(\int_0^t a_{\mathrm{ref}}^2 \, \mathrm{d}\tau \right)^{1/2}}. \tag{3.26}$$

This quantity is a measure of the difference between the EOF coefficient a_i of each stochastic realization in the ensemble and the coefficient a_{ref} of the reference solution, integrated over the specified time interval. The values of the RIMSE are shown in Fig. 3.16 for the four modes considered and compared to the adapted reference solution. The results suggest that using OU processes

and estimated pdfs is in general favoured over using Gaussian noise, largely independent of the lengthscale. For the higher modes, an initial rapid increase in the RIMSE is observed regardless of the employed method. The results obtained using Gaussian noise and OU processes show little difference for short lead times. For increased lead times the latter shows favourable results.

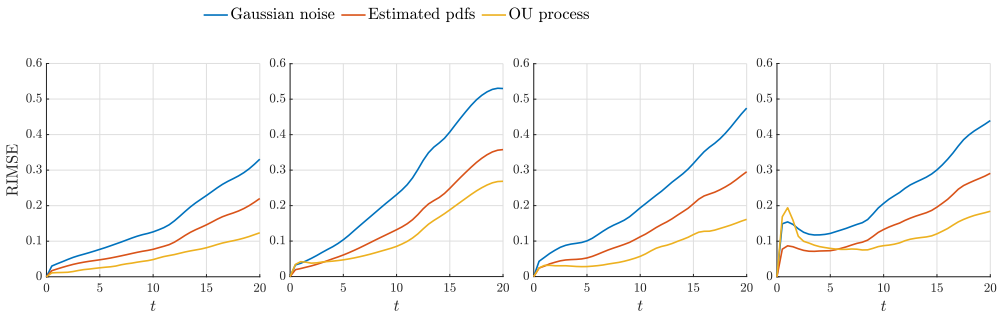


FIGURE 3.16: From left to right, RIMSE (3.26) for modes 1, 10, 50, 150, compared to the adapted reference solution. The results for each method are generated from an ensemble of 200 realizations.

An additional verification of the accuracy of the stochastic realizations is provided by comparing the means and variances of the EOF time series to those of the reference solutions. These values are shown in Fig. 3.17 for all EOF modes. Only the results using Gaussian noise are shown to keep the figures comprehensible. No significant differences were found for the other proposed models. The mean values of the stochastic realizations and the adapted reference solution are found to be nearly indistinguishable, whereas slight deviations from the mean values of the filtered DNS may be observed. The variances of the time series are also found to be in good agreement with those of the adapted reference solution, but differ marginally from the variances of the time series of the filtered DNS.

3.4 Concluding remarks

In this chapter, we have assessed three stochastic models for the simulation of the coarse-grained two-dimensional Euler equations. The closure is based on the so-called Stochastic Advection by Lie Transport (SALT) approach. The resulting SPDE contains a stochastic forcing term that requires to be modelled to close the equations. In particular, the forcing is decomposed into a deterministic basis (empirical orthogonal functions, or EOFs) multiplied by

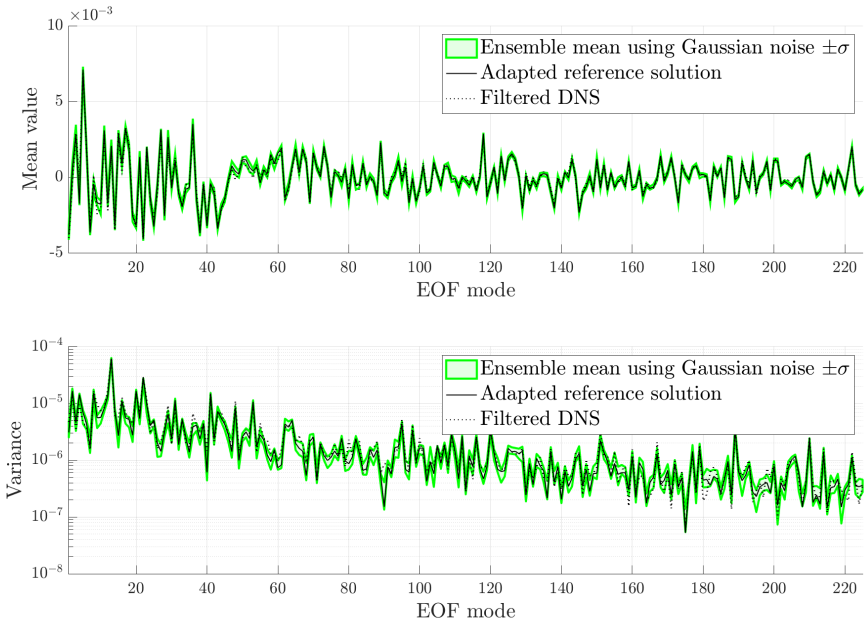


FIGURE 3.17: Mean values (top) and variances (bottom) of the EOF time series over the simulated time interval. The solid and dotted show the development of the adapted reference solution and the filtered DNS, respectively. The green bands are generated from an ensemble of 200 realizations using Gaussian noise and show the ensemble mean of the quantity of interest ± one ensemble standard deviation.

stochastic temporal traces. This decomposition is, by construction, fully determined from a fine-grid (DNS) dataset. However, to simulate outside the available dataset one is required to model the time traces. In the framework of SALT [40] the latter are regarded as Gaussian processes. Here we extend the stochastic forcing to more general processes, sampling from the data-estimated probability distribution functions (pdfs) and introducing correlation through Ornstein-Uhlenbeck (OU) processes. The latter two methods use additional data already available from the EOF time series. Between the methods no qualitative differences in the flow realizations were observed. However, the latter methods generally show favorable results compared to the former Gaussian method, in terms of ensemble mean and ensemble spread.
To meaningfully compare the different stochastic models we defined a maximal prediction horizon and an adapted reference solution. The prediction horizon sets the point in time beyond which a bundle of fine-grid solutions, starting

from the same initial condition on the coarse grid, deviates on order 1 due to high sensitivity to the initial conditions. This defines the time frame on which to assess the statistical quality of the coarse-grid predictions. The adapted reference solution was defined as the coarse-grid solution using the exact measured time series of the EOFs for the forcing. The latter allowed us to isolate the modelling error from other sources of error not taken into account in the considered model formulation, such as discretization error. The stochastic ensembles were compared to this reference solution using a global measure and pointwise values. For both the global and local measures, using either estimated pdfs or OU processes to define the forcing term yielded a smaller ensemble mean error and a smaller spread compared to using Gaussian noise.
Stochastic prediction ensembles on timescales relevant for data assimilation were further investigated by performing statistical tests, comparing ensembles of stochastic realizations to the adapted reference solution and the filtered DNS. A significantly smaller ensemble spread was found when using estimated pdfs or OU processes, compared to using Gaussian noise. Additionally, the observed mean ensemble error was lower for the former two methods. All three methods showed rapid growth in ensemble error when compared to the filtered DNS, suggesting that the filtered DNS contains not only the modelling error but also the the discretization error and the closure error. These results were further substantiated by rank histograms, showing that the ensembles were biased with respect to the filtered DNS, but were overdispersive compared to the adapted reference solution. In particular, using the estimated pdfs to define the stochastic forcing rarely resulted in the adapted reference solution not being contained in the ensemble. Finally, conditional distributions of the vorticity were computed and compared using the Hellinger distance. Here, using estimated pdfs or OU processes resulted in a smaller distance to the reference solution than using Gaussian noise, indicating a better statistical characterization of the vorticity dynamics.
The ensemble forecasts were assessed in spectral space to discriminate between the different lengthscales present in the solution. The stochastic ensembles were found to accurately capture the adapted reference solution on the considered scales. The overall prediction quality using OU processes was found to be favourable over using Gaussian noise, independent of the lengthscale. Additionally, the stochastic ensembles were found to accurately reproduce the mean values and variances of the EOF time series over the entire simulation period.
The methods presented in this chapter may be used in other flows where EOF-based stochastic modeling is relevant. These approaches are particularly appealing since all information used in these methods is readily available from the EOF decomposition and no additional data is required to construct the

models. The presented techniques are purely data-driven, they require no further assumption about the governing equations and can therefore be applied to other geophysical fluids. The short-time results indicate that a mean error reduction and smaller ensemble spread can be obtained using these methods, which can complement methods employed in data assimilation. Furthermore, the definition of the adapted reference solution motivates further research of the SALT method using different closure models and incorporating the discretization error.

Chapter 4

Data-driven spectral modeling for coarsening of the 2D Euler equations on the sphere

4.1 Introduction

Two-dimensional incompressible hydrodynamics models are fundamental for studying physical phenomena in atmospheric and oceanic flows. Typical examples include the two-dimensional Euler equations, quasi-geostrophic equations, and (rotating) shallow water equations. A characteristic feature of these flows is the formation of both large vorticity structures through the inverse energy cascade and small-scale vorticity filaments through the enstrophy cascade [173]. In realistic conditions, the energy spectrum extends over several orders of magnitude, making it computationally infeasible to fully resolve all scales that are present in the flow. Simplifications are required, either by reducing the complexity of the underlying mathematical model [111] or by reducing the spatial or temporal resolution with which the dynamics are resolved [23]. In this chapter, we will focus on high-fidelity coarsening of the two-dimensional Euler equations on the sphere by applying an *online/offline* approach to obtain accurate coarse-grained numerical solutions of statistically steady states. In particular, explicit information on well-resolved dynamics is obtained from high-resolution simulations in the offline phase, which is applied in an online control feedback model for accurate coarse-grained simulations.

There is considerable interest in achieving accurate numerical solutions of fluid flows at reduced computational costs [68]. This forms the main challenge of Large Eddy Simulation (LES), which aims to provide skillful large-scale predictions of complex flows by numerically solving spatially filtered momentum equations. Often, a model term is included to compensate for unresolved dynamics due to coarsening to retain a sufficiently detailed description of turbulent flows at high Reynolds numbers [74, 147, 70, 138, 146]. The growing

availability of computational resources has facilitated the use of high-resolution direct numerical simulations (DNS) as a source of data from which coarse-grid fluid models may be derived. Data-driven LES methods have successfully been developed in recent years, for example, by using neural networks to compute a variable eddy viscosity [11] to approximate a reference kinetic energy spectrum [101] or to model subgrid-scale forces [170]. Alternatively, approaches based on interpolation of small high-resolution patches of the spatial domain [28, 26] and data-driven residual modeling via global basis functions [57] have also shown computational efficiency and accuracy in coarse-grained numerical solutions.
Data assimilation provides an alternative method to achieving accurate coarse-grained results by combining predictions with real-time observations. In continuous data assimilation (CDA), observational data is incorporated into the prediction while the numerical model is being integrated in time [4, 31, 49]. Specifically, the difference between the numerical prediction and the corresponding observation determines a nudging term that is added to the governing equations. Studies on nudging of dissipative fluid models have shown that a range of nudging strengths may be chosen that all yield an accurate coarse-grained representation of the true solution [7, 4]. Adaptive nudging strengths based on energy balance have also been proposed [174] resulting in faster convergence towards the reference compared to a simulation that exploits a constant nudging strength. Since these models rely on observational data to achieve high-fidelity coarsened solutions, the uncertainty originating from measurement errors has to be taken into account, as well as possible accumulation of discretization errors [73].
Models of geophysical fluid flows often employ stochasticity as a means to model uncertainty inherent to flows [136]. Uncertainty arises predominantly from differences in initial conditions, errors in measurements, and model incompleteness. Low-dimensional models describing qualitative features of geophysical fluid flows often serve as a test bed for stochastic forcing. For example, stochastic forcing based on subgrid data in the two-scale Lorenz '96 system resulted in improved forecasting skill compared to deterministic parametrizations [6]. Ultimately, the exact way in which stochasticity is included in numerical simulations remains a modeling choice and may lead to qualitatively different effects on the dynamics [71]. These approaches have also been applied successfully to more complete geophysical models. Examples include the modeling of uncertainty through Casimir-preserving stochastic forcing for the two-dimensional Euler equations [40, 56, 35] and energy-preserving stochastic forcing in the quasi-geostrophic equations [145]. An alternative approach is based on statistics of subgrid data that lead to a stochastic forcing and eddy viscosity, which has been applied to the barotropic vorticity equation on the sphere [62]. This approach was found to accurately model uncertainty and

produce energy spectra on coarse computational grids that closely match reference high-fidelity simulations at much higher resolutions.
In this chapter, we propose an *online* data-driven standalone stochastic model for coarse numerical simulations of statistically steady states of the two-dimensional Euler equations on the sphere. Data of a statistical equilibrium is extracted from an *offline* high-resolution precursor simulation in the form of statistics of coefficients of spherical harmonic modes and is included as a stochastic forcing term closely following the formulation of the continuous-time limit of the 3D-Var algorithm [45] as presented in [17]. Similar to data-driven LES, a modeling term is added to the coarsened numerical simulation based on these *a priori* collected data. This term models the unresolved interactions between the modes as a linear stochastic process for each spherical harmonic coefficient separately and is designed to reproduce the energy spectrum of the high-resolution simulation. Like CDA, the model term is included as a feedback control term. This term nudges the coarse grid solution towards a known reference solution, chosen here as the statistically steady state. We opt for the nudging strength to be equal to the inverse of the characteristic time scale of the corresponding spherical harmonic mode. This choice has the benefit that it mimics the measured temporal correlation. The nudging procedure is performed via a prediction-correction scheme in which we first fully complete a time integration step involving all true fluxes and subsequently we apply the nudge as a correction to the predicted solution. This results in straightforward implementation in existing computational methods and leads to a numerical scheme of the same form as the diagonal Fourier domain Kalman filter [80, 115] with prescribed gain. Striking features of the high-fidelity reference solution were captured in the coarser model using this stochastic model.
The chapter is structured as follows. The two-dimensional Euler equations and the adopted numerical method are introduced in Section 4.2. In Section 4.3 we describe the model and focus in particular on how the model parameters are specified. In Section 4.4, we define the reference solution and apply the model at two coarse resolutions. The results are assessed qualitatively and by means of statistics of Fourier coefficients. Subsequently, we show that the model is capable of reproducing large-scale vortex dynamics of the reference solution. Section 4.5 concludes the chapter and suggests directions for further research.

4.2 Governing equations and numerical methods

The model that will be studied in this work is given by the two-dimensional Euler equations on the unit sphere $\mathbb{S}^2$. These equations arise as the two-dimensional Navier-Stokes equations in the inviscid limit and describe vortex

dynamics [173]. The dynamics are given in streamfunction-vorticity formulation by

$$\begin{aligned} \dot{\omega} &= \{\psi, \omega\}, \\ \Delta\psi &= \omega. \end{aligned} \tag{4.1}$$

Here ω is the vorticity, ψ is the streamfunction, and $\{\cdot,\cdot\}$ is the Poisson bracket. The vorticity and the streamfunction are related via the Laplace operator Δ. The vorticity relates to the fluid velocity $\mathbf{v}$ via $\omega = \operatorname{curl} \mathbf{v}$. These equations are part of a larger family of geophysical fluid models that can be derived from a variational principle and inherently reflect particular conservation laws [87]. The governing equations (4.1) form a Lie-Poisson system [117] with a Hamiltonian $\mathcal{H}$ and an infinite number of conserved quantities, known as Casimirs $\mathcal{C}_k$, given by

$$\mathcal{H}(\omega) = -\frac{1}{2}\int \omega\psi, \tag{4.2}$$

$$\mathcal{C}_k(\omega) = \int \omega^k, \quad k = 1, 2, \dots \tag{4.3}$$

where the integral is taken over the spatial domain.
A discrete system with a similar Lie-Poisson structure is obtained after so-called geometric quantization. This structure-preserving discretization is based on a finite truncation of the Poisson bracket, as proposed in [171, 172] and rests on the theory of quantization [88, 21, 20]. First, an $N > 1$ is chosen, which can be thought of as the numerical resolution. Subsequently, a total of $\frac{N(N+1)}{2} - 1$ global basis functions are determined explicitly before carrying out a simulation. These functions serve to construct the discrete vorticity representation W. A finite-dimensional approximation of the system (4.1) is obtained as

$$\begin{aligned} \dot{W} &= [P, W], \\ \Delta_N P &= W. \end{aligned} \tag{4.4}$$

Here W is the vorticity matrix, P is the stream matrix and $W, P \in \mathfrak{su}(N)$, that is, skew-Hermitian, traceless $N \times N$ matrices.

The discrete system (4.4) is interpreted as follows. A continuous vorticity field ω on the sphere can be expanded in a spherical harmonic basis $\{Y_{lm}\}$ as $\omega = \sum_{l,m} c_{lm} Y_{lm}$. The spherical harmonic coefficients c_{lm} are used to construct the matrix W. Namely,

$$W = \sum_{l=0}^{N-1} \sum_{m=0}^{l} c_{lm} T_{lm}^N, \tag{4.5}$$

Here, $\{T_{lm}^N\}$ is the so-called quantized spherical harmonic basis [37], which provides a particular discrete approximation to the spherical harmonic basis

$\{Y_{lm}\}$. In fact, the quantized representation enables the structure-preserving discretization [172, 120]. The basis element T^N_{lm} is a sparse skew-Hermitian traceless matrix, nonzero only on the m-th sub- and superdiagonal. We refer to [37] for a detailed description of the quantized basis. The quantized Laplacian Δ_N can be derived as a complicated expression, given in [89]. The matrix P then follows by applying the inverse quantized Laplacian to W. The bracket $[P, W] = PW - WP$ is the standard matrix commutator. In the limit of $N \to \infty$, the structure constants of the Lie algebra $\mathfrak{su}(N)$ converge to those of $C^\infty(\mathbb{S}^2)$ expressed in terms of spherical harmonics. This convergence implies that smooth functions on the sphere can be approximated by finite-dimensional matrices by means of Eq. (4.5) [37]. The discrete system is a Lie-Poisson system with a Hamiltonian H and N conserved quantities C_k,

$$H(W) = \frac{1}{2}\mathrm{Tr}\,(PW)\,, \tag{4.6}$$

$$C_k(W) = \mathrm{Tr}\left(W^k\right), \quad k = 1, \ldots, N. \tag{4.7}$$

Equations (4.4) are solved numerically using the second-order isospectral midpoint rule [163, 127], using the parallelized implementation described in [37]. This is a Lie-Poisson integrator, conserving the N discrete Casimir functions exactly. Given a time step size h, a time integration step proceeds as follows

$$\begin{aligned} W_n &= \left(I - \frac{h}{2}\Delta_N^{-1}\tilde{W}\right)\tilde{W}\left(I + \frac{h}{2}\Delta_N^{-1}\tilde{W}\right) \\ W_{n+1} &= \left(I + \frac{h}{2}\Delta_N^{-1}\tilde{W}\right)\tilde{W}\left(I - \frac{h}{2}\Delta_N^{-1}\tilde{W}\right), \end{aligned} \tag{4.8}$$

i.e., given W_n the intermediate solution $\tilde{W}$ is obtained first, after which W_{n+1} is determined to complete a time-step.

An example of the Euler equations integrated at high resolution using (4.8) is given in Fig. 4.1. This figure shows the vorticity fields as Hammer projections in order to display the entire spherical domain. The flow dynamics reveal that large-scale low-dimensional structures are present in the vorticity field at late times [128]. This motivates the use of coarse computational grids to capture the dynamics in the asymptotic time regime. A depiction of this is provided in Fig. 4.1, showing late stages in the evolution of a high-resolution numerical simulation initialized from a random vorticity field in which only large scales are present. After a period of vorticity mixing the solution reaches a statistically steady state, in which large-scale vorticity structures have emerged and persist.

To compare numerical solutions at different resolutions, we define a fine-to-coarse filter. Throughout the chapter the applied filter is a spectral cut-off filter, setting all coefficients corresponding to a wavenumber larger than a

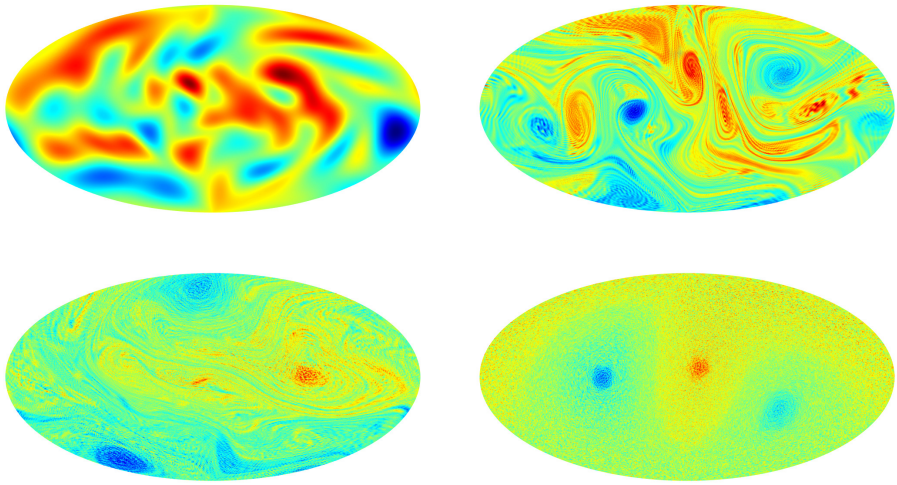


FIGURE 4.1: Snapshots of a high-resolution ($N = 512$) numerical simulation of system (4.4). The vorticity field is initialized as a random large-scale field (top left), after which it undergoes a period of vorticity mixing (top right, bottom left) before reaching a statistically steady state in which large-scale vorticity structures dominate the solution (bottom right).

specified wavenumber to zero. In the following, we consistently choose a cut-off wavenumber defined by the coarse-grid resolution, which yields a filtered solution containing only spatial scales resolvable on the corresponding coarse grid.

Significantly decreasing the resolution yields a qualitatively different statistically steady state, as shown in Fig. 4.2. Illustrated is a snapshot of the fine-grid solution ($N = 512$), a filtered version thereof (only the components up to $N = 64$ are shown), and a snapshot of a coarse-grid solution ($N = 64$) using the algorithm as outlined in (4.8). These simulations are initialized using the smooth vorticity field in Fig. 4.1. High-frequency components are visible in the snapshot of the high-resolution numerical solution, which develop as a result of the enstrophy cascade. By applying a spectral cut-off filter to the fine numerical solution we obtain a smooth vorticity field. By definition of the filter, this field can be fully resolved using the coarse resolution. Since the filtered fine solution is an orthogonal projection onto the coarse-resolvable subspace of solutions, it defines the best attainable result on the coarse grid and the result is a description of the large-scale components of the flow, influenced by all fine-grid resolvable scales.

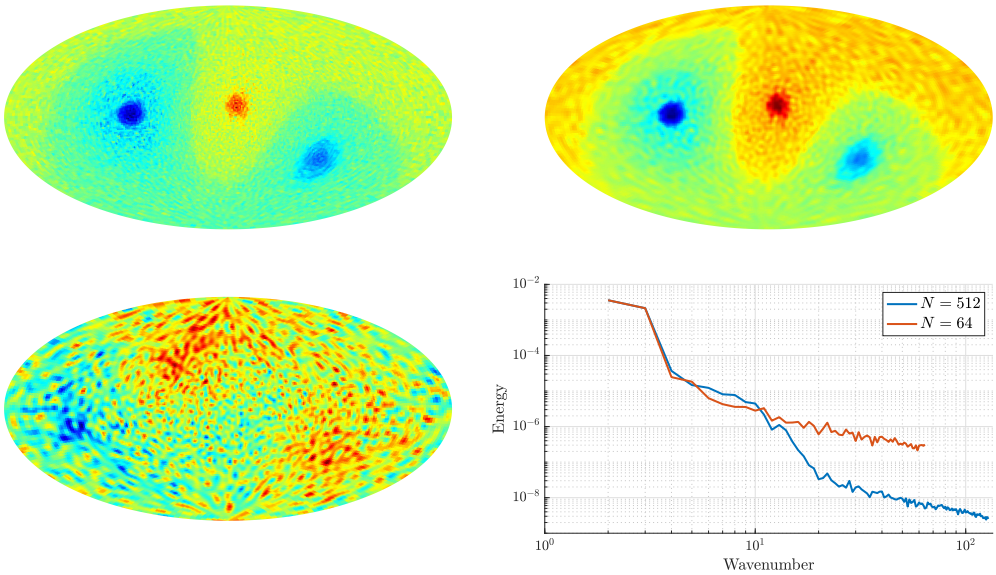


FIGURE 4.2: Snapshots of the fine vorticity field (top left), a filtered version thereof (top right) and a coarse vorticity field (bottom left) after reaching a statistically steady state. The energy spectra of the fine and coarse fields are shown in the bottom right panel.

A clear qualitative difference exists between the filtered high-resolution vorticity and the vorticity obtained at a lower resolution, which is best explained by analyzing the energy spectra. The energy spectrum of the coarse numerical solution deviates from the spectrum of the fine numerical solution at the smallest resolvable scales. Nonetheless, the energy in the large scales is captured well. Additionally, the energy decay at large wavenumbers follows the same decay of l^{-1} as observed in the high-resolution result, where l is the wavenumber. Despite this agreement, the instantaneous vorticity field at the coarse resolution differs significantly from the high-resolution result. The increased energy in high-frequency modes of the coarse numerical solution causes the small scales to dominate the vorticity field observed in Fig. 4.2. We note that the energy spectrum of the filtered reference solution exactly coincides with the spectrum of the reference solution until the cut-off frequency at wavenumber 64, by definition of the filter. The discrepancy between the energy levels of the coarse numerical solution and the filtered reference may therefore be reduced by an appropriate forcing term. In the next section, we introduce a data-driven forcing term that yields the desired energy level at each frequency and thus regularizes the numerical solution.

4.3 Data-driven spectrum-preserving forcing

In the previous section, we observed that a qualitative difference exists between the statistically steady states obtained at low and high resolutions. One of the defining features of a statistically steady state is its kinetic energy spectrum. The corresponding energy spectra of the solutions reveal a considerable difference between the energy levels of the small scales present in the flow. The discrepancy between the spectra may be reduced by introducing an appropriate forcing or correction term to the coarse numerical simulations. In a statistically steady state, the spectrum can be fully described using the mean and variance of the magnitude of the spectral coefficients. Therefore, the goal of the model is to reproduce these quantities accurately and, in doing so, recover the reference kinetic energy spectrum. In this section, we describe a forcing that achieves this goal, particularly in situations where the number of modes is kept low. For this purpose, we opt for a model that aims to match the mean and variance of the coefficient magnitudes to reference values. This approach is based on reference data corresponding to independently obtained highly resolved direct numerical simulations aiming to combine computational feasibility with accurate flow predictions.

To define the spectral forcing, we expand the vorticity matrix in the quantized spherical harmonic basis $\left\{T_{lm}^N\right\}$,

$$W(t) = \sum_{l=0}^{N-1} \sum_{m=0}^{l} c_{lm}(t) T_{lm}^N, \tag{4.9}$$

with complex coefficients $\{c_{lm}\}$. The energy in solution components at index l is then defined as

$$E_l(t) = \sum_{m=0}^{l} |c_{lm}(t)|^2. \tag{4.10}$$

The index l will also be referred to as 'wavenumber' following the analogy with an expansion in spherical harmonics and plane waves [122]. The expansion (4.9) allows us to formulate the equations of motion (4.4) in terms of the basis coefficients c_{lm}, as

$$\dot{c}_{lm} = L(\boldsymbol{c}, l, m), \tag{4.11}$$

where L is the spectral representation of $[P, W]$ and $\boldsymbol{c}$ the vector containing all basis coefficients. In particular, the evolution of the magnitude of c_{lm} will be expressed as $L_r(\boldsymbol{c}, l, m)$. A special feature of our approach is that the time stepping acts on W, while the model is applied in spectral space. In the actual algorithm, a mapping between elements of $\mathfrak{su}(N)$ and their representation as quantized spherical harmonic coefficients is needed for this purpose. Therefore

the operators L and L_r are not required to be explicitly defined or evaluated but serve only to simplify notation.

A mean-reverting forcing is introduced into the evolution of the coefficient magnitudes, to ensure that the magnitudes tend to a specified mean value. Forcing the magnitudes of the basis coefficients is pragmatic since these are stationary if the solution is in a statistically steady state. Mean reversion is realized by adding an Ornstein-Uhlenbeck (OU) process to the evolution of the coefficient magnitude. This way, the reference spectrum can be reproduced in a coarse numerical simulation. It has been shown [17] that the OU process arises in the governing equations as the continuous-time limit of the 3D-var data assimilation algorithm [45]. We thus propose

$$\mathrm{d}|c_{lm}| = L_r(\boldsymbol{c}, l, m)\mathrm{d}t + \frac{1}{\tau_{lm}}\left(\mu_{lm} - |c_{lm}|\right)\mathrm{d}t + \sigma_{lm}\mathrm{d}B^t_{lm}, \tag{4.12}$$

where μ_{lm} and τ_{lm} are means and correlation times extracted from a separate high-resolution simulation. In fact, from a sequence of solution snapshots time series are obtained for each of the basis magnitudes $|c_{lm}|$, of which μ_{lm} is the mean value and τ_{lm} is the characteristic time scale. The relaxation of the forcing is determined by the time scale τ_{lm}. Deviations of $|c_{lm}|$ from the mean μ_{lm} are nudged back in order to reduce the differences. Randomness is introduced via the term $\mathrm{d}B^t_{lm}$ in which B^t_{lm} is a general random process, defined for each pair l, m separately. The random process can be tailored to fit the measurement data [56], though the common choice is to let $\mathrm{d}B^t_{lm}$ be normally distributed with a variance depending on the time step size [84]. We choose the latter in what follows and include the variance scaling in σ_{lm}. The value of σ_{lm} depends on the sample variance of the time series, on τ_{lm} and on the adopted time step size and will be specified later in this section.

In the discrete setting, we apply the forcing defined by the OU process in (4.12) as a correction after time step is completed. This alters a time-advancement step as follows. Starting from the vorticity W^n at time level t^n, a prediction $\bar{W}^{n+1}$ of the vorticity at the next time level is obtained by integrating Eq. (4.4) over one time step using the algorithm (4.8). This prediction is then projected onto the basis of spherical harmonics to obtain the corresponding basis coefficients $\{\bar{c}^{n+1}_{lm}\}$. Finally, a correction is applied to these coefficients using (4.12) to obtain $\{c^{n+1}_{lm}\}$ which are then used to construct the vorticity field W^{n+1} at the new time level. We note that the correction is only applied to the magnitude of the basis coefficients. The parameter definitions in the implementation of (4.12) will now be described.

The correction procedure (4.12) will be referred to as *nudging*. We distinguish between *deterministic nudging*, using only the deterministic component of the forcing, and *stochastic nudging*, using both the deterministic and the

stochastic component. The former is described as

$$|c_{lm}^{n+1}| = |\bar{c}_{lm}^{n+1}| + \frac{\Delta t}{\tau_{lm}} \left(\mu_{lm,\text{det}} - |\bar{c}_{lm}^{n+1}| \right). \tag{4.13}$$

The stochastic nudge is defined as

$$|c_{lm}^{n+1}| = |\bar{c}_{lm}^{n+1}| + \frac{\Delta t}{\tau_{lm}} \left(\mu_{lm,\text{stoch}} - |\bar{c}_{lm}^{n+1}| \right) + \sigma_{lm} \Delta B_{lm}^{n}, \tag{4.14}$$

where ΔB_{lm}^{n} is drawn from a standard normal distribution for each l, m and n independently.

The nudging procedures in Eqs. (4.13, 4.14) can be characterized as a steady-state Kalman-Bucy filter [77] with prescribed gain $\Delta t/\tau_{lm}$. The value of τ_{lm} is chosen to be constant, similar to steady-state filters. At each time step, the 'observation' consists of coefficients for each spherical harmonic mode separately. The deterministic nudging procedure assumes the observation is a fixed value $\mu_{lm,\text{det}}$, whereas the stochastic approach adopts observations as distributed samples. Here, we use $\mathcal{N}\left(\mu_{lm,\text{stoch}}, \sigma_{lm}^2\right)$ as distribution and draw independent samples for each l, m, n separately. Thus, the unresolved interactions between different spherical harmonic modes are modeled as linear stochastic processes, independent for each value of l, m. This approach has been introduced as Fourier domain Kalman filtering [115]. For low-dimensional systems it was analyzed in Fourier space [80, 29] and also shown to be feasible for filtering high-dimensional systems.

In the continuous formulation (4.12) τ_{lm} can take on any positive value. In the discrete form (4.13, 4.14), τ_{lm} can take on values in the interval $[\Delta t, \infty)$. For $\tau_{lm} = \Delta t$ the forcing ensures that the magnitude $|c_{lm}|$ of the corresponding coefficient becomes constant in the case of deterministic nudging. In the case of stochastic nudging, this value of τ_{lm} ensures that $|c_{lm}|$ evolves as Gaussian noise with the specified mean and variance. In the limit of large τ_{lm} the forcing approaches zero and the unforced dynamics is retained.

The nudging procedures in Eqs. (4.13, 4.14) are treated as first-order autoregressive models with drift coefficient $(1 - \Delta t/\tau_{lm})$ and mean $\mu_{lm,\text{stoch}}$, which is a discretization of the OU process (4.12). The value of τ_{lm} is found by fitting the autocovariance function of the OU process to the sample autocovariance as obtained from the reference high-resolution simulation. The value of τ_{lm} is expected to decrease as larger wavenumbers l are considered. This increases the contribution of the model term to the dynamics of the coefficients c_{lm} at those wavenumbers. Therefore, with increasing spatial resolution one will resolve finer lengthscales associated with larger l, whose contributions correspond closer and closer to the direct observations. This is in accordance with

theoretical results for filter performance [115].
The values of $\sigma_{lm}, \mu_{lm,\text{stoch}}$ and $\mu_{lm,\text{det}}$ are chosen so that the reference energy spectrum is reproduced when the model is applied. Treating $|c_{lm}|$ as a stochastic variable, we observe that $\mathbb{E}(|c_{lm}|^2)$ is the expected energy content of the basis element T_{lm}^N. Through the definition of the variance we find that

$$\mathbb{E}(|c_{lm}|^2) = \text{var}(|c_{lm}|) + \mathbb{E}(|c_{lm}|)^2 \tag{4.15}$$

We define σ_{lm} so that variance of the autoregressive model coincides with the sample variance s_{lm}^2 of the reference time series, i.e.,

$$\sigma_{lm} = s_{lm}\sqrt{1 - \left(1 - \frac{\Delta t}{\tau_{lm}}\right)^2}, \tag{4.16}$$

where s_{lm} is the sample standard deviation of $|c_{lm}|$ as obtained from the high-resolution simulation. To obtain the desired energy content, $\mu_{lm,\text{stoch}}$ is subsequently chosen as $\mathbb{E}(|c_{lm}|)$. In the case of deterministic nudging, the variance of $|c_{lm}|$ vanishes when $\tau_{lm} = \Delta t$. To obtain the desired energy content in this limit, $\mu_{lm,\text{det}}$ is chosen as $\sqrt{\mathbb{E}(|c_{lm}|^2)}$. The mean, variance, and correlation time are estimated using standard unbiased estimators of which the mean squared error decreases linearly with the number of used samples. It is assumed that the mean and variance are constant in time, therefore requiring that the flow is in a statistically steady state.

For each basis function only three parameters need to be measured: the mean, the variance and the correlation time. This outlines the simplicity of the model. These parameters are inferred from the data, do not require additional tuning and are defined up to the resolution of the reference solution. Furthermore, the basis of spherical harmonics is resolution-independent. Therefore the forcing parameters only depend on the reference data and not on the choice of the coarse-grid resolution or time step size, implying that the model is self-consistent [63]. This is further corroborated in a later section of the chapter by applying the model at various low resolutions.

4.4 Numerical experiments

In this section, we apply the forcing proposed in Section 4.3 to coarse numerical simulations. We describe the reference solution and introduce the measured variables that constitute the model data. The forcing is applied at different coarse computational grids using several model configurations. The model results are compared to the reference solution and the no-model coarse numerical solution and are assessed in terms of statistical quantities of the resulting time

series of the basis coefficients. Finally, we illustrate that application of the model yields accurate long-time solutions on coarse computational grids.

4.4.1 Description of reference solution

The reference solution is acquired from the discretized equations described in Section 4.2 and adopts a resolution $N = 512$. The initial condition is the smooth vorticity field as shown in the left panel of Fig. 4.3, which is also adopted in later numerical simulations using lower resolutions. This initial condition is randomly generated and contains only large scales of motion. The vorticity is evolved until $t = 6500$, shown on the right panel of Fig. 4.3, at which a statistically steady state is reached. This was verified by averaging the kinetic energy spectrum over several time durations. High-resolution snapshots are collected every time unit after reaching this state. A total of 1000 snapshots is collected to ensure that estimates of the mean, variance, and correlation times are sufficiently accurate.
By projecting the snapshots onto the basis $\{T_{lm}^N\}$ a time series of coefficients for each spherical harmonic mode is obtained. These coefficients are complex-valued, however, in what follows we will only consider the time series of the corresponding magnitudes since these are the quantities that the proposed model acts on.
The forcing parameters are shown in Fig. 4.4, sorted per basis coefficient. Here, we show the measured means, standard deviations, and correlation times that are used in the model. On a grid of resolution N a total of $N(N+1)/2-1$ basis functions T_{lm}^N is available, which can be sorted in ascending order of l and m. Here only the first 2079 values are shown, corresponding to all resolvable modes for $N = 64$, a coarse resolution that will be investigated momentarily. A decreasing mean value and variance are observed as the scale size is decreased. This is seen until basis functions with $l = 23$ are considered, at the 275^{th} basis coefficient, after which the mean and variance remain roughly constant. This corresponds to the wavenumber at which the reference energy spectrum follows the l^{-1} decay. The measured correlation time τ_{lm} becomes smaller as larger wavenumbers are considered, which indicates that the smaller scales in the flow behave in an increasingly dynamic manner. The small values of τ_{lm} result in a relatively larger contribution from the model term to the dynamics of the smallest resolvable scales.

4.4.2 Coarse-grid flow simulations

In this subsection, the performance of the model is tested in coarse-grid numerical simulations of the flow. In particular, the model is applied at resolutions $N = 64$ and $N = 32$ to show that the forcing parameters are applicable at

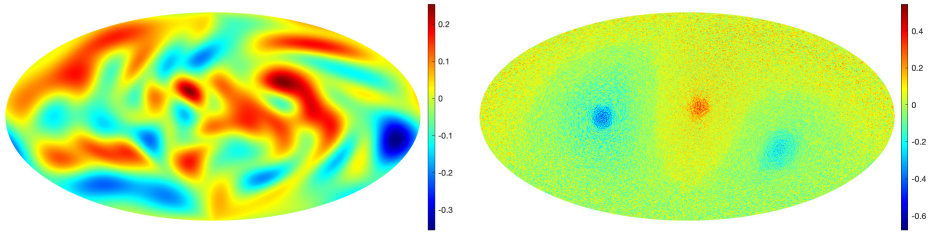


FIGURE 4.3: Left: Initial vorticity field used in the numerical simulations performed throughout the chapter. Right: Snapshot of the vorticity field after reaching a statistically steady state.

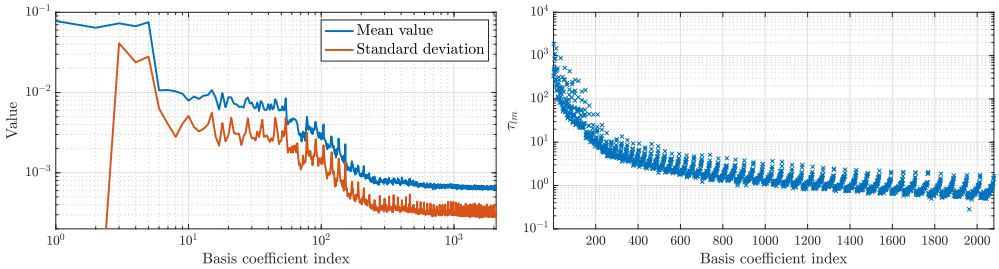


FIGURE 4.4: Left: measured means and standard deviations of the absolute value of each basis coefficient of the reference solution. Right: Estimated correlation time for each basis coefficient of the reference solution.

different coarse resolutions. The chosen levels of coarsening provide a significant reduction in computational costs. At the same time, the dominant flow patterns can be accurately resolved, as shown in Section 4.2. Four different settings for the model are studied by varying the minimal wavenumber at which the model is activated and by either enabling or disabling the stochastic model term. The scales at which the model is applied are $l \geq 1$ and $l \geq 8$ for resolutions $N = 64$ and $N = 32$, in order to capture the same flow complexity at different resolutions. The choice of $l \geq 1$ corresponds to applying the model at all available scales, whereas $l \geq 8$ only applies to small-scale flow features. For each resolution, we illustrate the need for modeling by providing snapshots of the filtered reference solution and the no-model coarse-grid solution. From these figures, the qualitative features of the solution at different resolutions become apparent.

We first consider the results at resolution $N = 64$. A qualitative comparison of the different numerical solutions is provided in Fig. 4.5. The top left panel shows a snapshot of the reference solution at the statistically steady

state, where the high-frequency components have been filtered from the solution. As before, the applied filter is a spectral cut-off filter where the cut-off wavenumber is defined by the coarse-grid resolution. Coherent large-scale vorticity structures that are resolvable on the coarse grid are visible in this snapshot. The merit of the forcing can be observed through the differences between the coarse numerical simulation results. The top middle panel shows the no-model coarse solution, which shows clear qualitative differences with the reference result. The top right panel and the bottom row show forced coarse numerical solutions at statistically steady states, using the different forcing settings. Evidently, the latter snapshots reveal a smoother vorticity field and a more accurate representation of the reference vorticity, compared to the coarse no-model simulation. In particular, a qualitative agreement in terms of large-scale vortex structures may be observed. In this specific case, a large connected positive vorticity structure (in red) and two smaller negative vorticity structures (in blue) are reproduced when applying the model. Interestingly, the proposed nudging concentrates some additional positive vorticity in the tail of the coherent structure (in red), whereas no such behavior is observed for the negative vorticity.

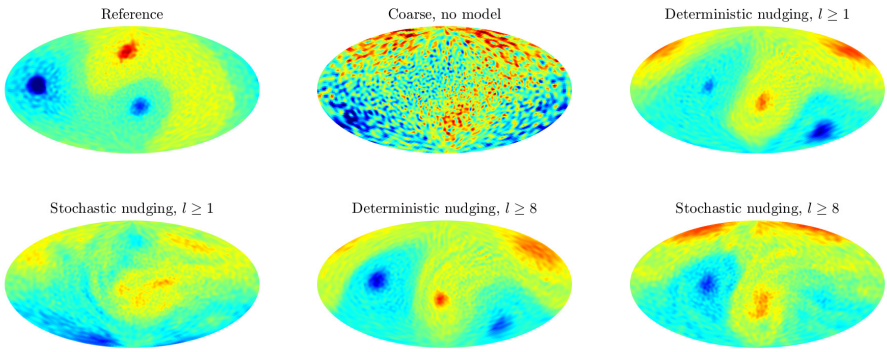


FIGURE 4.5: Snapshots of numerical solutions at a statistically steady state. Top left: filtered reference solution, displaying only modes resolvable for $N = 64$. Top middle: no-model coarse numerical solution. Top right and bottom row: coarse numerical solution with forcing applied, using the full model term or only the deterministic part, with varying minimal wavenumber at which the forcing is applied.

The qualitative differences are reflected in the energy spectra, visualized in Fig. 4.6, showing the energy spectra using the forcing for $l \geq 1$ and $l \geq 8$ in the two panels. By construction, nudging reduces the energy content in the small

scales of the flow. Accurate energy levels are observed for both the deterministic and stochastic nudging procedures. These results are observed for both choices of scales at which the forcing is applied. A good agreement in the energy at the large scales is observed for all performed simulations. Particularly, the energy spectra demonstrate a striking agreement at the smallest resolved scales when the model is applied. This suggests that the choice of parameters for the deterministic and stochastic forcing is well-suited for reproducing the energy spectra at these scales.

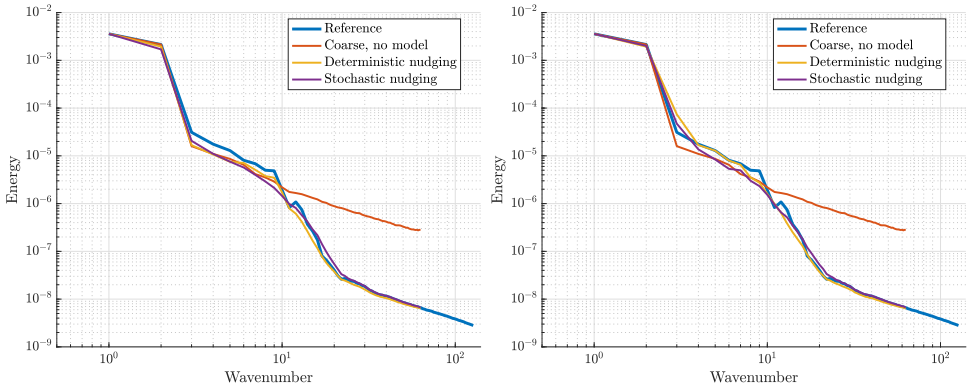


FIGURE 4.6: Average energy spectra for forced coarse solutions, using $N = 64$, compared to the energy spectra of the reference solution and the no-model coarse solution. The forcing is applied at wavenumbers $l \geq 1$ (left) and wavenumbers $l \geq 8$ (right).

A quantitative comparison of the statistics of the solutions is given in Figure 4.7. For each basis coefficient, the mean, standard deviation and estimated correlation time are shown. The mean and the standard deviations of the time series display similar qualitative behavior regardless of the minimal wavenumber at which the forcing is applied. For these quantities, both the deterministic nudging and the stochastic nudging lead to significant improvement compared to the no-model results. Including the stochastic component of the forcing, based on the high-resolution reference data, leads to an increased agreement at the smaller scales of the flow, indicating that the inclusion of additional variance in the forcing of the small scales leads to a truthful reproduction of these statistical quantities.

The estimated correlation times of the large-scale modes ($l \leq 8$) in Fig. 4.7 show that deterministic nudging of all modes yields an improved correlation time compared to the no-model case. However, the stochastic nudging procedure for $l \geq 1$ leads to smaller correlation times compared to the coarse no-model simulation, implying that the stochastic component of the forcing is

too strong. A qualitative improvement is observed when applying the model to wavenumbers $l \geq 8$, for both deterministic and stochastic nudging. These results suggest that the evolution of large scales in the flow benefits from an accurate statistical description of the evolution of small scales. This coincides with a basic premise underlying large-eddy simulation.

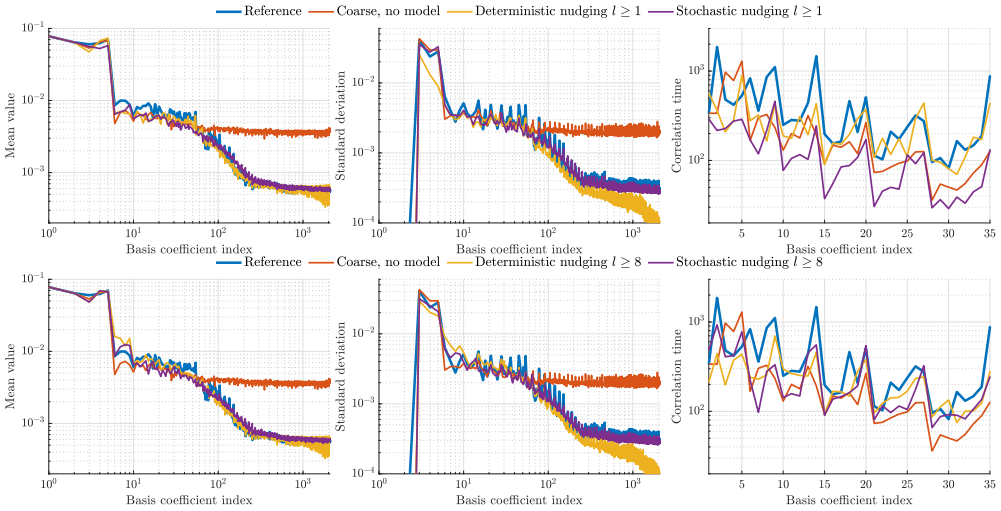


FIGURE 4.7: Statistics of the basis coefficient time series of the reference solution, no-model coarse solution, and coarse solution with the model applied for $N = 64$. Shown here are the results when applying the model at wavenumbers $l \geq 1$ (top row) and $l \geq 8$ (bottom row). The mean value (left) and standard deviation (middle) are shown for all wavenumbers. The correlation time (right) is shown for the large-scale components, with wavenumbers $l \leq 8$.

The numerical experiment is repeated at a resolution $N = 32$ to demonstrate that the proposed model yields forcing parameters that can be efficiently applied at different coarse resolutions. At this resolution, large spatial structures in the flow may still be resolved with acceptable accuracy. The forcing will be applied at wavenumbers $l \geq 1$ and $l \geq 8$, where the model affects all scales of motion in the former and only the small scales in the latter.

A qualitative comparison of the statistically steady states is given in Fig. 4.8. It may be seen that the model effectively produces a smooth vorticity field with qualitatively similar features as the reference solution. This is reflected by the decrease of energy in the smallest resolvable scales compared to the no-model formulation, as shown in the energy spectra in Fig. 4.9. As previously observed, all coarse-grid numerical simulations accurately capture

the energy in the largest scales of motion. Applying the model also leads to a notable agreement with the reference solution in the average energy levels of the smallest resolvable scales.

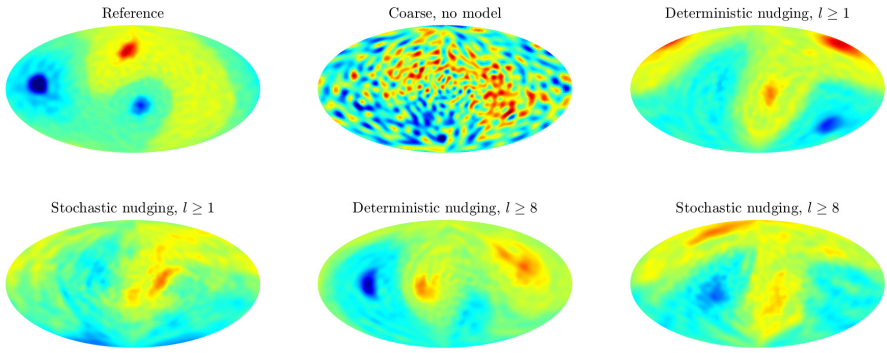


FIGURE 4.8: Snapshots of numerical solutions at a statistically steady state. Top left: filtered reference solution, displaying only modes resolvable for $N = 32$. Top middle: no-model coarse numerical solution. Top right and bottom row: coarse numerical solution with forcing applied, using the full model term or only the deterministic part, with varying minimal wavenumber at which the forcing is applied.

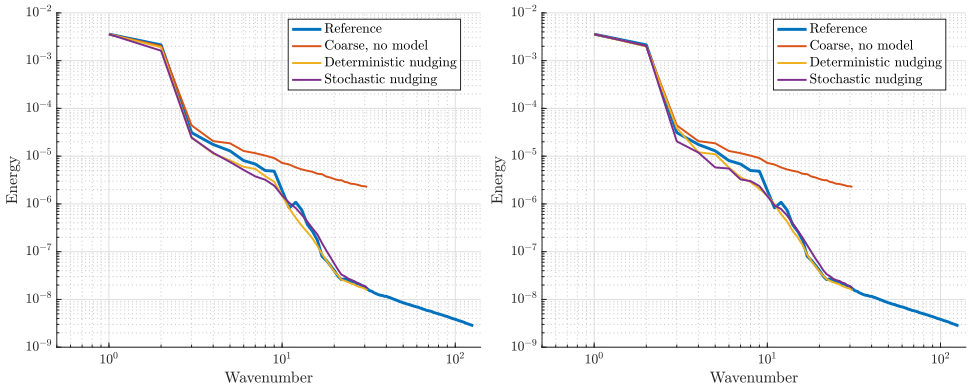


FIGURE 4.9: Average energy spectra for forced coarse solutions, using $N = 32$, compared to the energy spectra of the reference solution and the no-model coarse solution. The forcing is applied at wavenumbers $l \geq 1$ (left) and wavenumbers $l \geq 8$ (right).

A comparison of the mean value, standard deviation and estimated correlation times of the time series of the basis coefficients is given in Fig. 4.10. Applying the model leads to a clear improvement in the mean and variance of the coefficients, regardless of the choice of length scales at which the forcing is applied. Employing the deterministic forcing at all lengthscales yields a good agreement in the correlation times, whereas the stochastic forcing reduces the measured correlation times and yields no improvement. The correlation times are also found to improve when applying the deterministic model only to components with wavenumber $l \geq 8$. The stochastic forcing displays no significant improvement when applied at these wavenumbers.

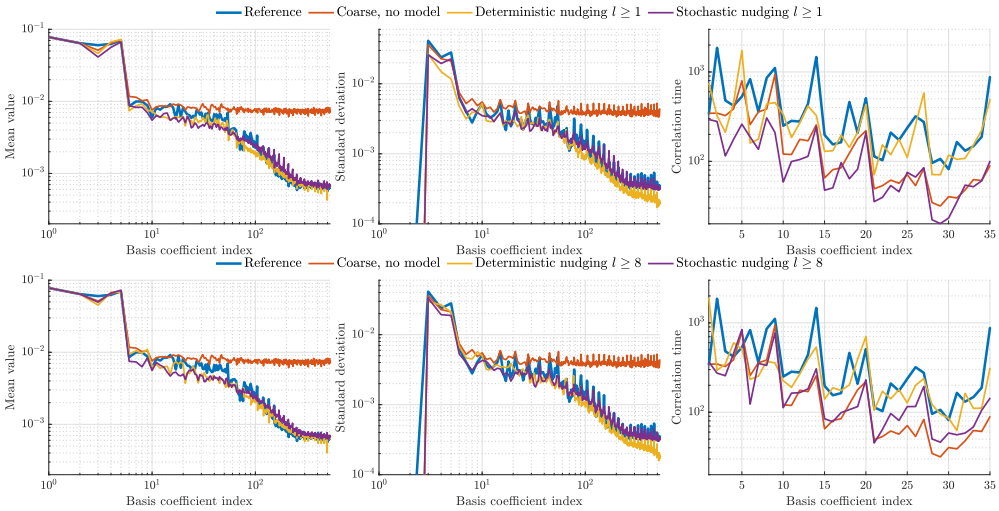


FIGURE 4.10: Statistics of the basis coefficient time series of the reference solution, no-model coarse solution, and coarse solution with the model applied for $N = 32$. Shown here are the results when applying the model at wavenumbers $l \geq 1$ (top row) and $l \geq 8$ (bottom row). The mean value (left) and standard deviation (middle) are shown for all wavenumbers. The correlation time (right) is shown for the large-scale components, with wavenumbers $l \leq 8$.

4.4.3 Large-scale vortex dynamics at statistically steady states

The qualitative predictions of coarse-grained modeled dynamics can be analyzed by means of the vortex trajectories over long integration times. Here, the vortex movement is tracked by locating the maximum and minimum attained vorticity value at each solution snapshot. According to [127], high-resolution numerical experiments indicate that the ratio between the angular momentum

and enstrophy governs the number of large-scale vortex structures in the final statistically steady state. This ratio is determined by the initial condition and remains constant throughout the no-model numerical simulations since the angular momentum and the enstrophy are conserved quantities in the discretized system. Additionally, the vortex trajectories are found to be stable. Thus, the long-term qualitative behavior of the coarse numerical solutions can be assessed by measuring the number of large-scale vortices and their trajectories. As we previously observed, the coarse-grained modeled vorticity fields show qualitative agreement in terms of the number of vortices. Here, we demonstrate the capability of the model to accurately yield stable long-time vortex dynamics by tracking vortex movement over long simulation times.
The long-time vortex trajectories for various numerical realizations are shown in Fig. 4.11. The reference trajectories are obtained from the high-resolution simulation as used in the previous section. The model results at resolutions $N = 32$ and $N = 64$ are obtained by applying the model to wavenumbers $l \geq 8$. The reference trajectories display stable movement along clearly defined trajectories about a fixed axis. Such behavior is not observed for the coarse no-model results, where instead the extreme values of the vorticity move in a seemingly unorganized fashion without distinct trajectories. Applying the model to either of the presented resolutions yields a noticeable qualitative improvement in the measured vortex movement. In particular, we identify trajectories about the same fixed axis as the reference trajectories but the model trajectories exhibit perturbations. The perturbations appear stronger when stochastic nudging is applied and when coarser grids are considered.

4.5 Concluding remarks

In this chapter, we have proposed and assessed a standalone data-driven model for the coarsening of the Euler equations on the sphere. High-resolution simulation snapshots were used as a reference. This data was decomposed into spherical harmonic modes and corresponding time series of coefficients were determined. A stochastic model was introduced to compensate for shortcomings introduced by severe coarsening. The model parameters were obtained from statistics of the spherical harmonic coefficients time series. In particular, the proposed model was designed to reproduce the kinetic energy spectrum of the reference data in statistically steady states by adopting a nudging strategy similar to continuous data assimilation.
The model is imposed using a prediction-correction scheme leading to a formulation similar to a steady-state Fourier domain Kalman filter. We opted for a separate nudging strength for each of the forced lengthscales, dependent on the corresponding measured characteristic timescale, and demonstrated that this

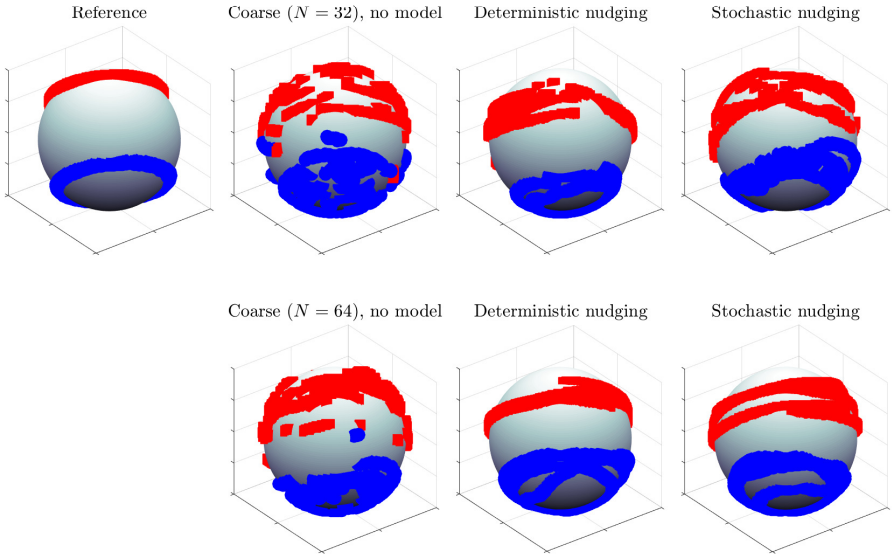


FIGURE 4.11: Trajectories of large-scale vortices for various numerical solutions. The red and blue lines denote the trajectories of the maximum and minimum vorticity values, respectively. Shown are the measured trajectories of the reference solution (top left). The coarse no-model numerical solution and the coarse solution with the model at resolution $N = 32$ are displayed on the top row. Both the deterministic and the stochastic models are applied to wavenumbers $l \geq 8$ at this resolution. The bottom row shows the realizations at resolution $N = 64$, where the model is applied to wavenumbers $l \geq 8$.

approach accurately recovers the energy levels in small resolved spatial scales and leads to stable long-time solutions. Moreover, no assumptions about the employed resolution are used in the derivation model. This was demonstrated by first measuring the forcing parameters and subsequently applying the model on several coarse computational grids. The proposed stochastic and deterministic models were not found to differ much in terms of results. Both approaches yielded accurate kinetic energy spectra at strong coarsening. In addition, the deterministic model yielded accurate correlation times of the magnitudes of the spherical harmonic basis coefficients, indicating accurate evolution of the large-scale flow features.

The results in this chapter show that the decomposition of a high-resolution reference signal into spatial global basis functions and temporal coefficients can be employed efficiently to obtain resolution-independent forcing parameters to

be used in models for coarse numerical simulations. The proposed model relies on several simplifying assumptions which will be scrutinized in future work. In particular, the robustness of the model in terms of stability and accuracy with respect to varying nudging strengths will be assessed. The connection to data assimilation algorithms and Kalman filtering theory may help to extend the model and weaken underlying assumptions, for example by including estimated covariance between different spherical harmonic modes in the model.
The approach presented here is general for flows in statistically stationary states and is not restricted to the two-dimensional Euler equations or the use of spherical harmonic modes as a global basis. Different flow settings may be considered by adopting, for example, a Fourier basis for periodic domains or, more generally, proper orthogonal decomposition (POD) modes when boundaries are present in the domain. Further work will be dedicated to extending the proposed model to different sophisticated flow settings such as two-dimensional Rayleigh-Bénard convection, the rotating Euler equations on the sphere, or the quasi-geostrophic equations on the sphere.

Chapter 5

Data-driven spectral turbulence modeling for Rayleigh-Bénard convection

5.1 Introduction

Turbulent flows are characterized by the distribution of kinetic energy over a vast range of scales. The nonlinearity in the Navier-Stokes equations ensures that large and small eddies interact with each other, resulting in a wide range of dynamic flow features [139]. This process transfers kinetic energy from the large energy-containing scales toward smaller scales, until the kinetic energy is finally dissipated by viscosity at the smallest scales. The energy cascade towards the smallest scales yields a significant challenge in computational fluid dynamics in the turbulent regime [74, 147]. In order to accurately simulate the flow, the fluid dynamical model should resolve the scales of turbulence down to the Kolmogorov length scale. A direct approach would then require very fine computational grids which is often intractable even with modern-day computing resources. A common way to alleviate the large computational requirements is by reducing the numerical resolution at which an approximate solution to the flow is obtained. To compensate for the lack of refinement of the computational approach, a model term is subsequently added to the governing equations to represent the influence of unresolved dynamics on the resolved components of the flow [70, 138, 146, 68].

In this chapter, we describe how prior knowledge of flow statistics obtained from an offline fully resolved simulation may be incorporated to construct an online high-fidelity model for coarse numerical simulations. The proposed reduced-order model acts on the numerical solution in spectral space, employing techniques from time series modeling and data assimilation. This model is designed to yield accurate kinetic energy spectra, despite the rather coarse flow

representation. We demonstrate the capabilities of this data-driven approach for two-dimensional Rayleigh-Bénard convection at high Rayleigh number.

Rayleigh-Bénard convection is a fundamental problem in fluid dynamics, describing buoyancy-driven flows heated from below and cooled from above [93, 97, 3, 100]. In particular, thermal convection is meaningful in geophysical processes, such as in describing convective processes in the atmosphere [81] or the ocean [118]. The large range of scales present in turbulence is also further affected by buoyancy effects. For example, a common phenomenon observed in Rayleigh-Bénard convection is the formation of spatially coherent structures in large-scale circulation [3] and, in larger spatial domains, the formation of thermal superstructures [153]. On the other hand, a thin boundary layer exists near either of the walls which becomes turbulent and increasingly thinner with growing temperature differences between the walls, i.e., growing Rayleigh number [98, 175]. Properly resolving the boundary layers requires large computational grids and poses a challenge even in two-dimensional Rayleigh-Bénard convection [175]. This stresses the conundrum of computational fluid dynamics, where one strives to find a balance between simulating flows at modest computational costs without adversely affecting the prediction of flow statistics.

Simulating flows at modest computational costs while retaining a high level of accuracy is the aim of large-eddy simulation (LES) [147, 74]. Instead of fully resolving all length scales of the flow, a computationally less intensive approximation is found by coarsening the flow description and simultaneously including a subgrid model to accommodate the loss of explicit finer details in the dynamics. The coarsening is accomplished by spatial filtering, which, through the specification of the filter-width, establishes the required level of refinement that should be included in the numerical simulations. The subgrid model then approximates the effect the unresolved dynamics has on the resolved scales, and serves as a closure for the filtered equations. This approximation depends on both the adopted filter [75] and selected closure model as well as the choice of discretization and level of coarsening [103, 11, 22].

With the increase of computational resources, direct numerical simulations (DNS) of turbulent flows are achievable to an ever-increasing extent and may serve to generate data from which LES models could be derived. This data-driven LES has been an active topic of research in recent years. For example, the decomposition of unresolved dynamics into fixed global basis functions and corresponding time series yields an efficient approach for which only the latter needs to be modeled. In [62] DNS data were employed of the barotropic vorticity equation to model the time series of spherical harmonics as stochastic processes with memory effects, leading to accurate kinetic energy spectra in coarse-grid simulations. Using proper orthogonal decomposition (POD), [57]

showed that applying corrections to coarse-grid numerical simulations may lead to significant error reduction. Machine-learning methods have also been successfully employed to find subgrid models [11], reporting improved results compared to traditional eddy-viscosity models. Examples include using artificial neural networks in two-dimensional decaying turbulence [119] and convolutional neural networks in three-dimensional homogeneous isotropic turbulence [101], yielding improved energy spectra and turbulent fluctuation distributions.

Incorporating data into numerical models to improve flow predictions is well-established in geophysical fluid dynamics, where data assimilation has been successfully employed for several decades. The aim is to improve forecasting by minimizing the differences between observed and modeled values while accounting for uncertainties [76, 48]. In particular, continuous data assimilation (CDA) aims to nudge the model solution toward an observed reference by means of a feedback control term acting as external forcing [8, 7]. This concept is also extended to linear stochastic differential equations, arising as the continuous-time limit of the 3DVAR data assimilation algorithm [17], which has been shown to successfully steer coarse-grained numerical solutions of the two-dimensional Navier-Stokes equations towards an observed reference solution. Additionally, the convergence of coarse numerical solutions augmented by CDA to an observed reference solution has been proven [60] and shown numerically for two-dimensional Rayleigh-Bénard convection [4].

The purpose of this chapter is to combine ideas from data assimilation with large-eddy simulation. In particular, we derive a model term based on statistical quantities from a reference high-resolution simulation and use this as a stand-alone model for coarse numerical simulations. Our proposed method incorporates Ornstein-Uhlenbeck processes in the evolution of the Fourier coefficients of the numerical solution, steering the solution towards an a priori measured statistically steady state. Only three parameters need to be defined for each Fourier mode, outlining the simplicity of the model. The parameters are inferred from data, do not depend on the adopted spatial or temporal discretization, and are defined such that the reference energy spectrum is closely reproduced in the coarse-grid simulations. The resulting prediction-correction scheme is of the form of the diagonal Fourier domain Kalman filter [80, 115] with a fixed prescribed gain. This identification enables future research that combines LES and data assimilation. The same approach has been applied in a recent study of coarse-grid modeling of the two-dimensional Euler equations on the sphere [55], where a decomposition of the vorticity field into spherical harmonic basis modes was employed in the coarse-grid model.

The chapter is structured as follows. The governing equations and adopted discretization are described in Section 5.2. The data-driven model is introduced in Section 5.3, detailing the forcing in Section 5.3.1 and complementary heat

flux correction in Section 5.3.2. The model performance for a variety of model configurations is assessed in Section 5.4. Conclusions are presented in Section 5.5.

5.2 Governing equations and numerical methods

In this section, we introduce the governing equations and the simulated case in Section 5.2.1, the employed numerical discretization in Section 5.2.2, and illustrate the effects of severe coarsening, without any modeling correction, on the quality of the solution in Section 5.2.3.

5.2.1 Governing equations

Rayleigh-Bénard (RB) convection is described by the incompressible Navier-Stokes equations coupled to buoyancy effects under the Boussinesq approximation. In non-dimensional form, the equations read

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} = \sqrt{\frac{Pr}{Ra}} \nabla^2 \mathbf{u} - \nabla p + T \mathbf{e}_y, \tag{5.1}$$

$$\nabla \cdot \mathbf{u} = 0, \tag{5.2}$$

$$\frac{\partial T}{\partial t} + \mathbf{u} \cdot \nabla T = \frac{1}{\sqrt{PrRa}} \nabla^2 T. \tag{5.3}$$

We restrict to two spatial dimensions in this work. Here, $\mathbf{u}$ denotes velocity, p the pressure and T the temperature. We denote by u and v the horizontal and vertical velocity components, respectively. Buoyancy effects are included in the momentum equation by means of the term $T\mathbf{e}_y$, where $\mathbf{e}_y$ denotes the vertical unit vector. The dimensionless parameters that determine the flow are the Rayleigh number $Ra = g\beta\Delta L_y^3/(\nu\kappa)$ and the Prandtl number $Pr = \nu/\kappa$. The Rayleigh number describes the ratio between buoyancy effects and viscous effects and is set to 10^{10} in order to set the focus on the challenging high-Ra convection regime. The Prandtl number determines the ratio of characteristic length scales of the velocity and the temperature and is set to 1. The gravitational acceleration is denoted by g, the thermal expansion coefficient by β, the temperature difference between the boundaries of the domain by Δ, the kinematic viscosity by ν, the thermal diffusivity by κ. The computational domain is rectangular with horizontal length L_x and vertical length L_y, which are chosen as 2 and 1, respectively. The domain is periodic for all variables in the horizontal direction and wall-bounded in the vertical direction. No-slip boundary conditions are imposed for the velocity at the walls. The non-dimensional temperature is prescribed at 1 at the bottom wall and 0 at the top wall.

Two-dimensional RB convection is fundamentally different from three-dimensional RB convection. The main advantage of restricting the flow to two spatial dimensions is that this enables direct numerical simulation (DNS) at a very large Rayleigh number [175]. Additionally, the large-scale circulation that appears in three-dimensional RB convection displays quasi-two-dimensional behavior and shows strong similarities with the large-scale circulation in two-dimensional RB convection [161].

5.2.2 Numerical methods

The adopted spatial discretization is an energy-conserving finite difference method [165] and is parallelized as in [34]. A staggered grid arrangement is used for the velocity, the pressure is defined at the cell centers, and the temperature is defined on the same grid as the vertical velocity. The latter choice ensures that no interpolation errors occur when computing the buoyancy term in Eq. (5.1) [160]. A uniform grid spacing is used along the horizontal direction whereas a hyperbolic tangent grid profile is adopted along the vertical direction. The non-uniform grid realizes refinement near the walls to resolve the boundary layer.

Time integration is performed using the fractional-step third-order Runge-Kutta (RK3) scheme for explicit terms combined with the Crank-Nicholson (CN) scheme for implicit terms, as presented in [160]. Every time step, from t^n to t^{n+1}, consists of three sub-stages, of which the steps are outlined below. The superscript j, $j = 0, 1, 2$, denotes the sub-stage, where $j = 0$ coincides with the situation at t^n. In each stage, a provisional velocity $\mathbf{u}^*$ is computed according to

$$\frac{\mathbf{u}^* - \mathbf{u}^j}{\Delta t} = \left[\gamma_j H^j + \rho_j H^{j-1} - \alpha_j \mathcal{G} p^j + \alpha_j \mathcal{A}_y^j \frac{(\mathbf{u}^* + \mathbf{u}^j)}{2} \right]. \tag{5.4}$$

The parameters γ, ρ, and α are the Runge-Kutta coefficients, given by $\gamma = [8/15, 5/12, 3/4], \rho = [0, -17/60, -5/12]$, and $\alpha = \gamma + \rho$ [34, 160, 141]. Moreover, H^j is comprised of the discrete convective terms, the discrete horizontal diffusion terms, and the source terms. Here, the only source term is the buoyancy term appearing in the evolution of the vertical velocity. The discrete gradient operator is denoted by $\mathcal{G}$. The discrete vertical diffusion, given by $\mathcal{A}_y$, is treated implicitly. The implicit treatment eliminates the severe viscous stability restriction caused by the use of a nonuniform mesh near the boundary [94]. Subsequently, the Poisson equation (5.5) is solved for the pressure to

impose the continuity constraint (5.2),

$$\nabla^2 \phi = \frac{1}{\alpha_j \Delta t} \left(\nabla \cdot \mathbf{u}^* \right). \tag{5.5}$$

Discretely, this equation takes the form

$$\mathcal{L}\phi = \frac{1}{\alpha_j \Delta t} \left(\mathcal{D}\mathbf{u}^* \right). \tag{5.6}$$

Here, $\mathcal{L}$ is the discrete Laplace operator $\nabla \cdot \nabla$ and $\mathcal{D}$ is the discrete divergence operator $\nabla \cdot$. The velocity and pressure are subsequently updated according to

$$\mathbf{u}^{j+1} = \mathbf{u}^* - \alpha_j \Delta t \left(\mathcal{G}\phi \right) \tag{5.7}$$

and

$$p^{j+1} = p^j + \phi - \frac{\alpha_j \Delta t}{2Re} \left(\mathcal{L}\phi \right), \tag{5.8}$$

after which the velocity $\mathbf{u}^{j+1}$ is divergence-free. Time integration of the energy equation (5.3) follows similarly. The newly obtained velocity is used to generate the energy field T^{j+1} analogously to Eq. (5.4).

The convective terms are discretized using the QUICK interpolation scheme [106]. The diffusive terms are discretized using a standard second-order finite difference method, for both spatial directions. Similarly, the discrete gradient $\mathcal{G}$, the discrete divergence $\mathcal{D}$, and the discrete Laplacian $\mathcal{L}$ are defined using finite differences.

5.2.3 Altered dynamics under coarsening

The DNS is carried out on a 4096×2048 grid, which has been shown to be a sufficiently high resolution for the chosen Rayleigh number [175]. The coarse-grid numerical simulations will be performed on a 64×32 grid. This coarsening introduces significant discretization errors and does not allow for an accurate resolution of the smaller coherent structures present in the flow. The truncation error of the numerical method on this coarse grid introduces artificial dissipation and additional high-pass smoothing native to the discretization method [73]. Fig. 5.1 shows a snapshot of the DNS and the coarse-grid simulation, both in statistically steady states, from which the huge implications of such significant coarsening become apparent.

The temperature at the mentioned resolutions is shown in the top row of Fig. 5.1. The coarsened temperature displays only qualitative large-scale agreement with the DNS temperature, both yielding similar plumes of temperature and similar large-scale circulation. Obviously, the persistent small-scale

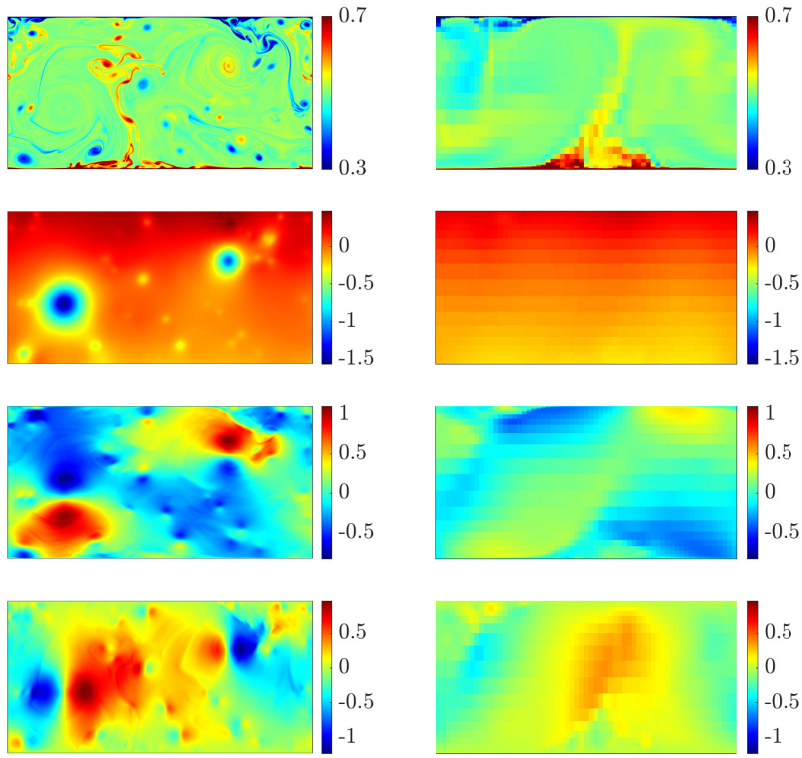


FIGURE 5.1: Snapshots of the DNS (4096×2048 grid, left) and coarse no-model simulation (64×32 grid, right) in statistically steady states. From top to bottom, we show temperature, pressure, horizontal velocity, and vertical velocity.

coherent structures visible in the DNS snapshot are lost on the coarse grid. This loss may also be observed for the pressure, depicted in the second row of Fig. 5.1. The high-resolution and low-resolution fields exhibit clear qualitative differences, especially considering the absence of distinct low-pressure regions in the coarse-grid result. A better qualitative agreement between the results at the different resolutions is observed for the velocity components, shown in the bottom rows of Fig. 5.1. At low resolution, qualitatively the same flow patterns can be observed as appear in the high-resolution results, albeit with decreased magnitude.

The discrepancies between the velocity and temperature fields at the different resolutions clearly pose the challenge we address in this chapter. In the next section, we therefore specify our new forcing approach which aims to rectify the observed differences largely. The extent to which this new approach is successful will be scrutinized in Section 5.4.

5.3 Spectrum-preserving forcing

In this section, we describe a data-driven forcing to augment coarsened numerical simulations of statistically steady states. The high-resolution and low-resolution snapshots presented in the previous section hinted at the need of introducing explicit forcing to more accurately represent average flow patterns on coarse computational grids. Here, we propose a model to reproduce the kinetic energy spectra in coarse numerical simulations.

The model parameters are extracted from the DNS performed at a 4096×2048 computational grid, from a sequence of 8040 solution snapshots each separated by 0.05 time units. With these numerical data we achieve sufficiently many snapshots to reliably recover statistical properties of the flow, which is a prerequisite for our model development. The Fourier components of horizontal cross-sections of the solution are computed for each snapshot, yielding time series for each streamwise wavenumber at each y-coordinate. The magnitudes of these complex time series yield mean values, variances, and correlation times that are used as model parameters. We next present the main steps in the construction of this model.

5.3.1 Reconstructing the energy spectra

The momentum equation (5.1) and temperature equation (5.3) can be written as a system of complex ODEs for the mode coefficients through projection onto a Fourier basis. In what follows, we only describe a spectrum-preserving forcing for the momentum. The same derivation is used to define a forcing for the temperature. We note that a spectral decomposition of the velocity and temperature fields is not straightforward due to the presence of boundaries along one spatial dimension. Therefore, we restrict ourselves to one-dimensional periodic horizontal profiles to ensure that the Fourier series is well-defined. This choice enables the model to explicitly identify wall-induced features of the flow. Alternatively, after taking into account the nonuniform grid spacing in the wall-normal direction the velocity field can be decomposed by applying a sine transform in the wall-normal direction or by periodically extending the domain and applying a Fourier transform, but these approaches are not pursued in this study. For a fixed vertical coordinate y_l, the profile $\mathbf{u}(x, y_l, t)$ is

decomposed into Fourier modes and corresponding complex coefficients $c_{k,l}$, where k denotes the wavenumber. The coefficients satisfy the system of ODEs

$$\frac{\mathrm{d}c_{k,l}}{\mathrm{d}t} = L\left(\mathbf{c}, k, l\right), \quad k = 0, \ldots, N_x/2, \tag{5.9}$$

where L involves the spectral representation of of $\mathbf{u} \cdot \nabla$, ∇, ∇^2 and the source term in Eq. (5.1). We have already assumed a finite truncation of the number of Fourier modes in this formulation. The vector $\mathbf{c}$ is comprised of all Fourier coefficients up to the largest resolvable wavenumber.

To arrive at a spectrum-preserving forcing, it is sufficient to consider only the magnitude of the Fourier coefficients $|c_{k,l}|$. These evolve according to a system of ODEs

$$\frac{\mathrm{d}|c_{k,l}|}{\mathrm{d}t} = L_r(\mathbf{c}, k, l), \quad k = 0, \ldots, N_x/2, \tag{5.10}$$

where L_r is introduced to simplify notation. In the model implementation, the definition and explicit computation of L and L_r is not required. Regarding $|c_{k,l}|$ as a stationary stochastic process, we observe that $\mathbb{E}\left(|c_{k,l}|^2\right)$ describes the mean energy content of the k^{th} Fourier mode at height y_l. By the definition of variance, we find that

$$\mathbb{E}\left(|c_{k,l}|^2\right) = \mathrm{var}(|c_{k,l}|) + \mathbb{E}\left(|c_{k,l}|\right)^2. \tag{5.11}$$

Thus, it is sufficient to obtain an accurate mean value and variance of $|c_{k,l}|$ to achieve the desired energy content in the k^{th} Fourier mode. We aim to reproduce the mean value and variance of $|c_{k,l}|$ by augmenting Eq. (5.10) with an Ornstein-Uhlenbeck (OU) process,

$$\mathrm{d}|c_{k,l}| = L_r(\mathbf{c}, k, l)\mathrm{d}t + \frac{1}{\tau_{k,l}}\left(\mu_{k,l} - |c_{k,l}|\right)\mathrm{d}t + \sigma_{k,l}\mathrm{d}W^t_{k,l}, \tag{5.12}$$

where $\mu_{k,l}$, $\tau_{k,l}$ and $\sigma_{k,l}$ are statistical parameters inferred from a sequence of snapshots of the reference solution. The desired mean value of $|c_{k,l}|$ is given by $\mu_{k,l}$, the term $\sigma_{k,l}\mathrm{d}W^t_{k,l}$ serves to match the measured variance. Here, $\mathrm{d}W^t_{k,l}$ is a general stochastic process which can be tailored to the data [56], although the common choice is to let it be a Gaussian process with a variance depending on the time step size [84]. We adopt the latter and include the variance scaling in the definition of $\sigma_{k,l}$. The forcing strength is determined by a timescale $\tau_{k,l}$ and can be specified per k, l separately. Detailed specifications of the adopted values of $\mu_{k,l}, \sigma_{k,l}$, and $\tau_{k,l}$ follow shortly. The combination of the original dynamics and the feedback control, including the stochastic term, arises as

the continuous-time limit of the 3DVAR data assimilation algorithm [17]. The model assumes that the unresolved dynamics can be accurately represented by independent stochastic processes. Interactions in the vertical direction are included via the fully resolved simulations which are basis to the forcing. We will demonstrate that this model is capable of producing accurate results.

The model will be applied as a prediction-correction scheme. In the case of the RK3 scheme adopted in this work, the following steps are performed for each RK sub-stage. First, the provisional velocity $\mathbf{u}^*$ is computed as in Eq. (5.4). The Fourier coefficients $\tilde{c}_{k,l}$ are computed for this provisional velocity, after which the model is applied in the form of a correction

$$|c_{k,l}| = |\tilde{c}_{k,l}| + \frac{\alpha_j \Delta t}{\tau_{k,l}} \left(\mu_{k,l} - |\tilde{c}_{k,l}|\right) + \sigma_{k,l} \Delta W_{k,l}, \tag{5.13}$$

where $\Delta W_{k,l}$ are samples from a standard normal distribution, independently drawn for each k and l. These are determined for each time step and kept constant throughout the sub-stages comprising the time step. The correction only affects the magnitudes of the Fourier coefficients, the phases of $c_{k,l}$ are the same as those of $\tilde{c}_{k,l}$. Velocity fields are subsequently obtained by applying the inverse Fourier transform to the corrected coefficients $c_{k,l}$. After this procedure, the Poisson equation (5.6) is solved using the newly obtained velocity fields and the remaining steps of the sub-stage are completed. Applying the model before solving the Poisson equation ensures that the flow is incompressible at the end of each RK sub-stage. The entire algorithm, with the exception of solving the Poisson equation, may also be applied to the temperature equation. This prediction-correction algorithm has the additional benefit that the model can be easily implemented into already existing computational methods.

The correction (5.13) will be referred to as *nudging*. We distinguish between stochastic nudging, which is described by Eq. (5.13), and deterministic nudging, for which the stochastic term in Eq. (5.13) is omitted. We define a mean $\mu_{k,l,\text{stoch}}$ and $\mu_{k,l,\text{det}}$ for these methods, respectively, and specify these below.

The mean $\mu_{k,l}$ is specified such that the desired energy content is reproduced for small values of $\tau_{k,l}$. The magnitude of the coefficients is fully determined by the model in the limit of small $\tau_{k,l}$ and, as a result, Eq. (5.11) can be used to derive the mean $\mu_{k,l}$. To attain the desired energy contents when using stochastic nudging, we require that $\mu_{k,l,\text{stoch}} = \mathbb{E}\left(|c_{k,l}|\right)$. In the case of deterministic nudging with small $\tau_{k,l}$, the magnitudes of the coefficients remain constant at the value of $\mu_{k,l,\text{det}}$. Thus, in this situation the variance of $|c_{k,l}|$ is set to zero and we require that $\mu_{k,l,\text{det}} = \sqrt{\mathbb{E}\left(|c_{k,l}|^2\right)}$.

Treating Eq. (5.13) as a first-order autoregressive (AR(1)) process allows

us to specify the noise magnitude $\sigma_{k,l}$. For small values of τ_{lm}, Eq. (5.13) is well-approximated by an AR(1) process, indicating that the model should work well at small scales. We observe that the drift coefficient is $(1-\alpha_j\Delta t/\tau_{k,l})$ and assume that the sample variance $s^2_{k,l}$ is known from the high-fidelity data for every k,l. The noise magnitude follows by matching the variance of the AR(1) process with the sample variance, leading to the expression

$$\sigma_{k,l} = s_{k,l}\sqrt{1-\left(1-\frac{\alpha_j\Delta t}{\tau_{k,l}}\right)^2}. \tag{5.14}$$

In this chapter, the time scale $\tau_{k,l}$ will be defined as the autocorrelation of the time series of $|c_{k,l}|$, as measured from the high-resolution data. We note that, with the adopted definitions of $\mu_{k,l}$ and $\sigma_{k,l}$, $\tau_{k,l}$ can take on a range of values whilst still yielding accurate energy spectra. Robustness of the model under variations of $\tau_{k,l}$ will be the subject of future work. In fact, $\tau_{k,l}$ can take on any positive value larger than or equal to $\alpha_j\Delta t$. Small lengthscales are expected to yield a small value of $\tau_{k,l}$, resulting in an increased weight towards the model term and an increased noise magnitude for the stochastic term in the nudging. The model term will have a decreased weight at scales for which a large $\tau_{k,l}$ is measured, which is often observed for large spatial scales. These would correspondingly follow the deterministic resolved dynamics more closely.

The proposed prediction-correction method is of the same form as Fourier domain Kalman filtering [115, 80]. The approach can be understood as a steady-state filter with a prescribed gain $\alpha_j\Delta t/\tau_{k,l}$, for each k,l separately. By defining a prediction and an observation, the approach can be placed in the context of data assimilation. At each sub-stage of the RK3 scheme, the prediction is obtained by evolving the velocity fields according to the coarse-grid discretization. The 'observation' then consists of velocity or temperature fields sampled from the reference statistically stationary state. For stochastic nudging, these are velocity or temperature fields where the magnitudes of the Fourier coefficients are drawn from normal distributions with mean $\mu_{k,l,\text{stoch}}$ and variance $\sigma^2_{k,l}$. In the deterministic case, the observation consists of these fields with Fourier coefficients of prescribed magnitudes $\mu_{k,l,\text{det}}$.

5.3.2 Heat transport correction

The heat transport in the turbulent flow is described by the Nusselt number and is considered the key response of the system to the imposed Rayleigh number [3]. The definition of the Nusselt number that we adopt here is

$$Nu = 1 + \sqrt{PrRa}\langle vT\rangle_\Omega, \tag{5.15}$$

which is well-suited for use on coarse computational grids. An alternative definition of Nu involves a gradient of temperature, which is more sensitive to coarse-graining. In Eq. (5.15) Ω denotes the domain with area $|\Omega|$ and $\langle\cdot\rangle_\Omega$ denotes the domain average. It is clear from definition (5.15) that vT needs to be modeled accurately to recover skillful predictions of the heat flux. To achieve this, we propose a constraint to be used in conjunction with the model described in Section 5.3.1.

The volume average in (5.15) is comprised of averages of the heat flux along horizontal cross-sections of the domain. For a fixed vertical coordinate y_l, we denote the heat flux along this cross-section by $\langle vT\rangle_l$. Along this cross-section, we indicate the Fourier coefficients of the velocity and temperature with a hat symbol $\hat{\cdot}$ and observe that

$$(\widehat{vT})_0 = \sum_k \widehat{v}_k^* \widehat{T}_k, \tag{5.16}$$

where the subscript k signifies the k^{th} Fourier coefficient. The subscript 0 indicates that we consider the zeroth mode of the Fourier series, which by definition equals the value of $\langle vT\rangle_l$. We assume that a mean heat flux along the horizontal cross-section is known from the reference high-resolution data and denote this value by F_l. Subsequently, the heat flux along the horizontal cross-section in a coarse numerical simulation is corrected by minimizing the error $\|F_l - \sum_k \widehat{v}_k^* \widehat{T}_k\|^2$ with respect to T. Here, we minimize the error by varying the phases of the Fourier coefficients $\widehat{T}_k$. We alter the temperature instead of the vertical velocity so that the velocity field remains divergence-free. Adapting the phases only ensures that the spectrum of the temperature along the horizontal cross-section is invariant under the heat transport correction. In total, the heat flux correction is an extension of the nudging procedure (5.13). It enables a correction of the temperature field solely based on a statistic of the reference solution, rather than on a dynamic equation. In doing so, the dependence between the vertical velocity and the temperature is taken into account in the nudging procedure. Thus, applying the heat flux correction ensures an improved average Nusselt number estimate and is therefore expected to improve the accuracy of the numerical solutions.

The error $\|F_l - \sum_k \hat{v}_k^* \hat{T}_k\|^2$ is minimized using a gradient descent algorithm. We note that the correction may in principle yield an arbitrarily good approximation of the reference heat flux, but this is not guaranteed to produce physically relevant results. Instead of aiming for an exact agreement of the mean heat flux, we apply the gradient descent algorithm until the heat flux in the horizontal cross-section is within a 10% margin of the reference value. This serves to demonstrate the added value of the correction. Preliminary tests

have shown that this already improves the heat flux significantly without qualitatively altering the temperature field. In the next section, coarse numerical simulations are performed both with and without the heat flux correction. The optimization of this procedure is beyond the scope of this chapter.

5.4 Model performance

In this section, we apply the model in eight different configurations to numerical simulations on the coarse grid. The configurations are listed in Table 5.1 and differ in the variable that is being forced, the wavelengths at which the forcing is applied, and whether the forcing is deterministic or stochastic. These configurations will be referred to as M0-7, inspired by the nomenclature used in the comparison of LES models in [166]. Here, the wavenumbers at which the forcing is applied are chosen as $l \leq 5$ and $l \leq 32$. The former implies that the model only explicitly acts on the large scales of motion and the latter implies that all resolved scales are directly affected by the model. This set of configurations is chosen to distinguish between the effects of large-scale forcing and small-scale forcing, deterministic forcing and stochastic forcing, and the choice of the forced variable. The model simulations are run with a time step size $\Delta t = 0.02$. The minimal obtained value of $\tau_{k,l}$ is found to be larger than $\alpha_j \Delta t$ as described in Section 5.3.1. Therefore, the dynamics in cases M1-7 will always be a combination of the discretized dynamics and the model and will, at any wavenumber, not depend solely on the model.

Table 5.1: Model configurations used in the coarse numerical simulations.

	Model	Forced variable	Wavelengths	Curve
	Filtered DNS			solid
M0	No model			dashed
M1	deterministic	u	$k \leq 5$	dash-dotted
M2	deterministic	u	$k \leq 32$	dotted
M3	stochastic	u	$k \leq 5$	$*$
M4	stochastic	u	$k \leq 32$	$+$
M5	deterministic	T	$k \leq 32$	$\times$
M6	deterministic	u, T	$k \leq 32$	$\square$
M7	deterministic, heat flux correction	u, T	$k \leq 32$	$\Diamond$

We first provide in Section 5.4.1 an impression of the qualitative improvements obtained when applying the model. In the ensuing subsections, a detailed quantitative comparison is carried out. Several quantities will be compared with the filtered DNS data to gain insight into the quality of the model. In Section 5.4.2, we first verify that the model approximates the average energy spectra of the filtered DNS by comparing the spectra of the velocity and the temperature near the wall and in the core of the domain. In Section 5.4.3, the mean temperature, the mean heat flux, and the root-mean-square deviation (rms) are measured as a function of wall-normal distance and compared to the reference. Finally, global flow statistics such as the total kinetic energy and the Nusselt number are examined in Section 5.4.4.

The rms, mean temperature and mean heat flux rely on averages along horizontal cross-sections of the domain. For a fixed value of y, we adopt the following definition

$$\mathrm{rms}(f, y, t) = \left[\frac{1}{|A|}\int_A \left(f(x, y, t) - \langle f(x, y, t)\rangle_A\right)^2 \mathrm{d}A\right]^{1/2}, \tag{5.17}$$

where $\langle\cdot\rangle_A$ denotes the average over the horizontal cross-section with length $|A|$ and f is the field of interest. The mean temperature and heat flux are computed as the mean $\langle\cdot\rangle_A$ of the corresponding fields. The global kinetic energy will be computed as

$$\mathrm{KE} = \int_\Omega \frac{1}{2}\left(u^2 + v^2\right)\,\mathrm{d}\Omega \tag{5.18}$$

and the Nusselt number follows from definition (5.15).

Our interest lies in the time average of the aforementioned quantities. The quality of coarse-grid models is therefore measured by comparing averaged quantities rather than instantaneous quantities [103, 166]. The energy spectra, rms values and mean temperature, and heat flux will be measured after the coarse-grid numerical simulations have reached a statistically steady state. The global quantities of interest are illustrated using a rolling average over time.

5.4.1 Qualitative model performance

A qualitative comparison of the model configurations M0-7 is given in Fig. 5.2 to Fig. 5.5. In these figures, the snapshots of the DNS and of M0 are the same as depicted in Fig. 5.1. A comparison of the temperature fields in statistically steady states is provided in Fig. 5.2. Here, we observe that the configurations M1-4 do not lead to significant qualitative changes in the temperature field when compared to the no-model configuration M0. In these configurations,

the temperature is not explicitly forced and suffers from artificial dissipation inherent to the coarsening. The model configurations M5-7, in which the temperature is forced directly, display more pronounced small-scale features. At the same time, the large-scale circulation pattern is still visible in these results. In addition, from the results of M6 and M7 we conclude that applying the heat flux correction does not lead to qualitatively different temperature fields.

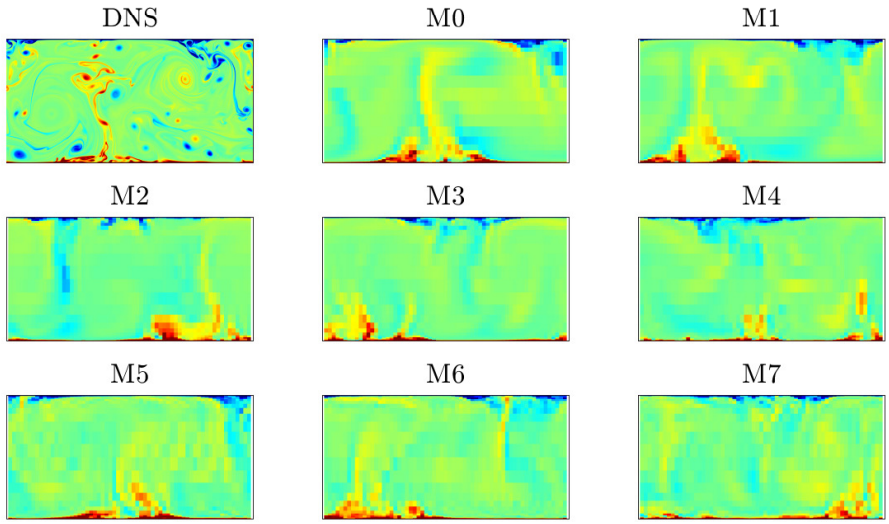


FIGURE 5.2: Temperature fields in statistically steady states. Shown are the reference solution and the results obtained with coarse numerical simulations M0-7. The color scheme is the same as used for the temperature fields shown in Fig. 5.1.

The pressure fields of the corresponding solutions are shown in Fig. 5.3. Here, we recall that any detail obsereved in the DNS pressure field is lost in the coarse no-model result M0. No improvements are observed in the pressure field when only the temperature is explicitly forced, as is done in M5. The remaining model configurations all yield a distinct qualitative improvement in the pressure fields. In particular, only applying a large-scale velocity correction already qualitatively changes the pressure field. This is observed for the deterministic and the stochastic forcing, given by M1 and M3, respectively. The addition of forcing the velocity at small scales or simultaneously forcing the temperature does not yield additional significant changes. A noticeable difference exists between the deterministic and stochastic methods. As becomes clear from M3-4, the random forcing leads to a fragmentation of the coherent structures in the pressure field.

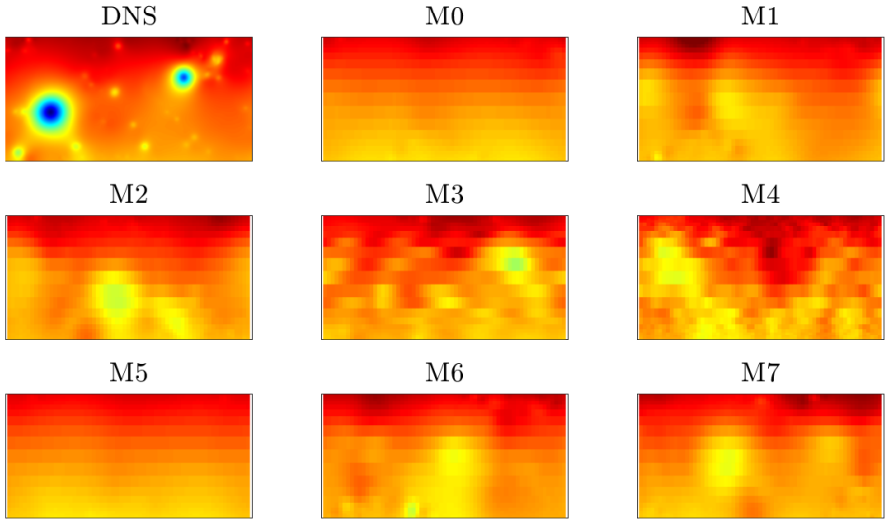


FIGURE 5.3: Pressure fields in statistically steady states. Shown are the reference solution and the results obtained with coarse numerical simulations M0-7. The color scheme is the same as used for the pressure fields shown in Fig. 5.1.

The horizontal velocity fields and vertical velocity fields are provided in Fig. 5.4 and Fig. 5.5, respectively. We observe that all coarse numerical solutions display agreement with the DNS in terms of large-scale coherent structures. Nonetheless, artificial dissipation leads to an underestimate of the velocity magnitude in cases M0 and M5. This suggests that only forcing the temperature is not sufficient for accurately reproducing the velocity fields. The other cases indicate that explicitly forcing the velocity leads to accurate velocity magnitudes.

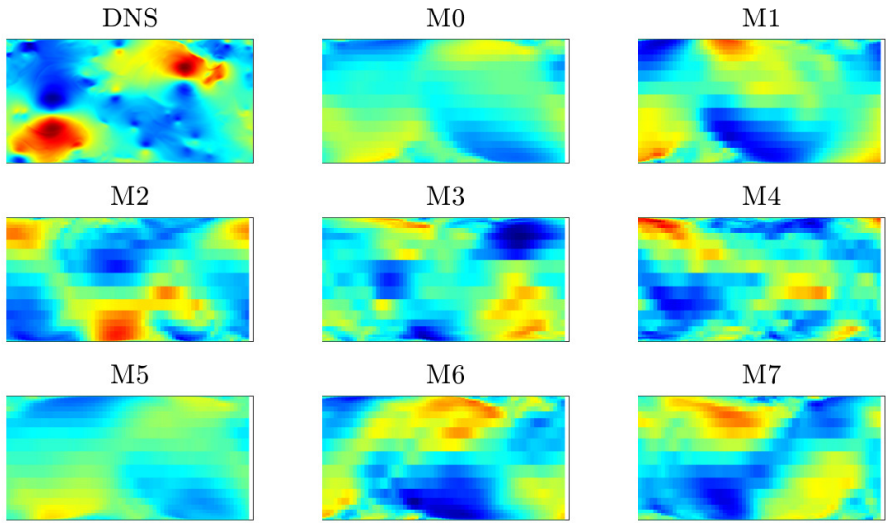


FIGURE 5.4: Horizontal velocity fields in statistically steady states. Shown are the reference solution and the results obtained with coarse numerical simulations M0-7. The color scheme is the same as used for the horizontal velocity fields shown in Fig. 5.1.

5.4.2 Energy spectra

We now establish that the model proposed in Section 5.3.1 improves the average energy spectra of the forced variables. The average energy spectra of the velocity components and the temperature are shown in Fig. 5.6, displaying the spectra along a horizontal cross-section near the wall and in the center of the domain. Both near the wall and in the center of the domain, respectively shown in the top and bottom row, the no-model M0 results exhibit significant differences compared to the filtered DNS. The measured energy levels of the velocity are too low with M0 at all resolved scales. In contrast, a significant discrepancy in the temperature spectra is observed only for wavenumbers larger than 10.

The discrepancies in the spectra of M0 and the reference are attributed to artificial dissipation caused by the coarsening. In particular, the numerical dissipation affects both the velocity and the temperature spectra at higher wavenumbers. Through the nonlinear interactions in the momentum equation the velocity is adversely affected at all wavenumbers. This is further corroborated by the results of M2 and M4, where all available lengthscales are forced only for the velocity. In the core of the domain, where the coarsening

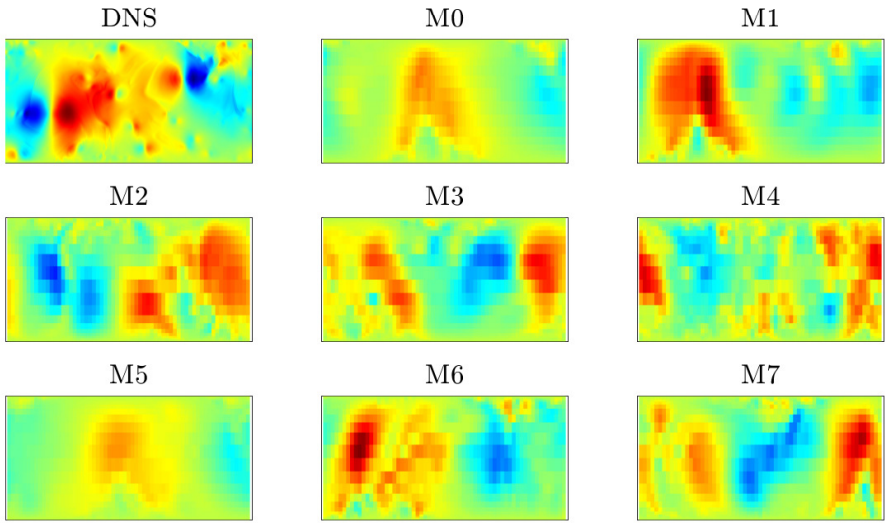


FIGURE 5.5: Vertical velocity fields in statistically steady states. Shown are the reference solution and the results obtained with coarse numerical simulations M0-7. The color scheme is the same as used for the vertical velocity fields shown in Fig. 5.1.

is strongest, these results display accurate velocity spectra but yield no improvement in the temperature spectra, suggesting that the temperature still suffers from artificial viscosity in these cases. Apparently, the improvements in the velocity spectra influence the prediction of the temperature only to a small degree.

The large-scale velocity forcing applied in M1 and M3 yields improved velocity energy levels at low wavenumbers. However, the improvement gradually vanishes at higher wavenumbers. These configurations exhibit no improvement in the temperature spectra. The cases M2 and M4 lead to an improved agreement on the velocity spectra at all wavenumbers in the center of the domain, establishing the spectrum-reconstructing property of the model described in Section 5.3.1. Nonetheless, all models underestimate the large-scale energy in the center of the domain. At these scales, the measured correlation time τ is large and therefore the model contribution is limited.

Near the wall, the horizontal velocity is accurately represented at all wavenumbers despite the fact that the energy of the vertical velocity deviates from the reference for wavenumbers larger than 15. The temperature spectra for M2 and M4 near the wall show good agreement with the reference. Comparing this to the results of M1 and M3 indicates that the prediction of near-wall

temperature is improved by the forcing of small-scale velocity despite no explicit forcing being applied to the temperature. No improvement is observed for these cases in the center of the domain, which we attribute to artificial dissipation.

The velocity spectra show no significant change when only the temperature is explicitly forced, as observed from the results of M5. This case produces an accurate temperature spectrum in the core of the domain and yields an improved spectrum near the wall. Additionally forcing the velocity significantly improves the velocity spectra, as is observed for cases M6 and M7. Here, we observe good agreement for the velocity and the temperature across all length scales in the center of the domain. In particular, a definite improvement is observed when comparing the temperature spectrum to those of M1-4. Near the wall, the horizontal velocity and the temperature are both captured accurately, while the vertical velocity still deviates for wavenumbers larger than 15. The similarity between the spectra obtained for M6 and M7 indicates that the heat flux correction described in Section 5.3.2 does indeed not lead to significant changes in the spectra.

5.4.3 Flow statistics

The mean temperature and mean heat flux are displayed in Fig. 5.7 as a function of the wall-normal distance. All models except M5 efficiently mitigate the small mean temperature discrepancy between M0 and the reference.

The mean heat flux of the no-model M0 case is consistently too low, which is a direct result of underestimating the vertical velocity. Applying the large-scale velocity forcing as done in cases M1 and M3 yields an improved heat flux. In particular, the measured heat flux near the wall shows good agreement with the reference. The mean heat flux is consistently overestimated when the velocity is forced at all wavenumbers, which is the case for M2, M4, and M6. Comparison of the results of M6 with M7 establishes that the heat flux correction described in Section 5.3.2 ensures a better prediction of the mean heat flux. Finally, only imposing the temperature spectrum deteriorates the measured heat flux, as shown by the results of M5.

These observations in combination with the energy spectra of the previous subsection expose the simplifying model assumptions discussed in Section 5.3.1. Despite accurate energy spectra of all variables, the M6 model does not yield an accurate heat flux. This indeed suggests that the energy spectra alone do not provide sufficiently strict modeling criteria for obtaining accurate coarse-grid numerical simulations, and instead benefit from additional cross-variable constraints such as the imposed heat flux.

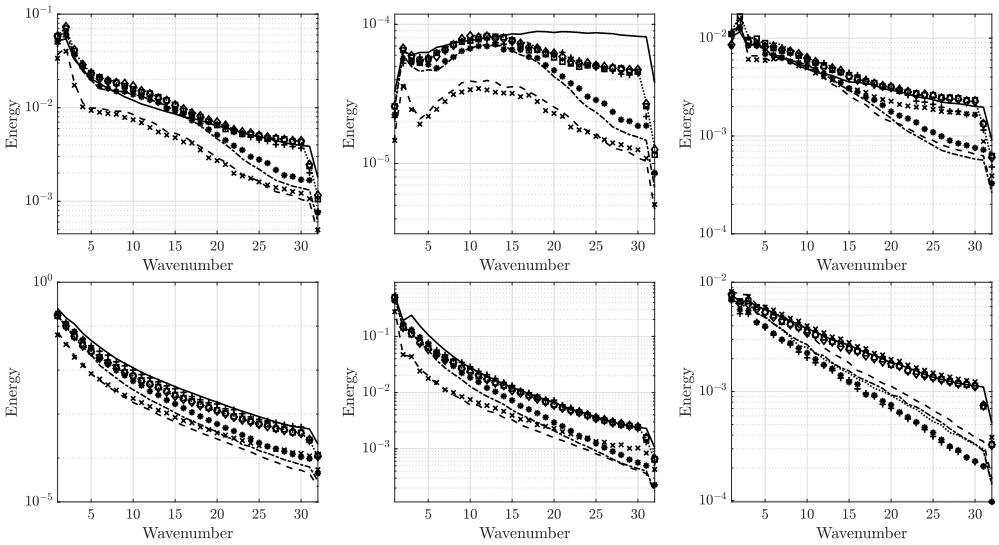


FIGURE 5.6: Time-averaged energy spectra measured along horizontal cross-sections of the domain for the horizontal velocity (left column), vertical velocity (middle column), and temperature (right column). The cross-sections are taken near the bottom wall (top row) and the core of the domain (bottom row). The cross-sections are taken at $y = 8.5 \times 10^{-4}, y = 5.5 \times 10^{-1}$ for the horizontal velocity and at $y = 5.0 \times 10^{-4}, y = 5.0 \times 10^{-1}$ for the vertical velocity and the temperature. The solid lines show the average spectra of the filtered DNS, the model results are displayed using the symbols in Table 5.1.

The rms of the velocity components are shown in the left and middle panels of Fig. 5.8 as a function of the wall-normal distance. A strong reduction of the turbulent intensity of the velocity is observed for the no-model M0 results. Similar to previous observations for case M5, only forcing the temperature does not lead to improvements in the rms of the velocity. All other model configurations lead to a comparable improvement in the rms of the horizontal velocity. A slight difference between the stochastically forced and deterministically forced solutions may be distinguished in the rms profiles of the horizontal velocity, visible in the results of M3-4. Comparable results are observed for the rms of the vertical velocity, where all models except M0 and M5 display good agreement with the reference.

The average temperature fluctuations are shown in the right panel of Fig. 5.8. We observe that all model configurations except M0 and M5 predict the wall-normal distance of the peak of the fluctuations accurately. However, the model overestimates the maximal predicted rms by 7.5% to 18%.

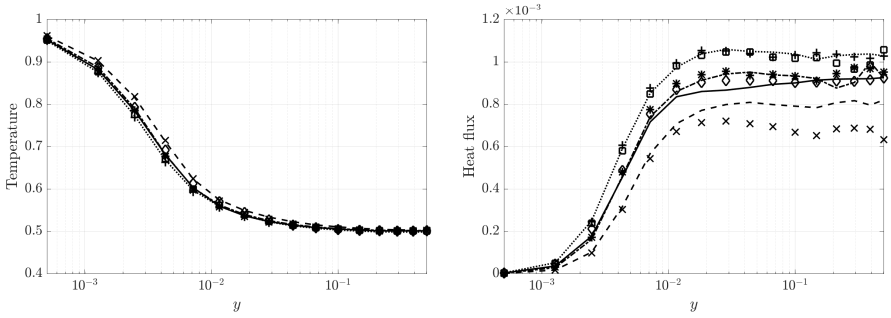


FIGURE 5.7: Comparison of the time-averaged temperature (left) and time-averaged heat flux (right) measured along horizontal cross-sections of the domain and displayed as a function of the wall-normal distance. The solid line shows the mean values of the filtered DNS, the model results are displayed using the symbols in Table 5.1.

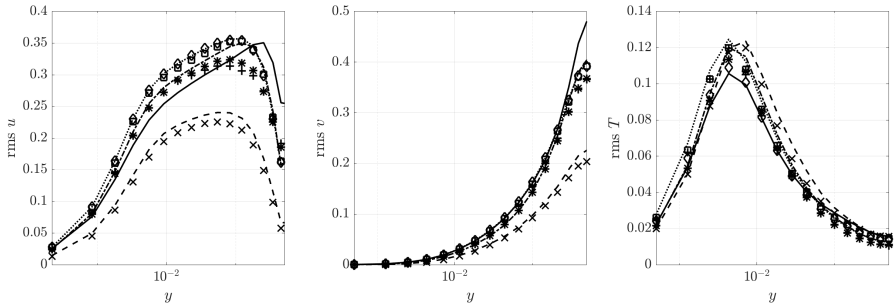


FIGURE 5.8: Root mean square (rms) of the horizontal velocity (left), vertical velocity (middle), and temperature (right), measured along horizontal cross-sections of the domain and displayed as a function of the wall-normal distance. The solid line shows the rms values of the filtered DNS, the model results are displayed using the symbols in Table 5.1.

5.4.4 Total kinetic energy and heat flux

A comparison of the rolling mean of the total kinetic energy (KE) is shown in Fig. 5.9. The improvement obtained by M1-4, M6, and M7 is evident. At $t = 400$, the mean of the KE for M1 is approximately 31% of the reference KE. Only forcing the temperature, shown by M5, deteriorates the total energy and yields roughly 27% of the reference value. The other models contain between 72% and 77% of the reference value. It is reasonable to assume that this discrepancy is predominantly caused by the model underestimating the energy

in large scales in the center of the domain, as was discussed in Section 5.4.2.

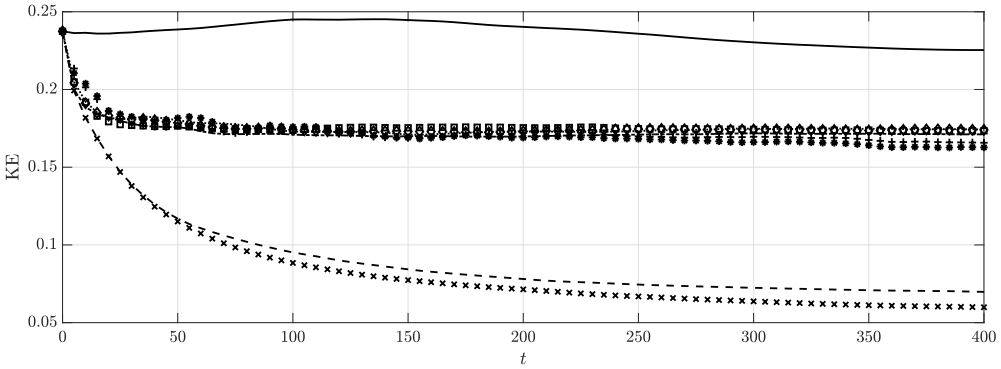


FIGURE 5.9: Comparison of the rolling mean of the kinetic energy (KE) over time. The solid line shows the KE of the filtered DNS, and the model results are displayed using the symbols in Table 5.1.

A quantification of the total heat flux in the domain is provided by comparing the time-averaged Nusselt number, shown in Fig. 5.10. Note that the reference value $Nu = 95$ is shown with 5% error margins. The no-model coarse-grid simulation leads to an underestimated heat flux resulting from the reduced velocity magnitude induced by artificial dissipation. The temperature forcing in case M5 was previously shown to not yield any improvements in the mean temperature or the velocity and does therefore not improve the Nusselt number estimate. A correction of the large-scale velocity features in configurations M1 and M3 leads to a very accurate Nusselt number estimate. Nonetheless, an accurate description of the velocity does not guarantee an accurate heat flux. This is underpinned by the results of M2, M4, and M6, which all exhibit an accurate representation of large and small velocity features, but consistently overestimate the Nusselt number. Finally, we observe that this adverse effect is efficiently mitigated by the heat flux correction, as becomes evident from the resulting Nusselt number estimate of M7.

5.5 Concluding remarks

In this chapter, we have proposed a data-driven model for coarse numerical fluid simulations and assessed its performance when applied to two-dimensional Rayleigh-Bénard convection. Statistical information of Fourier coefficients of a reference direct numerical simulation was used to infer model parameters, which constituted a forcing term for reproducing the reference energy spectra. The model parameters are defined such that the model weighs strongly towards

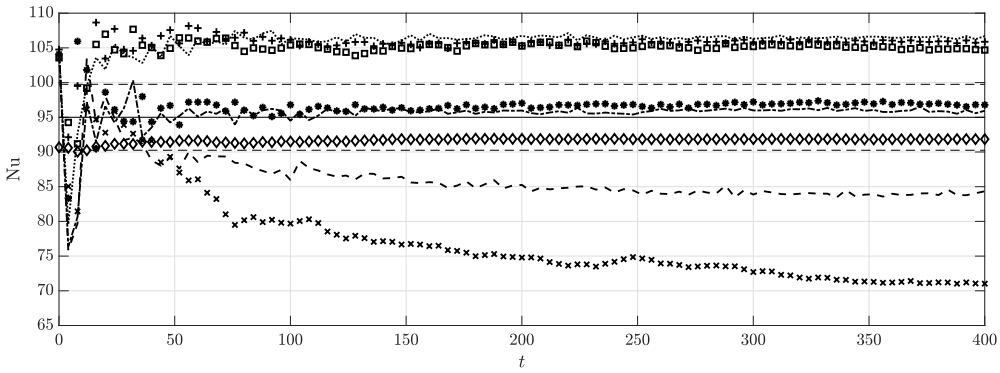


FIGURE 5.10: Comparison of the rolling mean of the Nusselt number over time. The solid line at $Nu = 95$ shows the theoretically predicted value, with 5% error margins given by the dashed lines. The model results are displayed using the symbols in Table 5.1.

the small scales. Various model configurations were applied to gain insight into the model performance, generally leading to improved results compared to using no model.

Applying the model at all wavelengths resulted in significant improvement of the spectra both near the walls and near the center of the domain, which established that the model had its desired effect on the numerical solution. Additionally, the application of the model was found to yield improved estimates of flow statistics. In particular, the average turbulent fluctuations and average temperature improved significantly compared to the no-model case. The total kinetic energy was found to improve upon using the model but highlighted that the large-scale features might benefit from less assumptive approaches. Finally, the measured total heat flux was accurately captured for several model configurations, although accurately reconstructing energy spectra was shown not to be a sufficient criterion for this purpose. The latter problem was efficiently alleviated by including a constraint on the average heat flux in the model.

Future work will be dedicated to expanding the proposed model by consulting Kalman filtering theory. Specifically, the interactions between the Fourier modes can be explicitly represented by including covariance estimates in the model. This would additionally serve to verify at which frequencies the Fourier modes evolve independently, which is expected to result in a better understanding of the modeling of small-scale flow features. Alternatively, spatially coherent structures can be included in the model by applying proper orthogonal decomposition to the reference data, as demonstrated in [58]. Although no

assumptions are made about the numerical method or adopted coarse resolution in the formulation of the model, further numerical experiments adopting a different resolution or discretization may be carried out to assess the robustness of the model.

Chapter 6

Conclusions and outlook

The goal of this thesis was the study and development of data-driven stochastic models for numerical simulations of geophysical fluid flows on coarse computational grids. This was achieved by decomposing spatiotemporal data into fixed spatial basis functions and corresponding time series. The latter were modeled as stochastic processes or served to define stochastic forcing, yielding systematic approaches to quantify uncertainty and model subgrid-scale effects on resolved dynamics. Here, we summarize our main findings and specify challenges and directions for future research.

Chapter 2 dealt with exact error measurements and consequent reduced-order corrections for coarsened numerical simulations. A numerical study was carried out for the one-dimensional shallow water equations, governed by the evolution of momentum and free surface height. A finite difference (FD) method and a finite element (FE) method were employed, representative of much-used numerical methods. We presented a measurement procedure that exactly captured the subgrid-scale processes on coarse computational grids over one time step, by using high-resolution numerical data. These measurements consisted of the effects of unresolved dynamics and discretization error on the resolved scales of motion and yielded, for the specified initial conditions, data with which the high-resolution numerical solution could be perfectly reconstructed. The proper orthogonal decomposition (POD) algorithm was subsequently applied to the subgrid-scale data to obtain a reduced-order correction. The POD modes (also referred to as empirical orthogonal functions or EOFs) differed for the methods investigated, due to the coarse-grid measurements being strongly dependent on the adopted discretization.

Application of the reduced-order correction showed that a significant error reduction can already be achieved by using a small number of modes in the correction term, which was demonstrated on two coarse computational grids. Additionally, the error reduction was found to depend strongly on the corrected variable. Namely, correcting only the velocity in the FD method and only the

free surface height in the FE method led to considerable error reduction. This corresponds to correcting the variables that dominated the discretization errors in the adopted methods. On the contrary, only correcting the free surface height in the FD method and the velocity in the FE method yielded at most a small improvement in the quality of the numerical solution. The chapter concluded with a test of the reduced-order correction applied in numerical simulations with perturbed initial conditions. While the measurements and corresponding correction terms were defined for one specific initial condition, such perfect estimation of forcing parameters is usually not available in numerical simulations of fluid flows. By perturbing the initial conditions, two new test cases were defined where the previously obtained measurements could be considered as an approximation of the true measurements. This may be regarded as a more realistic modeling setting than the earlier results presented in this chapter. In the cases with perturbed initial conditions, the prescribed correction was still found to yield a substantial error reduction. This result suggested that the quality of coarse-grid numerical simulations can already be considerably improved by using a small number of POD modes and that the time series modeling tolerates some level of approximation without a significant loss of accuracy.

Chapter 3 concerned a study of uncertainty quantification for the two-dimensional Euler equations on the unit square. An extension of the work of [40] was presented, where stochastic advection by Lie transport (SALT) [85] was employed to quantify uncertainty. In SALT, a stochastic forcing term is added to the advection velocity and, as a result, Casimir functions remain constant under the stochastic perturbations. A high-resolution numerical simulation was used to compute Lagrangian trajectories, which were compared to trajectories based on a filtered version of this solution. The difference between the Lagrangian trajectories defined a space-time array of measurements and acted as input for the POD algorithm. When executing the POD algorithm using the singular value decomposition (SVD), one simultaneously obtains the spatial profiles (POD modes), their captured energy (the eigenvalues), and the corresponding time series. Thus, the time series data can readily be used to derive a stochastic forcing without a significant increase in computational costs. The novelty of this work consisted of the derivation and application of two data-driven types of stochastic processes per POD mode, which lead to smaller ensemble spreads without loss of accuracy. The resulting stochastic forcing used either uncorrelated noise with an underlying distribution as the empirical probability density function (pdf) of the time series data or noise with a correlation time equal to that of the time series. This way, the generated signals mimicked the statistical properties of the measurements. The complete stochastic forcing term consisted of 225 POD modes accounting for 90% of the

variability of the measurements, out of a total of 4096 modes, each multiplied by noise generated from an independent stochastic process as described above.

A new reference solution was computed in order to assess the quality of the stochastic models. This reference was obtained by adding the exact measurements as a deterministic forcing term to the advection velocity, thereby eliminating the effects of discretization error in the comparison between the models and the reference. The main finding was that using the estimated pdfs or the time-correlated noise yielded stochastic ensembles with a smaller spread than when Gaussian noise was used. At the same time, the ensemble mean error of the newly proposed methods was generally found to be smaller than when Gaussian noise was employed for the stochastic forcing. These results suggest that a strict uncertainty quantification, in terms of ensemble spread, may be obtained without loss of accuracy and without an increase in computational costs by generating random signals that adhere to the statistical properties of the original measurements and using these signals to define a stochastic forcing term.

In Chapter 4, we turned our attention to data-driven stochastic subgrid-scale modeling. For this purpose, we introduced a forcing that could be used deterministically or stochastically, with forcing parameters that depended only on the available reference data. The forcing entered the governing equations at the level of the spectral coefficients of the numerical solution as a mean-reverting stochastic process. The quality of the model was assessed for the two-dimensional Euler equations on the sphere. The presented approach can be regarded as an extension of the work of Chapter 2. Instead of having a prescribed forcing specified for a single initial condition, the proposed method dynamically estimates a forcing term based on the current state of the dynamical system. The forcing acted on the coefficients of the spherical harmonic basis functions, which are determined up to the resolution of the reference data. Therefore, the forcing relied solely on the availability of high-resolution data and required no assumptions about the adopted discretization or resolution. All necessary data came in the form of the time series of spherical harmonic coefficients of the reference solution. Similar to data-driven large-eddy simulation, a priori collected data served to generate the forcing parameters. Subsequently, the forcing followed from a data assimilation method, where a deterministic mean-reverting term and stochastic term are included in the governing equations. The choice of parameters was such that the reference kinetic energy spectrum was reproduced in coarse numerical solutions and that the model weighed heavily toward small scales.

The proposed model was applied to the two-dimensional Euler equations on the sphere. Two low resolutions were considered to demonstrate that the forcing parameters were independent of the adopted resolution. On both coarse

computational grids, applying the model yielded a good agreement of the energy spectra, particularly at small spatial scales. This was a direct result of a striking agreement between the reference means and variances of the spectral coefficients and the corresponding values attained when applying the model. The measured correlation times of large-scale modes in the coarse simulations improved when only the deterministic component of the forcing was included, whereas no structural improvement was observed when also including the stochastic component. Finally, long-time simulation results showed that the model was capable of producing stable and accurate large-scale dynamics, indicating that the method is potentially useful for the study of the long-time behavior of the Euler equations.

Chapter 5 discussed data-driven subgrid-scale modeling of two-dimensional Rayleigh-Bénard convection. The model proposed in Chapter 4 was employed at the level of the Fourier coefficients of the numerical solution over horizontal cross-sections of the domain, thereby leading to a forcing that could be applied deterministically or stochastically. Physical effects in Rayleigh-Bénard convection, such as buoyancy effects and boundary layer formation near the walls, pose challenges for numerically resolving the flow. Nonetheless, the model was generally found to produce satisfactory results in terms of kinetic energy spectra and flow statistics. Several model configurations were studied by varying the variable being forced, the wavenumbers at which the model was applied, and whether the stochastic forcing term was included. In addition, a heat flux correction was introduced that could be used in conjunction with the spectrum-recovering model. Imposing this correction on the phases of the Fourier coefficients of the temperature field led to a flow constraint that respected the incompressibility of the velocity field and did not alter the energy spectra of the numerical solution, yielding an additional physical constraint for the model that did not adversely affect the solution.

Application of the model resulted in a significant improvement of measured flow statistics compared to the coarse-grid no-model setting. In fact, the model largely provided an accurate reconstruction of the energy spectra for the forced variables, both near the walls and in the center of the domain. This established that the method had its desired effect on the flow simulation. Additionally, improvements of the mean temperature profile and the root mean square profiles of the velocity components and the temperature were obtained when applying the model. The average heat flux in the domain showed improvement upon model application but was not guaranteed to follow the reference value closely without the inclusion of the heat flux constraint.

Outlook

The studies presented in this thesis were undertaken to explore the use of data-driven stochastic forcing in the context of uncertainty quantification and subgrid-scale modeling for fluid flows on coarse computational grids. We believe that the development of subgrid-scale data measurement procedures, algorithms for the efficient use of data in stochastic forcing, and a general method for data-driven stochastic forcing of fluid flows has contributed to tackling these challenges in computational fluid dynamics. Nonetheless, further work is required to investigate robustness, generalizability, and weaknesses of the approaches presented in this thesis. Below, we provide ideas for future research to address these points.

Machine learning of proper orthogonal decomposition basis coefficients. The measurement and subsequent decomposition of subgrid-scale processes presented in Chapter 2 can be applied in machine learning of turbulence closure models. The measurements represent the effect of coarsening on the dynamics, which can be compensated by a machine-learned closure model. Through the use of the singular value decomposition (SVD), any finite truncation of the proper orthogonal decomposition (POD) is the optimal truncation of the measurements at the specified number of degrees of freedom. Therefore, one can already obtain a reasonably accurate representation of the expected subgrid-scale processes for a given system state, using only a reduced number of POD modes. Machine learning techniques have shown to be capable of empirically finding the expected sub-grid scale contributions in numerical flow simulations [11, 101, 162]. We anticipate that applying machine learning techniques to estimate the coefficients of the POD modes of subgrid-scale contributions will also lead to satisfying results for this purpose. The benefit of estimating these coefficients is that the finite truncation of the number of modes can significantly reduce the number of degrees of freedom, even when compared to the number of degrees of freedom in a coarse-grid fluid problem. This reduces the complexity of the minimization problem underlying the machine learning algorithm and might lead to a reduced training time without a significant loss of accuracy.

Robustness of the proposed data-driven subgrid-scale model. The proposed data-driven model for coarse grid simulations used in Chapters 4 and 5 should be investigated more closely to study its robustness. Despite the satisfactory results obtained for the Euler equations on the sphere and two-dimensional Rayleigh-Bénard convection, the nudging strength and wavenumbers at which the model is applied remain a modeling choice. The aim of the model is to reconstruct reference kinetic energy spectra. The adopted definitions of the mean

and variance in the mean-reverting forcing ensure that this goal is achieved for sufficiently large nudging strengths. Furthermore, literature on continuous data assimilation has shown that a range of nudging strengths leads to convergence towards the observed reference solution [67]. Naturally, this warrants further research on the 'best' nudging strength. In the ideal case, the nudging strength should be determined systematically. For example, one can draw inspiration from ensemble Kalman filters [59], where the extent to which the correction is applied depends on measured covariances in the data and on model uncertainty. Alternatively, the nudging strength can be variable and defined such that measured reference inter-scale energy or enstrophy transfers are reproduced [157] or such that global quantities of interest are reproduced [54]. Complementary to this is the assessment of the model in the case of a limited amount of data. Sparse and noisy data affect the estimates of the forcing parameters and the effect on the model performance is worthwhile to investigate.

The proposed data-driven subgrid-scale model as a general method for coarsened computational fluid dynamics. The data-driven model presented in Chapters 4 and 5 requires no assumptions about the underlying governing equations or adopted discretization method. In fact, it only requires several statistics of the time series of basis coefficients, measured in a statistically steady state. The choice of basis functions has been shown to be flexible. For example, the forcing in Chapter 5 acted on the level of Fourier coefficients of the numerical solution, and similar results were obtained when POD modes were chosen as the basis for the solution [58]. In addition, current research efforts are dedicated to applying the model to other geophysical flow problems. Recent application of the model to the quasi-geostrophic equations on the sphere has shown that it leads to the formation of zonal jets, otherwise unattainable at coarse computational grids. Having obtained good results in a variety of flow settings, we can label the model approach as a general method for coarse numerical simulations of fluid flows. Its success encourages further work in the intersection of data assimilation and data-driven large-eddy simulation. In fact, application to a larger class of dynamical systems, outside of fluid dynamics, may also be considered since no assumptions are made about the underlying equations.

Subgrid-scale modeling from a geometric mechanical viewpoint. Many commonly used models in geophysical fluid dynamics (GFD) possess conservation laws that follow from the rich mathematical structure underlying the governing equations. In these models, this structure manifests in an infinite number of conserved quantities referred to as Casimirs. In this thesis, this particular aspect has been prominent in the use of stochastic advection by Lie transport in Chapter 3, which respects these conservation laws, and reappeared in Chapter

4 in the use of a Casimir-preserving numerical integrator for the Euler equations on the sphere. However, this feature of GFD models was not taken into account in the development of the data-driven subgrid-scale model presented in Chapter 4. For example, one could decompose the designed forcing into an energy-preserving component and a Casimir-preserving component and interpret their respective influence on the numerical solution from a geometric mechanical point of view. Approaching the subgrid-scale modeling challenge from this viewpoint might be beneficial for systematically and rigorously deriving subgrid-scale models for a large class of geophysical fluid systems. The use of structure-preserving integrators in such studies is beneficial since it allows for distinguishing the roles of energy preservation and Casimir preservation in geophysical fluid flows and subgrid-scale modeling. In addition, this may lead to an improved physical understanding of why structure preservation is desired.

Appendix A

Description of the compatible finite element method

We present the FE approach in a number of steps. Given the divergence-conforming space

$$H(\mathrm{div}, \Omega) = \{\mathbf{v} \in (L^2(\Omega))^d | \nabla \cdot \mathbf{v} \in L^2(\Omega)\}, \tag{A.1}$$

where Ω denotes the (periodic) domain and d its dimension, the function spaces $\mathbb{V}_u$ for the velocity field and $\mathbb{V}_\eta$ for total depth field are set up to satisfy

$$\begin{array}{ccc} H(\mathrm{div}; \Omega) & \xrightarrow{\nabla \cdot} & L^2(\Omega) \\ \downarrow{\scriptstyle \pi^1} & & \downarrow{\scriptstyle \pi^2} \\ \mathbb{V}_u(\Omega) & \xrightarrow{\nabla \cdot} & \mathbb{V}_\eta(\Omega) \end{array}$$

for bounded projections π^1, π^2 such that the diagram commutes. In the one-dimensional case, the divergence reduces to the single derivative ∂_x, and a pair of compatible spaces for u and η is given, e.g., by

$$\mathbb{V}_u = CG_k(\Omega), \qquad \mathbb{V}_\eta = DG_{k-1}(\Omega), \tag{A.2}$$

where $CG_k(\Omega)$ denotes the k^{th} polynomial order continuous Galerkin space and $DG_{k-1}(\Omega)$ the $(k-1)^{th}$ polynomial order discontinuous Galerkin space.

The governing shallow water equations (2.1) can now be discretized such that the divergence in the continuity equation is considered strongly, while the gradient in the momentum equation is imposed weakly, leading to the mixed

formulation

$$\langle w, u_t \rangle - \left\langle w_x, \frac{1}{2}u^2 + \frac{1}{\mathrm{Fr}^2}(\eta - b) \right\rangle = 0 \qquad \forall w \in \mathbb{V}_u, \tag{A.3}$$

$$\eta_t + F_x = 0, \tag{A.4}$$

where $\langle ., . \rangle$ denotes the L^2 inner product, and the flux F in (A.4) is given by the L^2-projection of ηu into the velocity space, i.e.,

$$\langle w, F - \eta u \rangle = 0 \qquad \forall w \in \mathbb{V}_u. \tag{A.5}$$

In this so-called compatible framework, the continuity equation is formulated in strong form, as the derivative in x maps the flux F into $\mathbb{V}_\eta$. Further, no surface integral is required for the spatial derivative's weak formulation in the momentum equation, since $w_x \in \mathbb{V}_\eta$ is well-defined everywhere. The above space discretization conserves mass locally as well as a discrete energy globally (for details, see e.g. [121]). Finally, we also incorporate transport stabilization for η without compromising on the latter two conservation properties, by modifying equations (A.3) - (A.4) according to [169]

$$\langle w, u_t \rangle + \langle P_x, w \rangle - \int_\Gamma [\![P]\!] \left\{ \frac{w}{\eta} \right\} \tilde{\eta} = 0 \qquad \forall w \in \mathbb{V}_u, \tag{A.6}$$

$$\langle \phi, \eta_t \rangle - \langle \phi_x, F \rangle + \int_\Gamma [\![\phi]\!] \left\{ \frac{F}{\eta} \right\} \tilde{\eta} \, dS = 0 \qquad \forall \phi \in \mathbb{V}_\eta, \tag{A.7}$$

where in a similar fashion to F, P is given by an L^2-projection of the form

$$\left\langle \phi, P - \left(\frac{1}{2}u^2 + \frac{1}{\mathrm{Fr}^2}(\eta - b) \right) \right\rangle = 0 \qquad \forall \phi \in \mathbb{V}_\eta. \tag{A.8}$$

The integrals are over all cell boundaries (which in 1D reduces to evaluations at single points), and $[\![.]\!]$ and $\{.\}$ denote difference and average values, respectively. Finally, $\tilde{\eta}$ denotes the upwind value along the given cell boundary. Note that in the adopted Runge-Kutta scheme, the projections F and P need to be evaluated separately before each evaluation of the dynamic contribution. In Chapter 2, we consider the lowest polynomial order $k = 1$ for the mixed compatible setup. The scheme and varying resolution mesh hierarchies are implemented using the automated finite element toolkit Firedrake, see [142, 126][1], which in turn relies on PETSc, see [9, 10].

[1] For further details, visit http://firedrakeproject.org

Bibliography

[1] RJ Adrian and J Westerweel. *Particle Image Velocimetry*. Cambridge University Press, 2011.

[2] Niraj Agarwal, Dmitri Kondrashov, Peter Dueben, Evgenii Ryzhov, and Pavel Berloff. A comparison of data-driven approaches to build low-dimensional ocean models. *Journal of Advances in Modeling Earth Systems*, 13(9):e2021MS002537, 2021.

[3] Guenter Ahlers, Siegfried Grossmann, and Detlef Lohse. Heat transfer and large scale dynamics in turbulent rayleigh-bénard convection. *Reviews of modern physics*, 81(2):503, 2009.

[4] MU Altaf, ES Titi, T Gebrael, OM Knio, L Zhao, MF McCabe, and Ibrahim Hoteit. Downscaling the 2D Bénard convection equations using continuous data assimilation. *Computational Geosciences*, 21(3):393–410, 2017.

[5] Akio Arakawa and Vivian R Lamb. Computational design of the basic dynamical processes of the UCLA general circulation model. *General circulation models of the atmosphere*, 17(Supplement C):173–265, 1977.

[6] HM Arnold, IM Moroz, and TN Palmer. Stochastic parametrizations and model uncertainty in the Lorenz'96 system. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1991):20110479, 2013.

[7] Abderrahim Azouani, Eric Olson, and Edriss S Titi. Continuous data assimilation using general interpolant observables. *Journal of Nonlinear Science*, 24(2):277–304, 2014.

[8] Abderrahim Azouani and Edriss S Titi. Feedback control of nonlinear dissipative systems by finite determining parameters - a reaction-diffusion paradigm. *arXiv preprint arXiv:1301.6992*, 2013.

[9] Satish Balay, Shrirang Abhyankar, Mark Adams, Jed Brown, Peter Brune, Kris Buschelman, Lisandro Dalcin, Alp Dener, Victor Eijkhout, W Gropp, et al. PETSc users manual. 2019.

[10] Satish Balay, William D Gropp, Lois Curfman McInnes, and Barry F Smith. Efficient management of parallelism in object-oriented numerical software libraries. In *Modern software tools for scientific computing*, pages 163–202. Springer, 1997.

[11] Andrea Beck, David Flad, and Claus-Dieter Munz. Deep neural networks for data-driven LES closure models. *Journal of Computational Physics*, 398:108910, 2019.

[12] Andrea Beck and Marius Kurz. A perspective on machine learning methods in turbulence modeling. *GAMM-Mitteilungen*, 44(1):e202100002, 2021.

[13] Gal Berkooz, Philip Holmes, and John L Lumley. The proper orthogonal decomposition in the analysis of turbulent flows. *Annual review of fluid mechanics*, 25(1):539–575, 1993.

[14] Erik Bernsen, Onno Bokhove, and Jaap JW van der Vegt. A (dis) continuous finite element model for generalized 2D vorticity dynamics. *Journal of Computational Physics*, 211(2):719–747, 2006.

[15] Hakima Bessaih, Eric Olson, and Edriss S Titi. Continuous data assimilation with stochastically noisy data. *Nonlinearity*, 28(3):729, 2015.

[16] Animikh Biswas, Ciprian Foias, Cecilia F Mondaini, and Edriss S Titi. Downscaling data assimilation algorithm with applications to statistical solutions of the Navier-Stokes equations. In *Annales de l'Institut Henri Poincaré C, Analyse non linéaire*, volume 36, pages 295–326. Elsevier, 2019.

[17] Dirk Blömker, Kody Law, Andrew M Stuart, and Konstantinos C Zygalakis. Accuracy and stability of the continuous-time 3DVAR filter for the Navier–Stokes equation. *Nonlinearity*, 26(8):2193, 2013.

[18] Guido Boffetta and S Musacchio. Evidence for the double cascade scenario in two-dimensional turbulence. *Physical Review E*, 82(1):016307, 2010.

[19] Thomas Bolton and Laure Zanna. Applications of deep learning to ocean data inference and subgrid parameterization. *Journal of Advances in Modeling Earth Systems*, 11(1):376–399, 2019.

[20] Martin Bordemann, Jens Hoppe, Peter Schaller, and Martin Schlichenmaier. $\mathfrak{gl}(\infty)$ and geometric quantization. *Communications in Mathematical Physics*, 138(2):209–244, 1991.

[21] Martin Bordemann, Eckhard Meinrenken, and Martin Schlichenmaier. Toeplitz quantization of Kähler manifolds and $\mathfrak{gl}(n), n \to \infty$ limits. *Communications in Mathematical Physics*, 165(2):281–296, 1994.

[22] F. van der Bos, JJW van der Vegt, and B.J. Geurts. A multi-scale formulation for compressible turbulent flows suitable for general variational discretization techniques. *Computer Methods in Applied Mechanics and Engineering*, 196(29-30):2863–2875, 2007.

[23] Fedderik Bos and Bernard J. Geurts. Computational error-analysis of a discontinuous Galerkin discretization applied to large-eddy simulation of homogeneous turbulence. *Computer Methods in Applied Mechanics and Engineering*, 199(13-16):903–915, 2010.

[24] Marguerite L Brown, Olivier Pauluis, and Edwin P Gerber. Scaling for saturated moist quasi-geostrophic turbulence. *Journal of the Atmospheric Sciences*, 2023.

[25] Roberto Buizza, M Miller, and Tim N Palmer. Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 125(560):2887–2908, 1999.

[26] Judith E Bunder, Jayaraman Divahar, Ioannis G Kevrekidis, Trent W Mattner, and Anthony J Roberts. Large-scale simulation of shallow water waves via computation only on small staggered patches. *International Journal for Numerical Methods in Fluids*, 93(4):953–977, 2021.

[27] John Burkardt, Max Gunzburger, and Hyung-Chun Lee. POD and CVT-based reduced-order modeling of Navier–Stokes flows. *Computer Methods in Applied Mechanics and Engineering*, 196(1-3):337–355, 2006.

[28] Meng Cao and AJ Roberts. Multiscale modelling couples patches of non-linear wave-like simulations. *IMA Journal of Applied Mathematics*, 81(2):228–254, 2016.

[29] Emilio Castronovo, John Harlim, and Andrew J Majda. Mathematical test criteria for filtering complex systems: plentiful observations. *Journal of Computational Physics*, 227(7):3678–3714, 2008.

[30] W Cazemier, RWCP Verstappen, and AEP Veldman. Proper orthogonal decomposition and low-dimensional models for driven cavity flows. *Physics of Fluids*, 10(7):1685–1699, 1998.

[31] Jule Charney, Milton Halem, and Robert Jastrow. Use of incomplete historical data to infer the present state of the atmosphere. *Journal of the Atmospheric Sciences*, 26(5):1160–1163, 1969.

[32] Chuchu Chen, David Cohen, Raffaele D'Ambrosio, and Annika Lang. Drift-preserving numerical integrators for stochastic Hamiltonian systems. *Advances in Computational Mathematics*, 46:1–22, 2020.

[33] HM Christensen, S-J Lock, IM Moroz, and TN Palmer. Introducing independent patterns into the stochastically perturbed parametrization tendencies (SPPT) scheme. *Quarterly Journal of the Royal Meteorological Society*, 143(706):2168–2181, 2017.

[34] P Cifani, JGM Kuerten, and BJ Geurts. Highly scalable DNS solver for turbulent bubble-laden channel flow. *Computers & Fluids*, 172:67–83, 2018.

[35] Paolo Cifani, Sagy Ephrati, and Milo Viviani. Sparse-stochastic model reduction for 2D euler equations. *arXiv preprint arXiv:2301.06326*, 2023.

[36] Paolo Cifani, Milo Viviani, Erwin Luesink, Klas Modin, and Bernard J Geurts. Casimir preserving spectrum of two-dimensional turbulence. *Physical Review Fluids*, 7(8):L082601, 2022.

[37] Paolo Cifani, Milo Viviani, and Klas Modin. An efficient geometric method for incompressible hydrodynamics on the sphere. *Journal of Computational Physics*, 473:111772, 2023.

[38] David Cohen and Gilles Vilmart. Drift-preserving numerical integrators for stochastic Poisson systems. *International Journal of Computer Mathematics*, 99(1):4–20, 2022.

[39] Colin Cotter, Dan Crisan, Darryl D Holm, Wei Pan, and Igor Shevchenko. Modelling uncertainty using circulation-preserving stochastic transport noise in a 2-layer quasi-geostrophic model. *arXiv preprint arXiv:1802.05711*, 2018.

[40] Colin Cotter, Dan Crisan, Darryl D Holm, Wei Pan, and Igor Shevchenko. Numerically modeling stochastic Lie transport in fluid dynamics. *Multiscale Modeling & Simulation*, 17(1):192–232, 2019.

[41] Colin Cotter, Dan Crisan, Darryl D Holm, Wei Pan, and Igor Shevchenko. A particle filter for stochastic advection by Lie transport: a case study for the damped and forced incompressible two-dimensional

euler equation. *SIAM/ASA Journal on Uncertainty Quantification*, 8(4):1446–1492, 2020.

[42] Colin J. Cotter, Dan Crisan, Darryl Holm, Wei Pan, and Igor Shevchenko. Data assimilation for a quasi-geostrophic model with circulation-preserving stochastic transport noise. *Journal of Statistical Physics*, 179(5):1186–1221, 2020.

[43] Colin J Cotter, Georg A Gottwald, and Darryl D Holm. Stochastic partial differential fluid equations as a diffusive limit of deterministic Lagrangian multi-time dynamics. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2205):20170388, 2017.

[44] Colin J Cotter and Jemma Shipton. Mixed finite elements for numerical weather prediction. *Journal of Computational Physics*, 231(21):7076 – 7091, 2012.

[45] Philippe Courtier, E Andersson, W Heckley, D Vasiljevic, M Hamrud, A Hollingsworth, F Rabier, M Fisher, and J Pailleux. The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: Formulation. *Quarterly Journal of the Royal Meteorological Society*, 124(550):1783–1807, 1998.

[46] Dan Crisan, Darryl D Holm, James-Michael Leahy, and Torstein Nilssen. Variational principles for fluid dynamics on rough paths. *Advances in Mathematics*, 404:108409, 2022.

[47] Daan Crommelin and Eric Vanden-Eijnden. Subgrid-scale parameterization with conditional Markov chains. *Journal of the Atmospheric Sciences*, 65(8):2661–2675, 2008.

[48] Roger Daley. Estimating model-error covariances for application to atmospheric data assimilation. *Monthly weather review*, 120(8):1735–1746, 1992.

[49] Roger Daley. *Atmospheric data analysis*. Number 2. Cambridge University Press, 1993.

[50] Andrew Dawson. eofs: A library for EOF analysis of meteorological, oceanographic, and climate data. *Journal of Open Research Software*, 4(1), 2016.

[51] Luc Devroye. Nonuniform random variate generation. *Handbooks in operations research and management science*, 13:83–121, 2006.

[52] Jesse Dorrestijn, Daan T Crommelin, A Pier Siebesma, and Harm JJ Jonker. Stochastic parameterization of shallow cumulus convection estimated from high-resolution model data. *Theoretical and Computational Fluid Dynamics*, 27:133–148, 2013.

[53] Wouter Edeling and Daan Crommelin. Towards data-driven dynamic surrogate models for ocean flow. In *Proceedings of the platform for advanced scientific computing conference*, pages 1–10, 2019.

[54] Wouter Edeling and Daan Crommelin. Reducing data-driven dynamical subgrid scale models by physical constraints. *Computers & Fluids*, 201:104470, 2020.

[55] Sagy Ephrati, Paolo Cifani, Milo Viviani, and Bernard Geurts. Data-driven stochastic spectral modeling for coarsening of the two-dimensional Euler equations on the sphere. *arXiv preprint arXiv:2304.12007*, 2023.

[56] Sagy R Ephrati, Paolo Cifani, Erwin Luesink, and Bernard J Geurts. Data-driven stochastic Lie transport modelling of the 2D Euler equations. *Journal of Advances in Modeling Earth Systems*, page e2022MS003268, 2023.

[57] Sagy R Ephrati, Erwin Luesink, Golo Wimmer, Paolo Cifani, and Bernard J Geurts. Computational modeling for high-fidelity coarsening of shallow water equations based on subgrid data. *Multiscale Modeling & Simulation*, 20(4):1468–1489, 2022.

[58] SR Ephrati, P Cifani, and BJ Geurts. Stochastic data-driven POD-based modeling for high-fidelity coarsening of two-dimensional Rayleigh-Bénard turbulence. In *13th ERCOFTAC Workshop on Direct & Large Eddy Simulation 2022*. ERCOFTAC, 2022.

[59] Geir Evensen et al. *Data assimilation: the ensemble Kalman filter*, volume 2. Springer, 2009.

[60] Aseel Farhat, Michael S Jolly, and Edriss S Titi. Continuous data assimilation for the 2D Bénard convection through velocity measurements alone. *Physica D: Nonlinear Phenomena*, 303:59–66, 2015.

[61] Jorgen S Frederiksen and Antony G Davies. Eddy viscosity and stochastic backscatter parameterizations on the sphere for atmospheric circulation models. *Journal of the Atmospheric Sciences*, 54(20):2475–2492, 1997.

[62] Jorgen S Frederiksen and Steven M Kepert. Dynamical subgrid-scale parameterizations from direct numerical simulations. *Journal of the Atmospheric Sciences*, 63(11):3006–3019, 2006.

[63] Jorgen S Frederiksen, Vassili Kitsios, Terence J O'Kane, and Meelis J Zidikheri. Stochastic subgrid modelling for geophysical and three-dimensional turbulence. In *Nonlinear and stochastic climate dynamics*, pages 241–275. Cambridge University Press, 2017.

[64] David John Gagne, Hannah M Christensen, Aneesh C Subramanian, and Adam H Monahan. Machine learning for stochastic parameterization: Generative adversarial networks in the Lorenz'96 model. *Journal of Advances in Modeling Earth Systems*, 12(3):e2019MS001896, 2020.

[65] William K George. Lectures in turbulence for the 21st century. *Chalmers University of Technology*, 550, 2013.

[66] Massimo Germano, Ugo Piomelli, Parviz Moin, and William H Cabot. A dynamic subgrid-scale eddy viscosity model. *Physics of Fluids A: Fluid Dynamics*, 3(7):1760–1765, 1991.

[67] Masakazu Gesho, Eric Olson, and Edriss S Titi. A computational study of a data assimilation algorithm for the two-dimensional Navier-Stokes equations. *Communications in Computational Physics*, 19(4):1094–1110, 2016.

[68] Bernard J. Geurts. *Direct and Large-Eddy Simulation.* De Gruyter, Berlin, Boston, 2023.

[69] Bernard J Geurts and Jochen Fröhlich. A framework for predicting accuracy limitations in large-eddy simulation. *Physics of Fluids*, 14(6):L41–L44, 2002.

[70] Bernard J. Geurts and Darryl Holm. *Alpha-modeling strategy for LES of turbulent mixing.* Springer - Turbulent flow computation, 2002.

[71] Bernard J Geurts, Darryl D Holm, and Erwin Luesink. Lyapunov exponents of two stochastic Lorenz 63 systems. *Journal of Statistical Physics*, 179(5):1343–1365, 2020.

[72] Bernard J Geurts and Fedderik van der Bos. Numerically induced high-pass dynamics in large-eddy simulation. *Physics of Fluids*, 17(12):125103, 2005.

[73] Bernard J Geurts and Fedderik van der Bos. Numerically induced high-pass dynamics in large-eddy simulation. *Physics of Fluids*, 17(12):125103, 2005.

[74] Bernardus J Geurts. *Elements of direct and large eddy simulation.* RT Edwards, Inc, 2003.

[75] B.J. Geurts and D.D. Holm. Regularization modeling for large-eddy simulation. *Physics of Fluids*, 15(1):L13–L16, 2003.

[76] Michael Ghil and Paola Malanotte-Rizzoli. Data assimilation in meteorology and oceanography. In *Advances in geophysics*, volume 33, pages 141–266. Elsevier, 1991.

[77] Mohinder S Grewal and Angus P Andrews. *Kalman filtering: Theory and Practice with MATLAB*. John Wiley & Sons, 2014.

[78] Thomas M Hamill. Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129(3):550–560, 2001.

[79] Abdel Hannachi, Ian T Jolliffe, and David B Stephenson. Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 27(9):1119–1152, 2007.

[80] J Harlim and AJ Majda. Filtering nonlinear dynamical systems with linear stochastic models. *Nonlinearity*, 21(6):1281, 2008.

[81] Dennis L Hartmann, Leslie A Moy, and Qiang Fu. Tropical convection and the energy balance at the top of the atmosphere. *Journal of Climate*, 14(24):4495–4511, 2001.

[82] Klaus Hasselmann. Stochastic climate models part I. Theory. *Tellus*, 28(6):473–485, 1976.

[83] Ernst Hellinger. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 1909(136):210–271, 1909.

[84] Desmond J Higham. An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM review*, 43(3):525–546, 2001.

[85] Darryl D Holm. Variational principles for stochastic fluid dynamics. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2176):20140963, 2015.

[86] Darryl D Holm and Erwin Luesink. Stochastic wave–current interaction in thermal shallow water dynamics. *Journal of Nonlinear Science*, 31(2):1–56, 2021.

[87] Darryl D Holm, Erwin Luesink, and Wei Pan. Stochastic mesoscale circulation dynamics in the thermal ocean. *Physics of Fluids*, 33(4):046603, 2021.

[88] Jens Hoppe. Diffeomorphism groups, quantization, and su(∞). *International Journal of Modern Physics A*, 4(19):5235–5248, 1989.

[89] Jens Hoppe and Shing-Tung Yau. Some properties of matrix harmonics on $\mathbb{S}^2$. *Communications in Mathematical Physics*, 195(1):67–77, 1998.

[90] Peter L Houtekamer, Louis Lefaivre, Jacques Derome, Harold Ritchie, and Herschel L Mitchell. A system simulation approach to ensemble prediction. *Monthly Weather Review*, 124(6):1225–1242, 1996.

[91] Sean Ingimarson, Leo G Rebholz, and Traian Iliescu. Full and reduced order model consistency of the nonlinearity discretization in incompressible flows. *Computer Methods in Applied Mechanics and Engineering*, 401:115620, 2022.

[92] P Joly, A Quarteroni, and J Rappaz. *Scientific Computation*. Springer, 2006.

[93] Leo P. Kadanoff. Turbulent heat flow: Structures and scaling. *Physics Today*, 54(8):34–39, 2001.

[94] John Kim and Parviz Moin. Application of a fractional-step method to incompressible Navier-Stokes equations. *Journal of Computational Physics*, 59(2):308–323, 1985.

[95] Vassili Kitsios and Jorgen S Frederiksen. Subgrid parameterizations of the eddy–eddy, eddy–mean field, eddy–topographic, mean field–mean field, and mean field–topographic interactions in atmospheric models. *Journal of the Atmospheric Sciences*, 76(2):457–477, 2019.

[96] Peter E Kloeden and Eckhard Platen. Stochastic differential equations. In *Numerical Solution of Stochastic Differential Equations*, pages 103–160. Springer, 1992.

[97] Gijs L Kooij, Mikhail A Botchev, Edo MA Frederix, Bernard J Geurts, Susanne Horn, Detlef Lohse, Erwin P van der Poel, Olga Shishkina, Richard JAM Stevens, and Roberto Verzicco. Comparison of computational codes for direct numerical simulations of turbulent Rayleigh–Bénard convection. *Computers & Fluids*, 166:1–8, 2018.

[98] Robert H Kraichnan. Turbulent thermal convection at arbitrary Prandtl number. *The Physics of Fluids*, 5(11):1374–1389, 1962.

[99] Robert H Kraichnan. Inertial ranges in two-dimensional turbulence. *The Physics of Fluids*, 10(7):1417–1423, 1967.

[100] R.P.J. Kunnen, B.J. Geurts, and H.J.H. Clercx. Turbulence statistics and energy budget in rotating Rayleigh–Bénard convection. *European Journal of Mechanics-B/Fluids*, 28(4):578–589, 2009.

[101] Marius Kurz, Philipp Offenhäuser, and Andrea Beck. Deep reinforcement learning for turbulence modeling in large eddy simulations. *International Journal of Heat and Fluid Flow*, 99:109094, 2023.

[102] Noé Lahaye and Vladimir Zeitlin. Decaying vortex and wave turbulence in rotating shallow water model, as follows from high-resolution direct numerical simulations. *Physics of Fluids*, 24(11):115106, 2012.

[103] Jacob A Langford and Robert D Moser. Optimal LES formulations for isotropic turbulence. *Journal of Fluid Mechanics*, 398:321–346, 1999.

[104] Kody Law, Andrew Stuart, and Kostas Zygalakis. Data assimilation. *Cham, Switzerland: Springer*, 214:52, 2015.

[105] Cecil E Leith. Climate response and fluctuation dissipation. *Journal of Atmospheric Sciences*, 32(10):2022–2026, 1975.

[106] Brian P Leonard. A stable and accurate convective modelling procedure based on quadratic upstream interpolation. *Computer Methods in Applied Mechanics and Engineering*, 19(1):59–98, 1979.

[107] Martin Leutbecher. Diagnosis of ensemble forecasting systems. In *Seminar on Diagnosis of Forecasting and Data Assimilation Systems*, pages 235–266, 2009.

[108] Edward N Lorenz. *Empirical orthogonal functions and statistical weather prediction*. Massachusetts Institute of Technology, Department of Meteorology Cambridge, 1956.

[109] Edward N Lorenz. Deterministic nonperiodic flow. *Journal of atmospheric sciences*, 20(2):130–141, 1963.

[110] Edward N Lorenz. Predictability: A problem partly solved. In *Proc. Seminar on predictability*, volume 1, 1996.

[111] Erwin Luesink. *Stochastic geometric mechanics of thermal ocean dynamics*. Thesis - Imperial College London, 2021.

[112] Erwin Luesink, Sagy Ephrati, Paolo Cifani, and Bernard Geurts. Casimir preserving stochastic Lie-Poisson integrators. *arXiv preprint arXiv:2111.13143*, 2021.

[113] Erwin Luesink and Bernard Geurts. An explicit method to determine Casimirs in 2D geophysical flows. *arXiv preprint arXiv:2302.11886*, 2023.

[114] John Leask Lumley. The structure of inhomogeneous turbulent flows. *Atmospheric turbulence and radio wave propagation*, pages 166–178, 1967.

[115] Andrew J Majda and John Harlim. *Filtering complex turbulent systems*. Cambridge University Press, 2012.

[116] Andrew J Majda, Ilya Timofeyev, and Eric Vanden Eijnden. A mathematical framework for stochastic climate models. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 54(8):891–974, 2001.

[117] Jerrold Marsden and Alan Weinstein. Coadjoint orbits, vortices, and Clebsch variables for incompressible fluids. *Physica D: Nonlinear Phenomena*, 7(1-3):305–323, 1983.

[118] John Marshall and Friedrich Schott. Open-ocean convection: Observations, theory, and models. *Reviews of geophysics*, 37(1):1–64, 1999.

[119] Romit Maulik, Omer San, Adil Rasheed, and Prakash Vedula. Subgrid modelling for two-dimensional turbulence using neural networks. *Journal of Fluid Mechanics*, 858:122–144, 2019.

[120] Robert I McLachlan. Explicit Lie-Poisson integration and the Euler equations. *Physical Review Letters*, 71(19):3043, 1993.

[121] Andrew T T McRae and Colin Cotter. Energy-and enstrophy-conserving schemes for the shallow-water equations, based on mimetic finite elements. *Quarterly Journal of the Royal Meteorological Society*, 140(684):2223–2234, 2014.

[122] Rami Mehrem. The plane wave expansion, infinite integrals and identities involving spherical Bessel functions. *Applied Mathematics and Computation*, 217(12):5360–5365, 2011.

[123] Etienne Mémin. Fluid flow dynamics under location uncertainty. *Geophysical & Astrophysical Fluid Dynamics*, 108(2):119–146, 2014.

[124] Miguel A Mendez, Andrea Ianiro, Bernd R Noack, and Steven L Brunton. *Data-Driven Fluid Mechanics: Combining First Principles and Machine Learning*. Cambridge University Press, 2023.

[125] Knud Erik Meyer, Jakob Martin Pedersen, and Oktay Özcan. A turbulent jet in crossflow analysed with proper orthogonal decomposition. *Journal of Fluid Mechanics*, 583:199–227, 2007.

[126] Lawrence Mitchell and Eike H Müller. High level implementation of geometric multigrid solvers for finite element problems: applications in atmospheric modelling. *Journal of Computational Physics*, 327:1–18, 2016.

[127] Klas Modin and Milo Viviani. A Casimir preserving scheme for long-time simulation of spherical ideal hydrodynamics. *Journal of Fluid Mechanics*, 884, 2020.

[128] Klas Modin and Milo Viviani. Canonical scale separation in two-dimensional incompressible hydrodynamics. *Journal of Fluid Mechanics*, 943:A36, 2022.

[129] Changhong Mou, Zhu Wang, David R Wells, Xuping Xie, and Traian Iliescu. Reduced order models for the quasi-geostrophic equations: A brief survey. *Fluids*, 6(1):16, 2020.

[130] Srikanth Derebail Muralidhar, Bérengère Podvin, Lionel Mathelin, and Yann Fraigneau. Spatio-temporal proper orthogonal decomposition of turbulent channel flow. *Journal of Fluid Mechanics*, 864:614–639, 2019.

[131] BT Nadiga and D Livescu. Instability of the perfect subgrid model in implicit-filtering large eddy simulation of geostrophic turbulence. *Physical Review E*, 75(4):046303, 2007.

[132] Andrea Natale, Jemma Shipton, and Colin Cotter. Compatible finite element spaces for geophysical fluid dynamics. *Dynamics and Statistics of the Climate System*, 1(1), 2016.

[133] Paul A O'Gorman and John G Dwyer. Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events. *Journal of Advances in Modeling Earth Systems*, 10(10):2548–2563, 2018.

[134] Tim Palmer. The ECMWF ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years. *Quarterly Journal of the Royal Meteorological Society*, 145:12–24, 2019.

[135] Tim N Palmer. Predicting uncertainty in forecasts of weather and climate. *Reports on progress in Physics*, 63(2):71, 2000.

[136] TN Palmer. Stochastic weather and climate models. *Nature Reviews Physics*, 1(7):463–471, 2019.

[137] Jonghwan Park and Haecheon Choi. Toward neural-network-based large eddy simulation: Application to turbulent channel flow. *Journal of Fluid Mechanics*, 914:A16, 2021.

[138] U. Piomelli, A. Rouhi, and Bernard J. Geurts. A grid-independent length scale for large-eddy simulations. *Journal of Fluid Mechanics*, 766:499–527, 2015.

[139] Stephen B Pope. *Turbulent flows*. Cambridge University Press, 2000.

[140] Sk M Rahman, Suraj Pawar, Omer San, Adil Rasheed, and Traian Iliescu. Nonintrusive reduced order modeling framework for quasigeostrophic turbulence. *Physical Review E*, 100(5):053306, 2019.

[141] Man Mohan Rai and Parviz Moin. Direct simulations of turbulent flow using finite-difference schemes. *Journal of Computational Physics*, 96(1):15–53, 1991.

[142] Florian Rathgeber, David A Ham, Lawrence Mitchell, Michael Lange, Fabio Luporini, Andrew T T McRae, Gheorghe-Teodor Bercea, Graham R Markall, and Paul HJ Kelly. Firedrake: automating the finite element method by composing abstractions. *ACM Transactions on Mathematical Software (TOMS)*, 43(3):1–27, 2016.

[143] Valentin Resseguier, Etienne Mémin, and Bertrand Chapron. Geophysical flows under location uncertainty, part II Quasi-geostrophy and efficient ensemble spreading. *Geophysical & Astrophysical Fluid Dynamics*, 111(3):177–208, 2017.

[144] Valentin Resseguier, Etienne Mémin, and Bertrand Chapron. Geophysical flows under location uncertainty, part III SQG and frontal dynamics under strong turbulence conditions. *Geophysical & Astrophysical Fluid Dynamics*, 111(3):209–227, 2017.

[145] Valentin Resseguier, Wei Pan, and Baylor Fox-Kemper. Data-driven versus self-similar parameterizations for stochastic advection by Lie transport and location uncertainty. *Nonlinear Processes in Geophysics*, 27(2):209–234, 2020.

[146] A. Rouhi, U. Piomelli, and B.J. Geurts. Dynamic subfilter-scale stress model for large-eddy simulations. *Physical Review Fluids*, 1(4):044401, 2016.

[147] Pierre Sagaut. *Large eddy simulation for incompressible flows: an introduction*. Springer Science & Business Media, 2006.

[148] Fabrizio Sarghini, G De Felice, and Stefania Santini. Neural networks based subgrid scale modeling in large eddy simulations. *Computers & fluids*, 32(1):97–108, 2003.

[149] Chi-Wang Shu and Stanley Osher. Efficient implementation of essentially non-oscillatory shock-capturing schemes. *Journal of Computational Physics*, 77(2):439–471, 1988.

[150] Joseph Skitka, JB Marston, and Baylor Fox-Kemper. Reduced-order quasilinear model of ocean boundary-layer turbulence. *Journal of Physical Oceanography*, 50(3):537–558, 2020.

[151] Joseph Smagorinsky. General circulation experiments with the primitive equations: I. The basic experiment. *Monthly weather review*, 91(3):99–164, 1963.

[152] Chris Snyder, Thomas Bengtsson, Peter Bickel, and Jeff Anderson. Obstacles to high-dimensional particle filtering. *Monthly Weather Review*, 136(12):4629–4640, 2008.

[153] Richard JAM Stevens, Alexander Blass, Xiaojue Zhu, Roberto Verzicco, and Detlef Lohse. Turbulent thermal superstructures in Rayleigh-Bénard convection. *Physical Review Fluids*, 3(4):041501, 2018.

[154] Oliver D Street and Dan Crisan. Semi-martingale driven variational principles. *Proceedings of the Royal Society A*, 477(2247):20200957, 2021.

[155] Patrick Tabeling. Two-dimensional turbulence: a physicist approach. *Physics reports*, 362(1):1–62, 2002.

[156] Hendrik Tennekes and John L Lumley. *A first course in turbulence*. MIT press, 2018.

[157] John Thuburn, James Kent, and Nigel Wood. Cascades, backscatter and conservation in numerical models of two-dimensional turbulence. *Quarterly Journal of the Royal Meteorological Society*, 140(679):626–638, 2014.

[158] Lloyd N Trefethen and David Bau III. *Numerical linear algebra*, volume 50. Siam, 1997.

[159] Geoffrey K Vallis. *Atmospheric and oceanic fluid dynamics*. Cambridge University Press, 2017.

[160] Erwin P Van Der Poel, Rodolfo Ostilla-Mónico, John Donners, and Roberto Verzicco. A pencil distributed finite difference code for strongly turbulent wall-bounded flows. *Computers & Fluids*, 116:10–16, 2015.

[161] Erwin P. van der Poel, Richard J. A. M. Stevens, and Detlef Lohse. Comparison between two- and three-dimensional Rayleigh–Bénard convection. *Journal of Fluid Mechanics*, 736:177–194, 2013.

[162] Toby van Gastelen, Wouter Edeling, and Benjamin Sanderse. Energy-conserving neural network for turbulence closure modeling. *arXiv preprint arXiv:2301.13770*, 2023.

[163] Milo Viviani. A minimal-variable symplectic method for isospectral flows. *BIT Numerical Mathematics*, 60(3):741–758, 2020.

[164] AW Vreman. An eddy-viscosity subgrid-scale model for turbulent shear flow: Algebraic theory and applications. *Physics of Fluids*, 16(10):3670–3681, 2004.

[165] AW Vreman. The projection method for the incompressible Navier–Stokes equations: the pressure near a no-slip wall. *Journal of Computational Physics*, 263:353–374, 2014.

[166] Bert Vreman, Bernard Geurts, and Hans Kuerten. Large-eddy simulation of the turbulent mixing layer. *Journal of Fluid Mechanics*, 339:357–390, 1997.

[167] Daniel S Wilks. Effects of stochastic parametrizations in the Lorenz'96 system. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 131(606):389–407, 2005.

[168] Daniel S Wilks. *Statistical methods in the atmospheric sciences*, volume 100. Academic press, 2011.

[169] Golo Wimmer, Colin Cotter, and Werner Bauer. Energy conserving upwinded compatible finite element schemes for the rotating shallow water equations. *Journal of Computational Physics*, 401:109016, 2020.

[170] Chenyue Xie, Jianchun Wang, and E Weinan. Modeling subgrid-scale forces by spatial artificial neural networks in large eddy simulation of turbulence. *Physical Review Fluids*, 5(5):054606, 2020.

[171] V Zeitlin. Finite-mode analogs of 2D ideal hydrodynamics: Coadjoint orbits and local canonical structure. *Physica D: Nonlinear Phenomena*, 49(3):353–362, 1991.

[172] V Zeitlin. Self-consistent finite-mode approximations for the hydrodynamics of an incompressible fluid on nonrotating and rotating spheres. *Physical Review Letters*, 93(26):264501, 2004.

[173] Vladimir Zeitlin. *Geophysical fluid dynamics: understanding (almost) everything with rotating shallow water models*. Oxford University Press, 2018.

[174] Camille Zerfas, Leo G Rebholz, Michael Schneier, and Traian Iliescu. Continuous data assimilation reduced order models of fluid flow. *Computer Methods in Applied Mechanics and Engineering*, 357:112596, 2019.

[175] Xiaojue Zhu, Varghese Mathai, Richard JAM Stevens, Roberto Verzicco, and Detlef Lohse. Transition to the ultimate regime in two-dimensional Rayleigh-Bénard convection. *Physical Review Letters*, 120(14):144502, 2018.

Summary

This thesis deals with data-driven stochastic modeling of coarsened computational geophysical fluid dynamics. Geophysical fluid dynamics concerns the study of fluid flows in large-scale geophysical systems, such as the Earth's atmosphere or oceans. Physical phenomena in these flows consist of motions described by a vast range of scales. Fully resolving all these scales in a numerical simulation requires tremendous computational resources, making accurate high-resolution numerical simulations computationally expensive and time-consuming. As a result, coarse numerical simulations can be used, which employ lower resolutions to accommodate for the large computational costs. However, the use of coarse computational grids introduces uncertainty and error to the numerical solution. Uncertainty is introduced via the loss of small-scale flow features on coarse grids. That is, the small-scale flow features often cannot be determined with absolute certainty for a given large-scale flow configuration. Errors arise in the coarsening process due to poorly resolved spatial derivatives, generally leading to a deterioration of the numerical solution. To address these challenges, stochastic modeling can serve both to compensate for the errors due to coarsening and to quantify the inherent uncertainty. The aim of the work presented in this thesis is to study stochastic modeling for coarsened flows when data of the fully resolved system is available.
In this thesis, the models enter the governing equations as a space- and time-dependent forcing term. This term is decomposed into fixed spatial basis functions and corresponding time series, of which only the latter are modeled based on high-resolution numerical data. Including a forcing term of this form immediately leads to modeling choices such as where the forcing enters the governing equations, what the spatial basis functions are, how the time series are modeled, and which high-resolution data is used. The work presented in this thesis explores these aspects of modeling. More specifically, we first investigate how to measure coarsening effects exactly. These measurements are inserted into a coarse numerical simulation as a reduced-order correction with prescribed time series to get an exact agreement with a pre-computed reference solution. The quality of the reconstruction depends strongly on the employed discretization method. Subsequently, Stochastic Advection by Lie Transport is used to

quantify uncertainty due to unresolved small-scale processes in coarse numerical simulations. Here, the time series are modeled as stochastic processes with statistics matching those of the measurements. Stochastic processes with similar temporal correlation as the measurements were found to lead to a reduced ensemble spread without a loss of accuracy when compared to uncorrelated noise, indicating a closer adherence of the former to the reference solution. Finally, a data-driven subgrid-scale model is proposed. The model is derived from a data assimilation algorithm and acts on the spectral coefficients of the solution with the aim of reconstructing a reference kinetic energy spectrum. This method is found to perform well in terms of flow statistics for two different fluid dynamical systems.

Samenvatting

Dit proefschrift behandelt datagedreven stochastisch modelleren van grove computationele geofysische vloeistofdynamica. Geofysische vloeistofdynamica betreft het onderzoek naar vloeistofstromen in grootschalige geofysische systemen, zoals de atmosfeer of oceanen. Fysische verschijnselen in deze stromingen bestaan uit bewegingen omschreven door een grote hoeveelheid tijd- en ruimteschalen. Het vergt een enorme hoeveelheid rekenkracht om deze schalen tot in het kleinste detail op te lossen, waardoor voldoende nauwkeurige simulaties duur en tijdrovend zijn. Als gevolg hiervan kunnen grove rekenroosters worden gebruikt, waarin een lagere resolutie gehanteerd wordt en zodoende minder rekenkracht vereist is. Echter, het gebruik van grove rekenroosters veroorzaakt onzekerheid en fouten in de numerieke oplossing. Onzekerheid ontstaat door het verlies van kleinschalige stroombewegingen op grove roosters, wat wil zeggen dat de kleinschalige bewegingen meestentijds niet met volle zekerheid kunnen worden bepaald aan de hand van de grootschalige stroombewegingen. Fouten ontstaan door het gebruik van grove roosters wegens slecht opgeloste ruimte-afgeleides, wat doorgaans tot een verslechtering van de numerieke oplossing leidt. Stochastische modellen kunnen gebruikt worden om deze problemen te verhelpen: deze dienen zowel om te compenseren voor fouten veroorzaakt door gebruik van grove roosters als om de inherente onzekerheid te kwantificeren. Het doel van het werk wat in dit proefschrift gepresenteerd is, is het bestuderen van stochastisch modelleren voor grove vloeistofstromingen wanneer metingen van het volledig opgeloste systeem beschikbaar zijn.
In dit proefschrift worden de stochastische modellen ingevoerd in de onderliggende vergelijkingen als een ruimte- en tijdsafhankelijke forceringsterm. Deze term wordt ontbonden in vaste ruimtelijke basisfuncties en bijbehorende tijdreeksen. Alleen deze tijdreeksen worden gemodelleerd aan de hand van beschikbare nauwkeurige metingen. Het invoegen van een dergelijke forceringsterm leidt direct tot allerlei modelleringskeuzes zoals waar de forcering in de onderliggende vergelijkingen moet worden geplaatst, wat de ruimtelijke basisfuncties zijn, hoe de tijdreeksen gemodelleerd moeten worden en welke metingen gebruikt moeten worden. Deze aspecten van modelleren worden in dit proefschrift onderzocht. Allereerst wordt onderzocht hoe de effecten van het gebruik van

grove roosters exact kunnen worden gemeten. Deze metingen worden vervolgens toegepast in een grove simulatie als een zogeheten 'reduced-order' correctie met voorgeschreven tijdreeksen, zodanig dat een exacte overeenkomst met de berekende referentie-oplossing behaald wordt. De kwaliteit van deze reconstructie is sterk afhankelijk van de gebruikte numerieke methode. Vervolgens wordt stochastische advectie door Lie transport gebruikt om de onzekerheid wegens onopgeloste kleinschalige bewegingen in grove simulaties te kwantificeren. Hier zijn de tijdreeksen gemodelleerd als stochastische processen met vergelijkbare statistische eigenschappen als de metingen. Stochastische processen met vergelijkbare tijdscorrelatie als de metingen bleken te leiden tot een verminderde spreiding in het voorspellingsensemble zonder een verlies van nauwkeurigheid wanneer deze vergeleken werden met processen zonder tijdscorrelatie, wat suggereert dat de voorgaande dichter bij de referentie-oplossing ligt. Ten slotte wordt een model voor processen van subgrid-schaal voorgesteld. Dit model is afgeleid van een algoritme voor data-assimilatie en werkt op de spectrale coefficienten van de numerieke oplossing met als doel een gemeten energiespectrum te reconstrueren. Deze methode is toegepast op twee verschillende vloeistofdynamicaproblemen en levert goede resultaten op in termen van statistische grootheden.