

# Thesis title

**Thomas D. Schanzer**

Supervisor: Prof. Steven Sherwood



A thesis submitted in partial fulfilment of the requirements  
for the degree of Bachelor of Advanced Science (Honours)

School of Physics and Climate Change Research Centre  
Faculty of Science

November 2023

### **Originality statement**

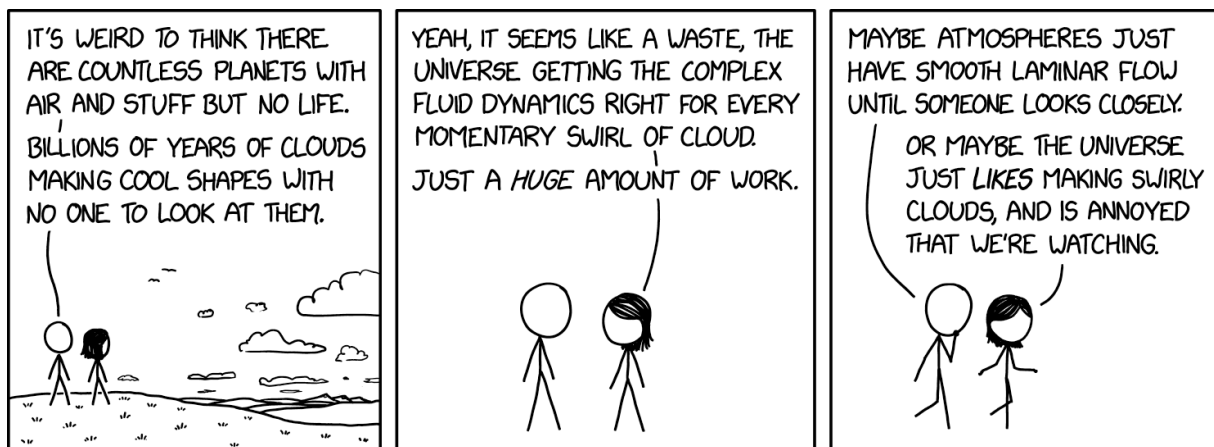
I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.

Thomas D. Schanzer  
10 November 2023

### **Data availability statement**

TODO:

# Acknowledgements



# Abstract

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Theoretical background . . . . .	1
1.1.1	The necessity of parametrisation . . . . .	1
1.1.2	Mathematical formulation of the parametrisation problem . . . . .	2
1.2	Practical background . . . . .	4
1.2.1	Physical processes requiring parametrisation . . . . .	4
1.2.2	Traditional solutions to the problem and their limitations . . . . .	4
1.3	Novel approaches to the parametrisation problem . . . . .	6
1.3.1	Stochastic parametrisation and memory . . . . .	6
1.3.2	Data-driven parametrisation and machine learning . . . . .	7
<b>2</b>	<b>Case study: the Lorenz '96 system</b>	<b>9</b>
2.1	The Lorenz '96 system . . . . .	9
2.2	Statistical models . . . . .	10
2.3	Evaluating parametrisation performance . . . . .	12
2.3.1	Offline testing . . . . .	12
2.3.2	Forecast performance . . . . .	12
2.3.3	Climate prediction performance . . . . .	13
2.4	Lessons learnt from Lorenz '96 . . . . .	14
2.5	Open questions and the case for more complex dynamical systems . . . . .	15
2.6	Beyond Lorenz '96 . . . . .	17
<b>3</b>	<b>Rayleigh-Bénard convection</b>	<b>19</b>
3.1	Governing equations . . . . .	19
3.2	Nondimensionalisation and scale analysis . . . . .	20
3.3	Hyperdiffusion terms . . . . .	21
3.4	Model configuration . . . . .	22
3.5	Choice of resolution: literature . . . . .	22
3.5.1	Resolution guidelines . . . . .	22

---

3.5.2	Metrics sensitive to under-resolution . . . . .	23
3.6	Choice of resolution: experiment . . . . .	24
<b>4</b>	<b>Calculation and modelling of subgrid tendencies</b>	<b>29</b>
4.1	Calculation of subgrid tendencies . . . . .	29
4.2	Choice of coarse-graining method . . . . .	30
4.3	Analysis of subgrid tendencies . . . . .	32
4.4	Modelling of subgrid tendencies . . . . .	32
<b>5</b>	<b>Evaluation of the parametrised model</b>	<b>33</b>
<b>6</b>	<b>Conclusion</b>	<b>35</b>
<b>A</b>	<b>Description of codebase and data</b>	<b>41</b>
A.1	Reproducibility . . . . .	41
<b>B</b>	<b>Details of numerical experiments</b>	<b>43</b>
B.1	Initial condition . . . . .	43
B.2	Model spin-up time . . . . .	43
B.3	Training dataset snapshot frequency . . . . .	44

*This page intentionally left blank*



FIRST WITCH:

When shall we three meet again  
In thunder, lightning, or in rain?

---

Shakespeare, *Macbeth*, Act I, Scene I

# Chapter 1

## Introduction

TODO: introductory paragraph with motivation (from proposal) and outline

### 1.1 Theoretical background

TODO: introductory paragraph

#### 1.1.1 The necessity of parametrisation

The primary task of general circulation models (GCMs) for Earth’s weather and climate is to simulate the dynamics of the atmosphere and ocean, which are governed by the Navier-Stokes equations. The algorithm chosen to solve these partial differential equations is known as a model’s *dynamical core* (McFarlane 2011). Since analytical solutions to the equations do not exist, the dynamical core necessarily approximates the continuous equations with finite-dimensional, exactly solvable alternatives using one of several possible discretisation schemes (e.g., the finite difference, finite element, finite volume and spectral methods) (Christensen and Zanna 2022). In practice, this usually involves representing the prognostic variables (i.e., those that affect the evolution of the flow) with sets of discrete samples in space and time, whose resolution is constrained by the available computing resources. The unavoidable consequence of discretisation is the loss of information about processes occurring on spatial and temporal scales smaller than the corresponding sampling intervals. These processes are said to be *unresolved*.

It is tempting to naïvely accept the loss of fine-scale information as a necessary sacrifice and hope the dynamical core will still make accurate predictions for the larger, resolved scales. Unfortunately, this too is impossible due to the *nonlinearity* of the governing equations. The reason for this may be seen by considering linear differential equations as a counterexample. If the governing equations were linear, they would allow arbitrary superpositions of solutions, meaning that any given solution could be partitioned into high- and low-frequency components, themselves also solutions. One would thus have the freedom to solve for the low-frequency components alone without compromise. This property may be understood more formally using the Fourier transform, defined for a function  $f$  of space and time by

$$\tilde{f}(\omega, \mathbf{k}) = \int dt d^3x e^{i(\mathbf{k} \cdot \mathbf{x} - \omega t)} f(t, \mathbf{x}),$$

which reduces any linear differential equation to an algebraic equation relating the frequency  $\omega$  to the wave vector  $\mathbf{k}$ . Each wavenumber component of the initial state propagates trivially according to its own time dependence  $e^{-i(\mathbf{k} \cdot \mathbf{x} - \omega(\mathbf{k})t)}$ , *independently of the other components*. If the equations of fluid dynamics were linear, one could safely neglect the fine-scale dynamics because they would have no influence on the coarse scales. In reality, the equations are nonlinear cannot be solved by Fourier transform.

The consequence of the nonlinearity of the equations governing atmospheric and oceanic flows is, therefore, a coupling of the resolved coarse scales to the unresolved fine scales (McFarlane 2011). This fact may be demonstrated more explicitly by applying so-called *Reynolds averaging* to the equations (Christensen and Zanna 2022). Reynolds averaging decomposes each field  $q$  into the sum of a coarse-grained (in space or time) or statistical-ensemble-averaged field  $\bar{q}$  and a perturbation  $q'$ . Note that  $\bar{q}' = 0$  by definition. The coarse-graining operation is assumed to be linear, commute with differentiation and satisfy  $\overline{\bar{p}q} = \bar{p}\bar{q}$  for any two fields  $p$  and  $q$  (Monin and Yaglom 2007). Following the example given by Christensen and Zanna (2022), consider the incompressible Navier-Stokes equations, which have the general form

$$\begin{aligned}\frac{\partial u_i}{\partial t} + u_j \frac{\partial u_i}{\partial x_j} &= \sum f_i, \\ \frac{\partial u_i}{\partial x_i} &= 0\end{aligned}$$

where  $u_i$  are the components of the velocity,  $x_i$  are the coordinates and  $f_i$  are various forces per unit mass. Summation over repeated indices is implied. Applying the decomposition and coarse-graining both sides of the equations yields

$$\begin{aligned}\frac{\partial}{\partial x_i} (\bar{u}_i + u'_i) &= 0 \\ \Rightarrow \quad \frac{\partial \bar{u}_i}{\partial x_i} + \underbrace{\frac{\partial u'_i}{\partial x_i}}_{=0} &= 0 \\ \Rightarrow \quad \frac{\partial \bar{u}_i}{\partial x_i} = \frac{\partial u'_i}{\partial x_i} &= 0\end{aligned}$$

and

$$\begin{aligned}\sum \bar{f}_i &= \overline{\frac{\partial}{\partial t} (\bar{u}_i + u'_i)} + \overline{(\bar{u}_j + u'_j) \frac{\partial}{\partial x_j} (\bar{u}_i + u'_i)} \\ &= \frac{\partial \bar{u}_i}{\partial t} + \underbrace{\frac{\partial u'_i}{\partial t}}_{=0} + \bar{u}_j \frac{\partial \bar{u}_i}{\partial x_j} + \bar{u}_j \underbrace{\frac{\partial u'_i}{\partial x_j}}_{=0} + \underbrace{u'_j}_{=0} \frac{\partial \bar{u}_i}{\partial x_j} + u'_j \frac{\partial u'_i}{\partial x_j} \\ &= \frac{\partial \bar{u}_i}{\partial t} + \bar{u}_j \frac{\partial \bar{u}_i}{\partial x_j} + \frac{\partial \bar{u}'_i u'_j}{\partial x_j} - \underbrace{u'_i \frac{\partial u'_j}{\partial x_j}}_{=0} \\ \Leftrightarrow \quad \frac{\partial \bar{u}_i}{\partial t} + \bar{u}_j \frac{\partial \bar{u}_i}{\partial x_j} &= \sum \bar{f}_i - \frac{\partial \bar{u}'_i u'_j}{\partial x_j}.\end{aligned}$$

The physical meaning of this last equation is that the evolution of the coarse-grained velocity field depends not only on itself and the coarse-grained forces, but also on the perturbations via  $\bar{u}'_i u'_j$ , which does not necessarily vanish. This dependence is a consequence of nonlinearity.

The theoretical discussion in this section establishes that physical processes occurring at one place in the spectrum of temporal and spatial scales are coupled to all the other processes in the spectrum (Franzke et al. 2015). In particular, to ignore the effect of processes not explicitly resolved in numerical models would introduce unacceptable systematic biases, or errors, in the model forecasts. GCMs therefore require parametrisation schemes to estimate the effects of unresolved processes as functions of the available large-scale information (and the parameters of these functions must be chosen appropriately—hence the name “parametrisation”).

### 1.1.2 Mathematical formulation of the parametrisation problem

In order to make any progress on the parametrization problem, it is necessary to formalise the notion of estimating the “effects” of unresolved processes. I will describe the general approach, which is well-

established (see, e.g., Hasselmann 1976; Palmer 2001; Demaeyer and Vannitsem 2018; Brajard et al. 2021), while acknowledging possible alternative conventions where they exist.

The Earth system may be considered as a dynamical system whose exact evolution is governed by equations of the form

$$\frac{dz}{dt} = \mathbf{F}(\mathbf{z}, t), \quad (1.1)$$

where the vector  $\mathbf{z}$  contains all the variables needed to fully specify the state of the system and  $\mathbf{F}$  is a nonlinear differential operator. One then employs a change of variables  $\mathbf{z} \rightarrow (\mathbf{x}, \mathbf{y})$ , where  $\mathbf{x}$  contains the information or variables that are explicitly resolved by the model being studied and  $\mathbf{y}$  contains the unresolved, ‘sub-grid’ information or variables. Mathematically,  $\mathbf{x}$  and  $\mathbf{y}$  are projections of the full state  $\mathbf{z}$  onto lower-dimensional subspaces of the original state space (Brajard et al. 2021). Practically,  $\mathbf{x}$  can be obtained by averaging the original variables over several adjacent grid points, with  $\mathbf{y}$  being the corresponding residuals (Zacharuk et al. 2018; Alcalá and Timofeyev 2021). **TODO: fix last sentence**

In principle, the change of variables splits (1.1) into two parts

$$\begin{aligned} \frac{d\mathbf{x}}{dt} &= \mathbf{G}(\mathbf{x}, \mathbf{y}, t), \\ \frac{d\mathbf{y}}{dt} &= \mathbf{H}(\mathbf{x}, \mathbf{y}, t) \end{aligned}$$

that (still exactly) describe the coupled evolution of the resolved and unresolved variables via nonlinear differential operators  $\mathbf{G}$  and  $\mathbf{H}$ . The goal of parametrisation is now to derive a new system of equations

$$\frac{d\tilde{\mathbf{x}}}{dt} = \tilde{\mathbf{G}}(\tilde{\mathbf{x}}, t) \quad (1.2)$$

whose solution  $\tilde{\mathbf{x}}(t)$  approximates the true  $\mathbf{x}(t)$  as well as possible without explicitly modelling  $\mathbf{y}(t)$ . The measure used to compare  $\tilde{\mathbf{x}}$  and  $\mathbf{x}$  depends on the intended function of the parametrised model; weather forecasting applications that prioritise short-term predictive skill might call for minimisation of the root-mean-square error (RMSE), while climate prediction would be more concerned with the closeness of their probability distributions (including both mean and extreme values).

Before proceeding, it is important to note two assumptions that are implicitly made in writing down (1.2). The first is that  $\tilde{\mathbf{G}}$  is a *deterministic* function of  $\tilde{\mathbf{x}}$ . The second is that  $\tilde{\mathbf{G}}$  depends only on the value of  $\tilde{\mathbf{x}}$  at time  $t$  and not earlier times (that is to say, the parametrisation is *memoryless*). Recent research has shown that relaxing these assumptions has the potential to greatly improve the accuracy and reliability of the parametrised model, and this will be discussed in detail in § 1.3 and Chapter 2. However, for the purpose of this basic introduction, I will temporarily follow the traditional approach and retain these assumptions.

There is more than one possible interpretation of  $\tilde{\mathbf{G}}(\mathbf{x}, t)$  and its relationship to  $\mathbf{G}(\mathbf{x}, \mathbf{y}, t)$ . Hasselmann (1976) describes  $\tilde{\mathbf{G}}(\mathbf{x}, t)$  as an ensemble average of  $\mathbf{G}(\mathbf{x}, \mathbf{y}, t)$ , taken over the distribution of all  $\mathbf{y}$  that are possible for a given  $\mathbf{x}$ , i.e.

$$\tilde{\mathbf{G}}(\mathbf{x}, t) = \langle \mathbf{G}(\mathbf{x}, \mathbf{y}, t) \rangle_{\mathbf{y}}.$$

Demaeyer and Vannitsem (2018) instead assert that the ultimate goal of parametrisation is to literally approximate  $\mathbf{y}$  as a function  $\boldsymbol{\xi}(\mathbf{x})$ , in which case

$$\tilde{\mathbf{G}}(\mathbf{x}, t) = \mathbf{G}(\mathbf{x}, \boldsymbol{\xi}(\mathbf{x}), t).$$

In practice,  $\mathbf{G}(\mathbf{x}, \mathbf{y}, t)$  usually separates into a known resolved part  $\mathbf{D}(\mathbf{x}, t)$  independent of  $\mathbf{y}$  and an unresolved coupling term  $\mathbf{C}(\mathbf{x}, \mathbf{y}, t)$ . It then suffices to approximate only the unresolved part by a parametrisation  $\mathbf{P}(\mathbf{x}, t)$  using one of the above methods, so that the parametrised model reads

$$\frac{d\tilde{\mathbf{x}}}{dt} = \mathbf{D}(\tilde{\mathbf{x}}, t) + \mathbf{P}(\tilde{\mathbf{x}}, t). \quad (1.3)$$

**TODO:  $\mathbf{x} = \langle \mathbf{z} \rangle$ , motivate data-driven method (see update presentation)**

## 1.2 Practical background

TODO: introductory paragraph

### 1.2.1 Physical processes requiring parametrisation

Following the broad theoretical argument in the previous sections, I now turn to concrete examples of processes that are often represented by parametrisations, restricting the discussion to atmospheric processes for brevity. This section will give a basic introduction to each process, its effect on the larger-scale behaviour of the climate system and the role of its corresponding parametrisation scheme. The purpose of these examples is to provide real-world context and motivation for current parametrisation research in more idealised settings, which will form one of the main topics of this review.

*Cloud microphysics* parametrisations model the composition of clouds in terms of the amount of water in the solid (e.g., hail), liquid (cloud droplets and rain) and gaseous phases, and the rate of transitions between these phases. Accurate representation of cloud formation and evolution is crucial for several reasons. First, the interaction of clouds with solar and terrestrial radiation has a major influence on the overall energy balance of the atmosphere (McFarlane 2011). Second, cloud water phase transitions lead to the formation of precipitation, and are a source and sink of latent heat that drives convection (McFarlane 2011). Furthermore, the statistics of cloud formation are linked to global temperatures in a feedback loop; uncertainty in the sign and magnitude of this feedback effect is a major contributor to uncertainty in the sensitivity of global temperatures to increases in atmospheric CO<sub>2</sub> concentration (Andrews et al. 2012; Christensen and Zanna 2022; Stevens and Bony 2013). Microphysical processes occur on the spatial scales of single water droplets and ice particles (of order  $10^{-6}$  m to  $10^{-2}$  m; Lamb (2003)), among the smallest scales in the atmosphere. All atmospheric models must parametrise the amount of cloud water in each phase, the size distribution and concentration of liquid and ice particles, and the resulting rate and type (rain, hail, snow, etc.) of precipitation (Christensen and Zanna 2022).

*Moist convection* encompasses vertical motions in the atmosphere that are accompanied (and driven) by phase changes of water, and must be parametrised when it occurs on sub-grid scales. Moist convection is generally triggered by warming and moistening at low levels, which create convective instability, and is manifested by narrow updrafts and downdrafts that can rarely be resolved explicitly (McFarlane 2011). Convection transports heat and moisture vertically, removing the instability and generating storms where the condensed water falls out as precipitation; more broadly, it is a key component of the global atmospheric circulation in spite of its small spatial scale (Christensen and Zanna 2022).

It should be noted that parametrisation also encompasses estimation techniques for processes that are too complex to model exactly for reasons unrelated to spatial and temporal resolution (McFarlane 2011). A good example of such a process is *radiative transfer*. Air and its constituent gases, as well as clouds, absorb, emit, reflect and/or scatter solar and terrestrial radiation. This leads to differential heating and cooling that drives the atmospheric circulation. While the theory of radiative transfer is understood well enough to allow precise calculations in principle, the prohibitive computational cost of such calculations necessitates a parametrisation based on simplifying assumptions (the details of which are beyond the scope of this review) (Christensen and Zanna 2022).

### 1.2.2 Traditional solutions to the problem and their limitations

With the examples of cloud microphysics, moist convection and radiative transfer in mind, I now broadly review the traditional approaches used to construct parametrisation schemes in practice. In particular, this section will identify the key assumptions upon which many traditional schemes are founded, and the circumstances under which the assumptions may be violated. This will motivate research into novel approaches.

A parametrisation scheme is traditionally constructed by formulating a simplified, easily solvable and

deterministic conceptual model of the physical process in question. The solution of this model is then used to estimate the effect of the process on the coarse-scale state of the parent model, known as the *unresolved tendency* (McFarlane 2011). For example, the earliest convective parametrisation was developed by Manabe, Smagorinsky, and Strickler (1965) and simply assumed that the net effect of convection is to relax the vertical structure of the atmosphere towards a neutrally stable state whenever it becomes convectively unstable. One obvious deficiency of simple conceptual models is that they cannot possibly capture the full range of variability in the processes they simulate. One major branch of modern parametrisation research therefore studies *data-driven* schemes that instead use observational or high-resolution simulated data to fit empirical models for the unresolved processes, naturally capturing a wider range of variability (Christensen and Zanna 2022). Data-driven parametrisation will be discussed in much further detail in the next section.

In general, a traditional parametrisation scheme aims to capture the net unresolved tendency due to all occurrences of the unresolved process (e.g., all convective updrafts and downdrafts) within each grid cell of the parent model. A deterministic prediction of this type is valid when each grid cell contains many independent realisations whose varying contributions may be expected to yield a reliable average tendency. This requires a *scale separation* between the unresolved process and the resolution of the parent model. Scale separation breaks down when model development and increases in available computing resources allow simulations at resolutions approaching the previously unresolved scales. In this case, knowledge of the coarse-scale state cannot be expected to uniquely determine the unresolved tendency because the process is only realised a few times in each grid cell. The resulting (seemingly) random nature of the true unresolved tendency motivates stochastic treatments (McFarlane 2011; Christensen and Zanna 2022; Berner et al. 2017). These will be discussed in the next section.

The conceptual models core to traditional parametrisations usually contain free parameters that are determined from the coarse-scale model state using additional assumptions called *closures*. Closures often postulate a state of quasi-equilibrium between the unresolved processes and their large-scale environment, such as a balance between the accumulation of convective available potential energy (CAPE) and its removal by convection, or between the horizontal convergence of moisture at low levels and its convective transport to higher levels (McFarlane 2011; Christensen and Zanna 2022; Palmer 2019). However, there is no guarantee that such an equilibrium exists; in fact, it has been demonstrated that the CAPE balance is violated by fluctuations on sub-diurnal time scales (Donner and Phillips 2003) and by midlatitude continental convection (Zhang 2002). Newer parametrisation schemes have allowed departures from equilibrium by representing the unresolved processes in a prognostic rather than diagnostic manner (i.e., allowing the processes to have their own self-governing time dependence rather than calculating them from the large-scale state at each time step independently of their values at the previous step) (Rio, Del Genio, and Hourdin 2019; Berner et al. 2017). This creates *memory* or *latency* in the parametrised tendencies, meaning that the tendencies have some nonzero response time when subjected to sudden changes in the large-scale state. Memory will be discussed further in the next section.

Parametrisation schemes commonly suffer from several other issues that I will briefly address here. Firstly, while the division of the general atmospheric dynamics into a set of separately parametrised processes (microphysics, convection, etc.) is physically motivated, it remains somewhat arbitrary because these processes, strictly speaking, form a continuum without well-defined boundaries (Christensen and Zanna 2022; McFarlane 2011). It is a goal of contemporary research to unify the parametrisations as much as possible. Secondly, given the importance of future climate projections, it is natural to ask whether the parametrisation schemes that have been developed and tuned on today's climate remain valid as the climate changes over decade- to century-long simulations. This is a matter of particular concern for data-driven parametrisations, since there is little reason to trust empirical models once they are extrapolated beyond the range of the data originally used to fit them (Christensen and Zanna 2022). Finally, unless very special care is taken, parametrisation schemes can cause the parent model to violate known physical conservation laws (e.g., mass, energy and momentum) (Christensen and Zanna 2022). Efforts to resolve this issue are ongoing.

## 1.3 Novel approaches to the parametrisation problem

Since the 1990s, the limitations identified in § 1.2.2 have prompted many to reconsider the principles upon which parametrisation schemes are founded. The main advance has been the development of stochastic parametrisations that incorporate random noise in the predicted tendencies. More recently, others have approached the problem from an entirely new direction, developing data-driven parametrisations that learn to predict the tendencies empirically. This section will introduce the stochastic and data-driven approaches in general terms while omitting the technical details of individual implementations, the aim being to contextualise and motivate the study of these approaches in more idealised frameworks in Chapter 2. Other methods (such as superparametrisation) exist but are beyond the scope of this review.

### 1.3.1 Stochastic parametrisation and memory

The potential value of stochasticity for climate modelling was first established by Hasselmann (1976), whose seminal paper sought to explain the characteristics of long-term climate variability. Knowing that climate depends on interactions between all components of the Earth system (atmosphere, ocean, cryosphere, biosphere, etc.), Hasselmann argued that the effect of the more rapidly-evolving atmosphere on the other, more slowly-varying components is that of a stochastic forcing. Owing to their long response time, the slowly-varying components effectively integrate this stochastic forcing, allowing long-term climate variability to be characterised as a type of random walk process akin to the Brownian motion of a massive particle in a fluid.

Further motivation for stochastic parametrisation in particular stems from the need to reliably estimate uncertainties in weather forecasts and climate projections. Weather forecasting centres typically propagate initial condition uncertainties (due to imperfect observations) through to the final forecast in a Monte Carlo fashion, by initialising an ensemble of model runs with perturbed initial conditions and measuring the spread of the resulting forecasts. It has been observed that deterministic models produce systematically under-dispersed ensembles that fail to span the range of actual weather outcomes, indicating that these models are failing to capture additional sources of variability (Palmer et al. 2005; Berner et al. 2017; Palmer 2019). Knowing that poor scale separation and departures from quasi-equilibrium should preclude deterministic relationships between unresolved tendencies and the large-scale state (see § 1.2.2), it should seem highly likely that deterministic parametrisation contributes to this deficiency.

The principle of stochastic parametrisation is, therefore, that unresolved tendencies should be randomly sampled from an appropriate distribution at each point in space and time, not simply set to the mean of the distribution (Franzke et al. 2015). This choice has now been theoretically justified using statistical mechanical arguments; most notably, Wouters and Lucarini (2012, 2013) showed that, assuming a weak coupling to the resolved variables, the effect of unresolved dynamics should be parametrised by a combination of deterministic and stochastic terms, as well as a *non-Markovian* memory term depending on the past states of the resolved variables.

The simplest and earliest approach to stochastic parametrisation is the method of *stochastically perturbed parametrisation tendencies* (SPPT), which takes an existing deterministic parametrisation and randomly scales its output with a multiplicative noise field (Palmer 2019; Christensen 2020). Re-using the notation of § 1.1.2, an SPPT model for the resolved variables  $\mathbf{x}$  takes the form

$$\frac{d\mathbf{x}}{dt} = \mathbf{D}(\mathbf{x}, t) + [I + \text{diag}(\mathbf{e})]\mathbf{P}(\mathbf{x}, t),$$

where  $\mathbf{P}$  is the existing deterministic parametrisation,  $\mathbf{e}$  is a mean-zero random vector with the same length as  $\mathbf{x}$  and  $I$  is the identity matrix. The choice of multiplicative (and thus inherently state-dependent) rather than additive noise is intuitively motivated by the expectation that variability in the effect of unresolved processes should be greatest when those processes are most active (Franzke et al. 2015). Christensen (2020) performed a comparison of high-resolution simulations to parametrised single-column model output which justified the use of multiplicative perturbations.



In GCMs, where the variables being modelled have spatial dependence, the multiplicative perturbation takes the form of a random *field*  $e(x, y, t)$ . It has been argued that this random field should be spatially and temporally correlated (in contrast to uncorrelated “white” noise) in order to emulate the organisation of unresolved processes on larger scales and their persistence in time (Christensen and Zanna 2022; Franzke et al. 2015). In particular, perturbations that explicitly depend on their own past values constitute a type of memory, albeit distinct from the memory term advocated by Wouters and Lucarini (2012, 2013), which would instead couple the perturbations to the past values of the large-scale variables.

Stochastic parametrisations have several known advantages over their deterministic counterparts, and are now implemented in operational weather forecast models. They have been found to remedy the aforementioned issues of ensemble underdispersion and prediction unreliability (Palmer et al. 2005; Berner et al. 2017), and (as of 2009) even make the skill of five-day forecasts comparable to that of deterministically parametrised two-day forecasts (Palmer 2019). Despite the zero-mean nature of the noise, it has been shown that stochastic parametrisations can reduce systematic model biases (“noise-induced drift”; Palmer et al. (2005)) and stabilise the simulation of regime-based behaviour (such as the El Niño–Southern Oscillation) in the climate system (Berner et al. 2017).

Reviewing the field, Palmer (2019) identifies outstanding issues for further research. The main concern is the lack of rigour in most stochastic schemes; SPPT, for example, modifies existing deterministic schemes *ad hoc* rather than incorporating stochasticity *ab initio* in the development process. More objective approaches are yet to gain widespread acceptance. This motivates both data-driven methods and further testing using simpler dynamical systems where objectivity is more feasible.

### 1.3.2 Data-driven parametrisation and machine learning

The fitting of predictive statistical models to data is, of course, ubiquitous in the sciences, but attempts to use such models as parametrisation schemes and couple them into fully fledged GCMs are a relatively new phenomenon—certainly more so than stochastic parametrisations. The prerequisite for all data-driven parametrisations is obviously training data: mathematically, an approximate solution of (1.1) of sufficient accuracy and resolution to be considered “truth” for the application at hand. Training data are typically derived from high-resolution simulations, such as regional weather model runs or large eddy simulations (LES). The other ingredient is the imperfect low-resolution model that will later be augmented by parametrisation.

To generate an appreciation of how these data might be used to construct a parametrisation in practice, I roughly follow the argument and notation of Brajard et al. (2021). Define a map  $\mathcal{M}$  that takes each state  $\mathbf{z}(t)$  in the training dataset to the state  $\mathbf{z}(t + \delta t)$  at the next time step. Similarly, denote by  $\mathcal{M}^r$  the low-resolution model (the superscript *r* meaning “reduced”), which maps a low-resolution state  $\mathbf{x}(t)$  to a prediction for  $\mathbf{x}(t + \Delta t)$ , where  $\Delta t$  is not necessarily equal to  $\delta t$  (usually larger). The spaces of high- and low-resolution states, having different dimensions, may be linked by an operator  $\langle \cdot \rangle$  that projects high-resolution states onto the low-resolution state space. Practically, the projection operation is simply a coarse-graining of  $\mathbf{z}$  by averaging values at adjacent grid points to match the resolution of  $\mathbf{x}$  **TODO: not necessarily!**. Now, for each state  $\mathbf{z}$  in the training dataset, one may compute the difference

$$\epsilon(\mathbf{z}) = \frac{\langle \mathcal{M}(\mathbf{z}) \rangle - \langle \mathbf{z} \rangle}{\delta t} - \frac{\mathcal{M}^r(\langle \mathbf{z} \rangle) - \langle \mathbf{z} \rangle}{\Delta t}$$

between the coarse-grained true tendency (on the left) and the tendency predicted by the coarse model when it sees the same state (on the right). This is the unresolved tendency. Knowledge of  $\langle \mathbf{z} \rangle$  alone does not uniquely determine  $\epsilon$  because  $\langle \mathcal{M}(\mathbf{z}) \rangle$  depends on the original  $\mathbf{z}$ . However, if one can fit a statistical model  $\mathbf{P}$  to the dataset of  $(\langle \mathbf{z} \rangle, \epsilon)$ , then  $-\mathbf{P}(\mathbf{x}(t))\Delta t$  will be able to serve as an estimate of the error incurred by the coarse model in predicting  $\mathbf{x}(t + \Delta t)$  given  $\mathbf{x}(t)$ . This motivates the construction of a parametrised model  $\mathcal{M}^p$ , defined by

$$\mathcal{M}^p(\mathbf{x}) = \mathcal{M}^r(\mathbf{x}) + \mathbf{P}(\mathbf{x})\Delta t$$

that simply subtracts the estimated error at each time step of the coarse model. In theory, the solution  $\mathbf{x}(t)$  obtained by iteration of  $\mathcal{M}^P$  will be more accurate than that obtained by iteration of the unparametrised  $\mathcal{M}^r$ .

The current method of choice for constructing  $\mathbf{P}$  in weather and climate modelling contexts is machine learning (ML). I will briefly introduce the key concepts, strengths and weaknesses of ML, following Beucler et al. (2022). ML encompasses a broad class of algorithms that programmatically and autonomously develop and apply rules for performing tasks, such as image classification, removing the need for a human to explicitly program the rules. They do this using supplied *training data* that exemplify the task to be performed. When it comes to data-driven parametrisation, the task is function approximation or regression, and the relevant algorithms usually require a training dataset of example inputs (i.e., coarse-scale variables) *labelled* with the desired outputs (i.e., unresolved tendencies). One commonly-used class of algorithms called *neural networks* consist of layers of inter-communicating calculation nodes called neurons and learn to approximate functions by optimising sets of weights associated with the neurons. Another type of algorithm is the *random forest*, which approximates functions by passing the inputs through a series of nested `if/else` decision trees.

The distinguishing advantage offered by ML data-driven parametrisations is their ability to systematically capture complex relationships without relying on artificially simplified physical models derived by humans (Irrgang et al. 2021; Beucler et al. 2022). In addition to constructing new parametrisations from scratch, ML has been used to emulate the action of existing parametrisation schemes in GCMs (e.g., Gentine et al. 2018). These ML emulators have the potential to reduce the computational burden that parametrisation usually imposes, leaving more resources for other tasks such as calling the parametrisation scheme more frequently or increasing the model resolution (Beucler et al. 2022). Furthermore, the ability to emulate multiple existing schemes at once is a step towards unifying the parametrisation of unresolved processes (a long-standing issue; see § 1.2.2).

ML parametrisations are not without their issues; they are known to be prone to overfitting and often cause their host models to become numerically unstable. They are also, by their very nature, far more difficult to interpret than traditional schemes. It is therefore very difficult to determine *a priori* whether or not they will continue to be valid in climate conditions outside the bounds of their training datasets (Irrgang et al. 2021; Beucler et al. 2022). “Interpretable” or “explainable” ML is an active area of research that aims to address these issues. Another active research topic is “physics-guided ML”, the aim of which is to augment ML with known physical constraints and conservation laws that would otherwise be violated (e.g., Yuval, O’Gorman, and Hill 2021).

TODO: summary?



## Chapter 2

# Case study: the Lorenz '96 system

**TODO: Fix this paragraph** Historically, it has been common practice to develop and test new parametrisation techniques using simpler dynamical systems that share the key nonlinear, multi-scale and chaotic properties of the climate system. These analogue systems provide testbeds where the parametrisation problem is more tractable, avoiding the expense, effort and technicalities of fully-fledged climate models. The simplicity of these systems also helps to ensure the reproducibility of the results obtained. This section will review the progress that has been made using simple analogue systems and argue that, considering the outstanding issues identified in §§ 1.2 and 1.3, further research using these systems may yet be warranted. The relevant techniques for constructing and testing parametrisation schemes in simple model frameworks, and the remaining open questions, will be discussed in order to inform future research.

### 2.1 The Lorenz '96 system

The Lorenz '96 model (henceforth L96) was proposed by eminent meteorologist Edward N. Lorenz in a 1995 paper (Lorenz 1995) on the growth of errors in solutions of dynamical systems and the resulting limits of predictability for those systems. Lorenz proposed to mimic the multi-scale nature of the climate system by coupling two dynamical systems with different characteristic time scales. Following his notation, there is a discrete set of “slow” variables  $\{X_k\}_{k=1}^K$ , each of which has an associated discrete set of “fast” variables  $\{Y_{j,k}\}_{j=1}^J$ . The indices of the slow variables are periodic ( $X_{K+1} = X_1$  and  $X_0 = X_K$ ) and the sets of fast variables are arranged end-to-end so that  $Y_{J+1,k} = Y_{1,k+1}$  and  $Y_{0,k} = Y_{J,k-1}$ . The intuition is that the variables represent samples of a field along a latitude circle, as shown in Figure 2.1 (a reproduction of Figure 1 by Russell et al. (2017)). The evolution of the system is governed by the ordinary differential equations

$$\frac{dX_k}{dt} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F - \frac{hc}{b} \sum_{j=1}^J Y_{j,k}, \quad (2.1a)$$

$$\frac{dY_{j,k}}{dt} = -cY_{j+1,k}(Y_{j+2,k} - Y_{j-1,k}) - cY_{j,k} + \frac{hc}{b} X_k. \quad (2.1b)$$

The constant  $c \geq 1$  dictates the ratio of the time scale of the  $Y$  variables to the time scale of the  $X$  variables,  $b$  the typical magnitude of  $X$  relative to  $Y$ ,  $h$  the strength of the coupling between the two subsystems and  $F$  the constant forcing applied to each  $X$  variable.

As discussed in § 1.1.2, the objective is to find  $P_k(X_1, \dots, X_K) \approx -(hc/b) \sum_{j=1}^J Y_{j,k}$  such that the solution of the parametrised system

$$\frac{dX_k}{dt} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F + P_k(X_1, \dots, X_K)$$

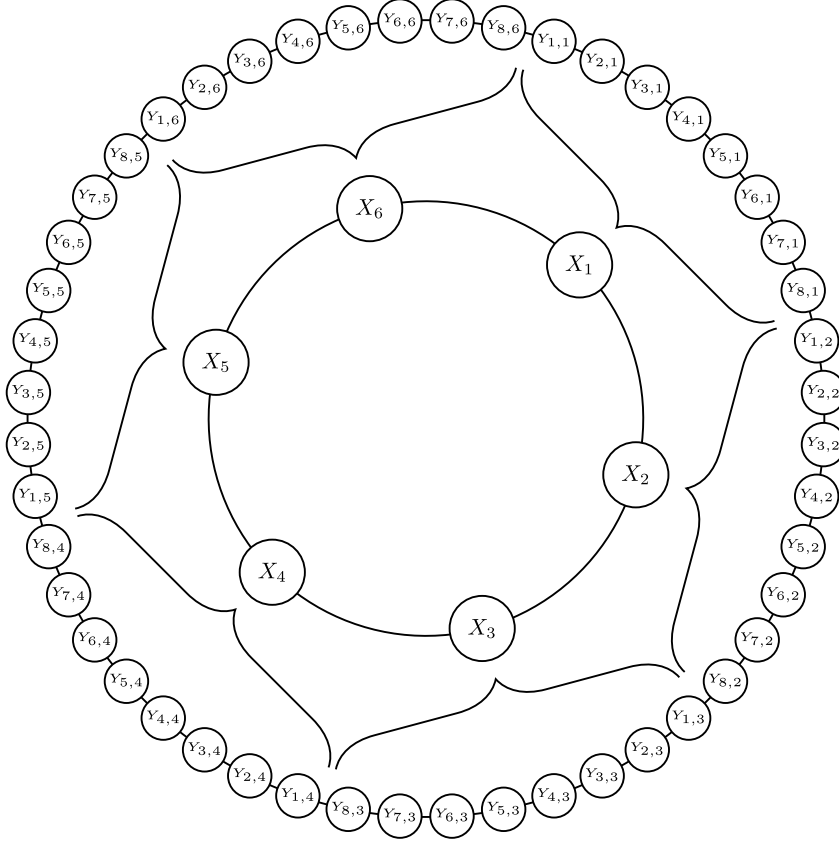


Figure 2.1: Illustration of the periodic arrangement of the L96 variables, with the slow variables  $X_k$  arranged in a circle. Each slow variable is coupled to a neighbouring subset of the similarly arranged fast variables  $Y_{j,k}$ . Reproduced from Russell et al. (2017), Figure 1.

approximates as accurately as possible the true  $X_k$  obtained by solving (2.1). The simple form of L96 allows one to generate a training dataset by numerically solving (2.1) and directly diagnosing the unresolved tendency of  $X_k$  as

$$U_k(t) = \frac{X_k(t + \Delta t) - X_k(t)}{\Delta t} - [-X_{k-1}(X_{k-2} - X_{k+1}) - X_k]. \quad (2.2)$$

As we shall see, it is usually assumed that  $P_k$  depends only on  $X_k$  (i.e., the parametrisation is *local*) but may also depend on the value of  $X_k$  at earlier times. The symmetry of (2.1) under shifts of the  $k$  index implies that all the  $X_k$  have the same long-term statistics, so it suffices to aggregate all the  $U_k$  into a single training dataset and use the resulting parametrisation for all the  $X_k$  rather than constructing a separate parametrisation for each variable.

## 2.2 Statistical models

Constructing a data-driven parametrisation scheme for L96 requires first choosing the structure of the scheme and then using the training dataset to find the optimal values of the associated free parameters. The simplest structure, popularised in an influential paper by Wilks (2005), consists of a polynomial regression of  $U$  against  $X$  as a deterministic base, modified by stochastic noise. Wilks' scatterplot of  $U$  and  $X$  and quartic polynomial least-squares regression, shown in Figure 2.2, confirm that such a structure is reasonable.

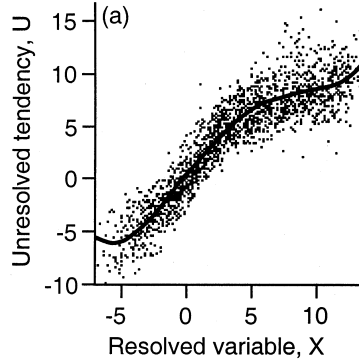


Figure 2.2: Scatterplot of unresolved tendencies  $U$  against large-scale variables  $X$  in L96 (black dots), and corresponding quartic polynomial least-squares regression (black line). Data are for forcing  $F = 18$ . Reproduced from Figure 2a of Wilks (2005).

Denote the deterministic polynomial part of the parametrisation by  $P_{\text{det}}(X)$ . Wilks' proposal was to then model the residuals as an additive, mean-zero noise term  $e(t)$ , independent of  $X$  and updated at each time step, so that

$$P(X) = P_{\text{det}}(X) + e(t). \quad (2.3)$$

Arnold, Moroz, and Palmer (2013) and later Christensen, Moroz, and Palmer (2015) considered the SPPT method (see § 1.3.1), where instead

$$P(X) = [1 + e(t)]P_{\text{det}}(X). \quad (2.4)$$

The most common form for  $e(t)$  in both (2.3) and (2.4) is that of a *first-order autoregressive* or AR(1) model

$$e(t) = \phi e(t - \Delta t) + \sigma z \quad (2.5)$$

where  $\phi \in [0, 1]$  and  $\sigma \geq 0$  are constants and  $z$  is drawn independently from the standard normal distribution at each time step. The AR(1) model introduces memory by having  $e(t)$  relax from its value at the previous time step towards zero, while also adding independent random jumps with standard deviation  $\sigma$ . AR(1) noise is commonly also referred to as *red* noise. The two important special cases of an AR(1) model are  $\phi = 0$ , which reduces  $e(t)$  to white noise without memory, and  $\phi = \sigma = 0$ , which results in a deterministic parametrisation  $P(X) = P_{\text{det}}(X)$ . The best estimates of  $\phi$  and  $\sigma$  are straightforwardly obtained by examining the autocorrelation and standard deviation of the residual time series  $U(t) - P_{\text{det}}(X(t))$  in the training dataset (see Arnold, Moroz, and Palmer (2013) and Wilks (2011, Chapter 9) for details).

It must be noted that the additive and SPPT schemes (2.3) and (2.4) are inherently limited by their simple form. The additive scheme assumes that the variance of the unresolved tendency is independent of the value of  $X$ , and SPPT implies that the variance vanishes when  $P_{\text{det}}(X) = 0$ . Inspection of Figure 2.2 indicates that neither of these conditions is strictly satisfied. Arnold, Moroz, and Palmer (2013) therefore proposed two possible modifications of (2.3), with the standard deviation of  $e(t)$  being a linear function of either  $|X(t)|$  or  $|P_{\text{det}}(X(t))|$ .

Other studies have experimented with more complex statistical models for the unresolved tendencies. Chorin and Lu (2015) tested NARMAX (nonlinear autoregression moving average with exogenous inputs), a model which represents the tendency as a function of (i) the tendencies estimated at previous time steps (autoregression), (ii) the current and previous values of  $X$  (nonlinear; exogenous data), (iii) independent Gaussian noise and (iv) the previous values of the Gaussian noise (moving average). The motivation for NARMAX is that it may be able to capture more complex relationships and memory effects by using more predictors and free parameters. Crommelin and Vanden-Eijnden (2008) and Kwasniok (2012) used Markov chain models, which approximate the continuous range of possible unresolved tendencies  $U$  for a given  $X$  by a discrete set of allowed states. The model transitions from state to state according

to a set of probabilities that depend on  $X$  and are estimated from the training dataset. The latest studies have tested machine learning algorithms for L96, reflecting the increasing research interest in ML-based parametrisation schemes for GCMs (see § 1.3.2). Gagne II et al. (2020) used generative adversarial networks (GANs), which involve pairs of competing neural networks and have the advantage of being stochastic with no need for *ad hoc* perturbations. Bhouri and Gentine (2023) used deterministic but memory-based neural networks trained to directly optimise short-term forecast accuracy without requiring the calculation of the unresolved tendencies  $U$  in the training dataset.

## 2.3 Evaluating parametrisation performance

After choosing and fitting a parametrisation scheme, the next step is to assess its performance. It is important to recognise that there are several contexts in which one might want a parametrisation to perform well. The first important distinction is between *offline* and *online* testing. Offline testing involves feeding large-scale states into the parametrisation scheme from a pre-computed test dataset that was not used for training (such as a new high-resolution simulation) and measuring the level of agreement between the tendencies predicted by the scheme and the true tendencies. Online testing, on the other hand, involves coupling the parametrisation scheme into the host model and comparing the output of the parametrised model to the corresponding “truth” solution (which might be the output of a high-resolution unparametrised model). Online performance can further be assessed for short-term (“weather”) forecasts or long-term (“climate”) predictions. The reliability of ensemble forecasts (see § 1.3.1) must also be determined. As we shall see, the extent to which good performance in more than one category is achievable is a topic of ongoing research.

### 2.3.1 Offline testing

Offline testing is rarely documented in the L96 literature because it is somewhat more straightforward than online testing. The aim is to ensure that the parametrisation scheme accurately captures the distribution of unresolved tendencies that exist in the true solution. Gagne II et al. (2020) achieved this by integrating the full system (2.1) to obtain a test dataset and estimating the time-aggregated probability density functions (PDFs) of the true unresolved tendencies (2.2) and the tendencies predicted by the parametrisation schemes. They then expressed the difference between these distributions as a scalar quantity using the Hellinger distance, defined for two PDFs  $p$  and  $q$  by

$$H(p, q) = \frac{1}{2} \int dx \left( \sqrt{p(x)} - \sqrt{q(x)} \right)^2. \quad (2.6)$$

### 2.3.2 Forecast performance

When evaluating online forecast accuracy for L96, studies are primarily concerned with the agreement between a “truth” (or verification) solution of (2.1) and the mean of an ensemble of independently realised, stochastically parametrised forecasts starting from the same or nearby initial conditions. This reflects standard practice in operational numerical weather prediction. A standard approach is to calculate the root mean square error (RMSE)

$$\text{RMSE}(t) = \left( \left\langle \sum_k (X_k^{\text{ens}}(t) - X_k^{\text{ver}}(t))^2 \right\rangle \right)^{1/2}$$

between the ensemble mean forecast  $X_k^{\text{ens}}(t)$  and the verification  $X_k^{\text{ver}}(t)$ , where the mean  $\langle \cdot \rangle$  is taken over many forecast-verification pairs with a range of initial conditions (Crommelin and Vanden-Eijnden 2008; Gagne II et al. 2020).

An alternative for single-integration (non-ensemble) forecasts is to take the mean over time, from the initial time up to time  $t$  (Bhouri and Gentine 2023). Another accuracy metric is the anomaly correlation (ANCR), which is the Pearson correlation coefficient

$$\text{ANCR}(t) = \left\langle \frac{\sum_k A_k^{\text{ens}}(t) A_k^{\text{ver}}(t)}{[(\sum_k A_k^{\text{ens}}(t)^2)(\sum_k A_k^{\text{ver}}(t)^2)]^{1/2}} \right\rangle$$

between the forecast anomaly  $A_k^{\text{ens}}(t) = X_k^{\text{ens}}(t) - \langle X_k^{\text{ver}}(t) \rangle_t$  and the verification anomaly  $A_k^{\text{ver}}(t) = X_k^{\text{ver}}(t) - \langle X_k^{\text{ver}}(t) \rangle_t$ , again averaged over forecast-verification pairs (Crommelin and Vanden-Eijnden 2008). Other studies (Kwasniok 2012; Arnold, Moroz, and Palmer 2013) have used more complex skill scores to assess forecast accuracy.

Recall from § 1.3.1 that an ensemble forecast is said to be reliable if the spread of the ensemble members is representative of the range of weather states that could possibly be realised, given the initial condition uncertainty and model errors. In the context of L96, there are two ways to measure reliability. The first (Wilks 2005; Crommelin and Vanden-Eijnden 2008; Kwasniok 2012) is to sort the states  $X_k$  of the ensemble members at the chosen lead time from largest to smallest (or vice versa), and find the rank that the truth state would occupy within the list. The ranks from many ensemble forecasts are aggregated and their distribution plotted on a *rank histogram*, which should ideally be uniform for a well-dispersed ensemble. A U-shaped rank histogram indicates underdispersion; that is, the truth too frequently lies at the extreme ends of or outside the ensemble. An inverted U shape indicates the opposite (overdispersion). Another method is to simply compare the ensemble standard deviation and the RMSE, which should be roughly equal for a reliable forecast (Arnold, Moroz, and Palmer 2013; Gagne II et al. 2020).

### 2.3.3 Climate prediction performance

There are several classes of metrics and methods available for assessing the accuracy of long-term climate predictions made by parametrised L96 models. The simplest is to compare the low-order moments (mean, variance) of the long-term distributions of  $X$  between the parametrised and truth solutions. However, it is often the case for the relatively simple L96 model that parametrised models reproduce these moments quite accurately (Wilks 2005), motivating comparison of the entire PDFs—a more stringent test. In addition to visual comparison of the PDFs, one can use scalar measures such as the Hellinger distance (2.6) (Arnold, Moroz, and Palmer 2013; Gagne II et al. 2020) and the Kolmogorov-Smirnov statistic

$$D_n = \max_X |\Psi_{\text{true}}(X) - \Psi_{\text{model}}(X)|,$$

which is the maximum absolute difference between the cumulative distribution functions  $\Psi$  of  $X$  in the truth and parametrised models (Wilks 2005; Chorin and Lu 2015; Kwasniok 2012). As a diagnostic tool, it is also possible to compare the joint distributions of  $X$  and the unresolved tendency  $U$  between the truth and parametrised models (Gagne II et al. 2020).

Another approach is to study the correlation and spectral properties of  $X$ . In the time domain, one can compute the autocorrelation function  $\langle X_k(t) X_k(t + \tau) \rangle_{t,k}$  or the cross-correlation  $\langle X_k(t) X_{k+1}(t + \tau) \rangle_{t,k}$  between neighbouring variables and compare these to their truth counterparts (Crommelin and Vanden-Eijnden 2008; Kwasniok 2012; Chorin and Lu 2015; Gagne II et al. 2020). Gagne II et al. (2020) applied the continuous wavelet transform (CWT) to the  $X(t)$  time series to measure the amount of power in the signal as a function of oscillation period. Similar analyses may be performed in “space” by computing the autocorrelation  $\langle X_k(t) X_{k+n}(t) \rangle_{t,k}$  with respect to shifts in the  $k$  index (Gagne II et al. 2020) or by taking a discrete Fourier transform of  $X_k$  with respect to  $k$  and deriving wavenumber statistics (Crommelin and Vanden-Eijnden 2008; Kwasniok 2012).

## 2.4 Lessons learnt from Lorenz '96

Now that the statistical models and analysis techniques for L96 parametrisation have been surveyed, I shall review the results of previous studies, focusing on key conclusions that have potential implications for more complex dynamical systems. The common objective in many studies has been to determine the value added by stochasticity and memory. The general consensus is that parametrisations with white noise outperform purely deterministic ones, and that adding memory via red noise or dependence on past large-scale states gives a further advantage over white noise (Wilks 2005; Arnold, Moroz, and Palmer 2013; Gagne II et al. 2020). The advantages of stochasticity and memory include improvements in the skill and reliability of ensemble forecasts and in the accuracy of climatological probability distributions. Arnold, Moroz, and Palmer (2013) additionally found that parametrisations with state-dependent noise have the potential to produce more skilful forecasts than those with state-independent additive noise. This was only observed under conditions of small time-scale separation between the resolved and unresolved variables (i.e.,  $c \sim 1$  in (2.1)); it appeared that simple additive noise was sufficient for larger scale separation. A representative illustration of the conclusions for short-term forecasting and climate prediction are respectively given by Figures 6 and 10 of Arnold, Moroz, and Palmer, which are reproduced here as Figures 2.3 and 2.4.

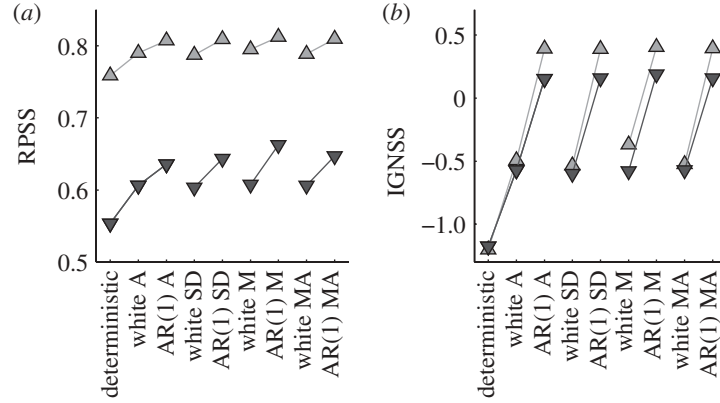


Figure 2.3: Reproduction of Figure 6 by Arnold, Moroz, and Palmer (2013), showing a comparison of ensemble forecast skill for parametrisations with different types of noise. Skill is measured by the ranked probability skill score (RPSS; panel (a)) and ignorance skill score (IGSS; panel (b)); higher values indicate better skill. The deterministic parametrisation is a cubic polynomial regression and added noise is either white or generated by an AR(1) process. Acronyms indicate the method used to introduce the noise; A = “additive”, SD = “state-dependent” (noise standard deviation a linear function of  $|X|$ ), M = “multiplicative” (SPPT) and MA = “multiplicative and additive” (noise standard deviation a linear function of  $|P_{\text{det}}(X)|$ ). Upright light grey triangles indicate results obtained under conditions of larger time scale separation ( $c = 10$  in (2.1)) and inverted dark grey triangles indicate smaller time scale separation ( $c = 4$ ). It is clear that schemes with white noise outperform deterministic ones, and schemes with AR(1) noise outperform those with white noise.

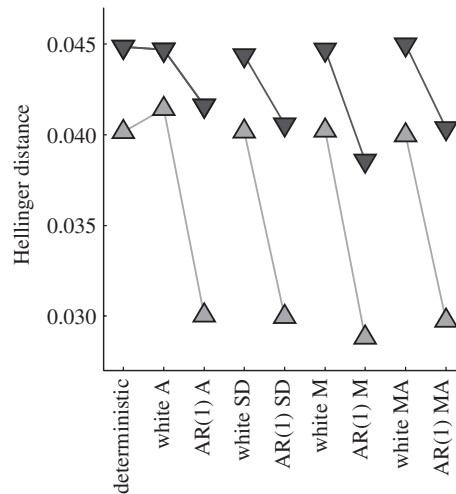


Figure 2.4: Reproduction of Figure 10 by Arnold, Moroz, and Palmer (2013), comparing the accuracy with which parametrisations with different types of noise reproduce the true climatological PDFs of the resolved variables. Accuracy is measured by the Hellinger distance between the modelled and true PDFs; lower values indicate higher accuracy. The labels on the horizontal axis and the markers have the same meaning that they do in Figure 2.3. A clear improvement in accuracy is given by the schemes with AR(1) noise.

Some studies have obtained results that conflict with the aforementioned consensus. Firstly Arnold, Moroz, and Palmer (2013) found no appreciable difference in climatological PDF accuracy between their white noise and deterministic schemes; this is evident in Figure 2.4. Analogously, Christensen, Moroz, and Palmer (2015) similarly found no significant difference between white noise and deterministic schemes in terms of their ability to reproduce regime behaviour and predict the system’s response to changes in the external forcing parameter  $F$ . Gagne II et al. (2020), who constructed parametrisations using generative adversarial networks, came to the unexpected conclusion that using temporally correlated red noise may degrade the accuracy of ensemble forecasts via an accumulation of noise-induced error and have little benefit over white noise for climate prediction accuracy.

Several authors have postulated general interpretations of these results that may be relevant to the parametrisation problem in dynamical systems beyond L96. Arnold, Moroz, and Palmer (2013) attributed the apparent superiority of memory-based parametrisations to a potential for subgrid-scale processes to influence the resolved state on scales much larger than the grid scale at which the host model is truncated; memory captures this spatial and temporal correlation or persistence. Christensen, Moroz, and Palmer (2015) agreed with this assessment on the basis of their results. They also argued that the persistence of correlated noise is necessary to drive regime transitions that would otherwise be underrepresented. Observing that stochastic parametrisation improved the prediction of responses to changes in forcing, they further argued that stochastic data-driven parametrisations are generally more likely to function well in changed climate conditions outside the range of their training datasets.

## 2.5 Open questions and the case for more complex dynamical systems

To conclude this section on data-driven parametrisation for L96, I will draw on the key messages and conclusions in the literature to motivate further research of the same kind in models of intermediate complexity.

Generally speaking, many parametrisation approaches of varying complexity have shown promise when



applied to L96; all the studies mentioned in § 2.2 were able to construct a scheme that reproduced the properties of interest without egregious errors. Indeed, some of the more recent and complex approaches have shown only marginal improvements over the original regression-based approach introduced by Wilks (2005). Gagne II et al. (2020) attributed this to the simplicity of L96, which only considers the evolution of a single field in one discrete spatial dimension. There is hope that applying the L96 methods to dynamical systems with more prognostic variables, spatial dimensions and degrees of freedom may reveal deficiencies that were inconsequential in L96 and thus give researchers more discriminative power in choosing approaches for application to full GCMs.

Several specific findings in L96 studies raise other questions that also motivate a change of focus to more complex dynamical systems. The first is the observation by Crommelin and Vanden-Eijnden (2008) that the advantage of their conditional Markov chain parametrisation over the regression-based approach was greatest under conditions of small time-scale separation and diminished with increasing scale separation. Arnold, Moroz, and Palmer (2013) drew a similar conclusion regarding the difference between stochastic and deterministic regression-based schemes. These findings are consistent with the expectation that the net effect of subgrid-scale processes should become increasingly unpredictable as resolution increases and the size of the sample of small-scale events decreases (see § 1.2.2). This is another piece of evidence that more realistic fluid models, which do not have such a clean separation between “resolved” and “unresolved” variables, will be more discriminating and allow the best approaches to show their full potential.

An open question of importance for real-world modelling concerns the extent to which parametrisation schemes that result in accurate weather forecasts tend can be expected to also produce accurate climate predictions (Christensen and Berner 2019). Another asks whether a scheme’s offline performance is a good indicator of its online performance once coupled into a model. Gagne II et al. (2020) addressed these questions by testing a collection of GAN parametrisations with different predictors and types of noise (see Figure 2.5), with the results suggesting an affirmative answer for the former and a negative answer for the latter. The verification of these conclusions would be a worthwhile goal for further research using other dynamical systems. Other possible goals could be to more clearly assess the ability of data-driven parametrisation schemes to adapt to unseen climate conditions (Christensen, Moroz, and Palmer 2015) and to explore and exploit the potential for data-driven schemes to correct numerical errors in addition to modelling unresolved processes (Bhouri and Gentile 2023).

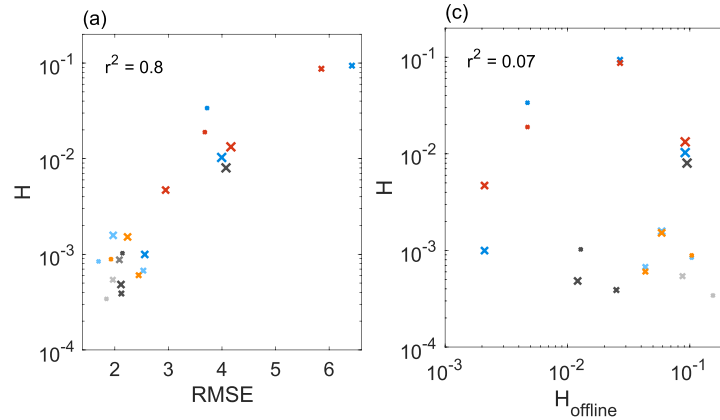


Figure 2.5: Reproduction of Figures 11a and 11c by Gagne II et al. (2020). Panel (a) compares ensemble mean weather forecast accuracy (measured by RMSE) to climatological PDF accuracy (measured by the Hellinger distance  $H$ ) for a collection of GAN parametrisations with different predictors and types of noise. Panel (c) performs a similar comparison with offline performance (measured by the Hellinger distance between the PDFs of the true and predicted tendencies) taking the place of RMSE.



## 2.6 Beyond Lorenz '96

The transition from L96 to more complex systems necessarily gives rise to several technicalities that must be addressed. The most pressing of these is the fact that increasing the number of degrees of freedom in the host model will greatly increase the number of predictors available for parametrisations to use. Models with more than one prognostic variable will also require separate parametrisations for each one. Without judicious simplifications, the problem of constructing a rigorous data-driven parametrisation will quickly become intractable. Crommelin and Vanden-Eijnden (2008) give an insightful discussion of this issue as it relates to Markov chain parametrisations.

Another complication is that the “spatial” dimension in L96 is discrete, while the dynamical systems being parametrised in weather and climate modelling (i.e., fluid flows) have continuous spatial dimensions. While it is well established that there is little benefit in using stochastic parametrisations with spatially correlated noise for L96, spatial correlation is expected to be a much more important consideration in spatially continuous systems (Arnold, Moroz, and Palmer 2013). One might also ask whether it will be beneficial to construct spatially nonlocal schemes where the tendency predicted at each point also depends on the values of the large-scale variables at neighbouring points (potentially capturing the gradients of these variables).

*This page intentionally left blank*

Talia flammato secum dea corde volutans  
nimborum in patriam, loca feta furentibus austris,  
Aeoliam venit. Hic vasto rex Aeolus antro  
luctantes ventos tempestatesque sonoras  
imperio premit ac vinclis et carcere frenat.  
Illi indignantes magno cum murmure montis  
circum claustra fremunt; celsa sedet Aeolus arce  
sceptra tenens, mollitque animos et temperat iras.  
Ni faciat, maria ac terras caelumque profundum  
quippe ferant rapidi secum verrantque per auras.  
Sed pater omnipotens speluncis abdidit atris,  
hoc metuens, molemque et montis insuper altos  
imposuit, regemque dedit, qui foedere certo  
et premere et laxas sciret dare iussus habenas.

Thus inwardly brooding with heart inflamed,  
the goddess came to Aeolia, motherland of  
storm clouds, tracts teeming with furious  
blasts. Here in his vast cavern, Aeolus, their  
king, keeps under his sway and with prison  
bonds curbs the struggling winds and the roar-  
ing gales. They, to the mountain's mighty  
moans, chafe blustering around the barriers.  
In his lofty citadel sits Aeolus, sceptre in hand,  
taming their passions and soothing their rage;  
did he not so, they would surely bear off with  
them in wild flight seas and lands and the  
vault of heaven, sweeping them through space.  
But, fearful of this, the father omnipotent hid  
them in gloomy caverns, and over them piled  
high mountain masses and gave them a king  
who, under fixed covenant, should be skilled  
to tighten and loosen the reins at command.

---

Virgil, *Aeneid* I, ll. 50–63  
trans. H. R. Fairclough

## Chapter 3

# Rayleigh-Bénard convection

TODO: introductory paragraph - explain where this is going

### 3.1 Governing equations

Rayleigh-Bénard convection is the overturning motion of a fluid confined between two horizontal isothermal plates, the bottom plate being warmer than the top plate. The fluid is traditionally assumed to be incompressible, with density  $\rho$  related to temperature  $T$  by the linear equation of state

$$\rho = \rho_0[1 - \alpha(T - T_0)],$$

$\alpha$  being the volume coefficient of thermal expansion and  $\rho_0$  and  $T_0$  a reference density and temperature. Provided that the density variations are small ( $\alpha(T - T_0) \ll 1$ ), one may employ the *Boussinesq approximation*, neglecting density variations everywhere except in their contribution to the weight force. This leads to the Boussinesq equations

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} = -\frac{1}{\rho_0} \nabla p' + \alpha(T - T_0)g\hat{\mathbf{z}} + \nu \nabla^2 \mathbf{u}, \quad (3.1)$$

$$\frac{\partial T}{\partial t} + \mathbf{u} \cdot \nabla T = \kappa \nabla^2 T, \quad \text{and} \quad (3.2)$$

$$\nabla \cdot \mathbf{u} = 0, \quad (3.3)$$

which represent the momentum balance, the conservation of energy and the assumption of incompressibility, respectively. The prognostic variables are the fluid velocity  $\mathbf{u}$  and the temperature  $T$  (the pressure perturbation  $p'$  is implicitly determined by (3.3)). The parameters of the model are the kinematic viscosity  $\nu$  and thermal diffusivity  $\kappa$ .  $\hat{\mathbf{z}}$  is the upward unit vector. The Boussinesq approximation has led to the appearance of a buoyant force (per unit mass)

$$\alpha(T - T_0)g = \frac{\rho_0 - \rho}{\rho_0}g$$

in the momentum equation, in agreement with Archimedes' principle. The reader is referred to Chandrasekhar (1961) for a detailed derivation of the Boussinesq equations; I have merely summarised the main assumptions and approximations involved.

In this work, I consider solutions of (3.1)–(3.3) in a two-dimensional domain  $[0, d] \times [0, L]$  with Cartesian coordinates  $x$  and  $z$ , subject to no-slip, isothermal boundary conditions

$$\mathbf{u} = \mathbf{0}, \quad T = T_0 + \frac{\delta T}{2} \quad \text{at } z = 0 \text{ and} \quad (3.4)$$

$$\mathbf{u} = \mathbf{0}, \quad T = T_0 - \frac{\delta T}{2} \quad \text{at } z = d \quad (3.5)$$

on the top and bottom plates, and periodic boundary conditions

$$\mathbf{u}(x=0) = \mathbf{u}(x=L) \quad \text{and} \quad T(x=0) = T(x=L) \quad (3.6)$$

in the horizontal.  $\delta T$  is the constant temperature difference between the plates.

## 3.2 Nondimensionalisation and scale analysis

It is helpful to nondimensionalise the governing equations (3.1)–(3.6); this is not only useful for numerical work but also gives insight into the different flow regimes that are possible. A range of nondimensionalisations are used in fluid dynamics literature; I adopt a common one (see, e.g., Grötzbach 1983; Ouertatani et al. 2008; Stevens, Verzicco, and Lohse 2010) which is suitable for the turbulent convective regime.

The first step is to choose representative time, length and temperature scales. For low-viscosity, turbulent flow, a suitable time scale is the *free-fall time*  $t_0$ , which is the time for a fluid element with constant temperature  $T = T_0 - \delta T$  to fall from the top plate to the bottom plate under the influence of buoyancy ( $-g\alpha\delta T$ ) alone. It is simple to show that

$$t_0 \sim \left( \frac{d}{g\alpha\delta T} \right)^{1/2},$$

ignoring a factor of  $\sqrt{2}$ . The obvious length and temperature scales are the plate separation  $d$  and temperature difference  $\delta T$ , respectively.

Making the substitutions  $p'/\rho_0 \rightarrow \pi$  and  $T - T_0 \rightarrow \theta$  in (3.1)–(3.6) and expressing all variables in units of  $t_0$ ,  $d$  and  $\delta T$  leads to the dimensionless equations

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} = -\nabla \pi + \left( \frac{\text{Pr}}{\text{Ra}} \right)^{1/2} \nabla^2 \mathbf{u} + \theta \hat{\mathbf{z}}, \quad (3.7)$$

$$\frac{\partial \theta}{\partial t} + \mathbf{u} \cdot \nabla \theta = (\text{Ra Pr})^{-1/2} \nabla^2 \theta, \quad \text{and} \quad (3.8)$$

$$\nabla \cdot \mathbf{u} = 0, \quad (3.9)$$

which are solved in the domain  $[0, \Gamma] \times [0, 1]$  with boundary conditions

$$\mathbf{u} = \mathbf{0}, \quad \theta = +\frac{1}{2} \quad \text{at } z = 0, \quad (3.10)$$

$$\mathbf{u} = \mathbf{0}, \quad \theta = -\frac{1}{2} \quad \text{at } z = 1, \quad (3.11)$$

$$\mathbf{u}(x=0) = \mathbf{u}(x=\Gamma) \quad \text{and} \quad \theta(x=0) = \theta(x=\Gamma). \quad (3.12)$$

There are three dimensionless parameters: the aspect ratio of the domain

$$\Gamma \equiv \frac{L}{d},$$

the *Prandtl number*

$$\text{Pr} \equiv \frac{\nu}{\kappa}$$

which measures the relative importance of viscosity (momentum diffusivity) and thermal diffusivity, and the *Rayleigh number*

$$\text{Ra} \equiv \frac{g\alpha d^3 \delta T}{\kappa \nu}.$$

The Rayleigh number can be interpreted as the ratio of the time scale for thermal transport by conduction to the time scale for thermal transport by convection. It determines the importance of diffusion for the evolution of  $\mathbf{u}$  and  $\theta$ ; inspection of (3.7) and (3.8) indicates that low Ra implies strong diffusion and high Ra weak diffusion. Stability analysis (see, e.g., Chandrasekhar (1961) and the seminal work by Lord Rayleigh (1916)) reveals that there exists a critical Rayleigh number (dependent on boundary conditions but of order  $10^3$ ), below which the equations have a stable equilibrium with the fluid at rest and a linear conductive temperature profile. Above the critical Rayleigh number, the equilibrium is unstable and small perturbations lead to the formation of a regular series of steady convection cells. If the Rayleigh number is increased much further (Le Quéré (1991) cites  $\text{Ra} \approx 2 \times 10^8$ ), the solution becomes unsteady and increasingly turbulent. This work is concerned with the turbulent regime, since Rayleigh numbers for atmospheric deep moist convection can be as large as  $10^{22}$  (Chillà and Schumacher 2012).

### 3.3 Hyperdiffusion terms

In Chapter 5 I will compare the output of a parametrised low-resolution model of (3.1)–(3.12) to the output of the unparametrised base model (i.e., the control). However, it was found in the early stages of this work that unparametrised low-resolution models were prone to numerical instability, preventing the obtainment of a suitable control solution. This is to be expected; at high Rayleigh numbers, when dissipation is weak, the solutions develop large gradients near small-scale features that cannot be properly resolved by coarse models, causing the model to crash. To be more precise, the energy spectra of turbulent flows exhibit an *energy cascade* whereby energy is transferred from larger-scale motions to smaller-scale motions; only at sufficiently small scales is energy removed from the system by the dissipative terms in the equations (Pope 2000). In coarse models where these smallest scales are not resolved, excess energy accumulates at the highest resolved wavenumbers.

I choose to stabilise the numerical model by introducing artificial dissipative terms that act on the smallest scales. The modified equations, with the new terms highlighted in red, are

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} = -\nabla \pi + \left( \frac{\text{Pr}}{\text{Ra}} \right)^{1/2} (\nabla^2 \mathbf{u} + \tilde{\nu} f(z) |\nabla^2 \mathbf{u}| \nabla^2 \mathbf{u}) + \theta \hat{\mathbf{z}}, \quad (3.13)$$

$$\frac{\partial \theta}{\partial t} + \mathbf{u} \cdot \nabla \theta = (\text{Ra Pr})^{-1/2} (\nabla^2 \theta + \tilde{\kappa} f(z) |\nabla^2 \theta| \nabla^2 \theta), \quad \text{and} \quad (3.14)$$

$$\nabla \cdot \mathbf{u} = 0, \quad (3.15)$$

where the constants  $\tilde{\nu} = \tilde{\kappa} = 2 \times 10^{-3}$  were chosen by trial and error to achieve sufficient dissipation. The function  $f(z)$  is necessary to reduce the artificial terms to zero near the top and bottom of the domain (otherwise the model crashes) while leaving them relatively unchanged in the interior, and is defined by

$$f(z) = \left[ 1 - \exp \left( - \frac{\min(z, 1-z)}{0.052} \right) \right]^4.$$

The definition of  $f(z)$  is inspired by the *van Driest damping function* commonly used in large-eddy simulations of wall-bounded flows (Pope 2000).

The new additions are akin to regular viscous and diffusive terms with viscosity and diffusivity coefficients that depend on second derivatives of the flow variables. Small-scale features, where  $|\nabla^2 \mathbf{u}|$  and/or  $|\nabla^2 \theta|$  are large, will experience a larger effective viscosity and/or thermal diffusivity respectively, and therefore be dissipated more rapidly than larger-scale features.

It must be emphasised that this work is *not* concerned with whether or not these modified equations are an accurate representation of physical reality; they simply serve as an example dynamical system for the study of parametrisation.

### 3.4 Model configuration

The nondimensionalised Boussinesq equations (3.13)–(3.15) and (3.10)–(3.12) are solved numerically using Dedalus (v3), a spectral code built in Python (Burns et al. 2020). Spectral methods represent the solution of a partial differential equation as a linear combination of linearly independent basis functions, solving for the coefficients of the basis function expansion rather than the values of the solution in real space. The resolution of the model is thus set by choosing the number of basis functions used. This work uses a Fourier (sine/cosine) basis in the horizontal direction and a basis of Chebyshev polynomials (of the first kind) in the vertical direction. The Fourier basis has the special property that all linear combinations respect the periodic boundary conditions. For storage and analysis, the model output is transformed into real space, where the fields are defined on a *collocation grid* that has uniform horizontal spacing but non-uniform vertical spacing (smaller near the top and bottom of the domain and larger in the middle).

Timestepping is performed using a second-order Runge-Kutta scheme. The Rayleigh number was set to  $10^9$ , a value at which the flow was found to be unsteady and feature transient eddies. The aspect ratio of the domain was  $\Gamma = 8$  and the Prandtl number was set to 1 for simplicity. The complete model source code is publicly available [TODO: data availability](#).

[TODO: link to animations](#)

### 3.5 Choice of resolution: literature

The next important step is to choose the resolution of the fine and coarse models. The fine model will serve as “truth”, providing the training dataset to which the parametrisation model will be fitted and the test dataset against which the parametrised coarse model will be evaluated. Its resolution should therefore be high enough that the (yet-to-be-)chosen evaluation metrics are relatively insensitive to small changes in resolution (i.e., there should be near-convergence of long-term statistics). Note, however, that neither a perfect solution of the fluid equations nor perfect agreement with real-world experimental measurements are necessary; the fine model merely serves as a reference dynamical system whose behaviour I aim to reproduce by parametrisation the coarse model. In order to create a worthwhile parametrisation problem, the coarse model’s resolution should therefore be low enough that it exhibits statistically significant biases relative to the fine model, as measured by the chosen evaluation metrics. However, as explained in § 3.3, the coarse model must still be numerically stable so that it can provide an unparametrised control solution.

With the above constraints in mind, § 3.5.1 will review the guidelines that have been established in the literature for producing well-resolved simulations of Rayleigh-Bénard convection. § 3.5.2 will then review the metrics that are known to be sensitive to under-resolution; these may be used to experimentally determine appropriate resolutions for the fine and coarse models, and later (in Chapter 5) to evaluate parametrisation performance.

#### 3.5.1 Resolution guidelines

Grötzbach (1983) is recognised as the first to formulate resolution requirements for accurate simulations of Rayleigh-Bénard convection (Chillà and Schumacher 2012; Scheel, Emran, and Schumacher 2013). He identified separate constraints on the mean (i.e., averaged in each spatial direction) grid spacing and the vertical spacing near the plates; I first discuss the former. Grötzbach reasoned that a numerical model that neglects subgrid-scale effects must have a geometric mean grid spacing  $h$  (i.e.  $h = (\Delta x \Delta y \Delta z)^{1/3}$  in

3D) such that

$$h \leq \underbrace{\pi \left( \frac{\text{Pr}}{\text{Ra}} \right)^{3/8} \langle \epsilon_k \rangle^{-1/4}}_{\eta}, \quad (3.16)$$

where  $\eta$  is the *Kolmogorov length*, the universal smallest relevant length scale for general turbulent flow, and  $\langle \epsilon_k \rangle$  is the spatial and temporal average of the kinetic energy dissipation rate defined by

$$\epsilon_k(\mathbf{x}, t) \equiv \frac{1}{2} \left( \frac{\text{Pr}}{\text{Ra}} \right)^{1/2} \sum_{ij} \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right)^2 \quad (3.17)$$

(Chillà and Schumacher 2012). The inequality (3.16) between  $h$  and  $\eta$  can be understood using the Nyquist-Shannon theorem, which states that a sampling frequency  $f \geq k/\pi$  is needed to unambiguously reconstruct a signal with maximum wavenumber  $k$ ; substituting  $f = 1/h$ ,  $k = 1/\eta$  leads to the claimed relation.

Grötzbach recognised that the above reasoning was only valid for the mean grid spacing; large gradients in temperature and velocity near the top and bottom plates require finer resolution in those regions. The notion of nearness can be formalised by considering the thickness of the *thermal boundary layer*, the region at each plate where large temperature gradients exist. Consider a Taylor series expansion of the horizontally- and time-averaged dimensionless temperature profile about  $z = 0$ ; recalling the boundary condition (3.10), this reads

$$\langle \theta \rangle_{A,t} \approx \frac{1}{2} + \left. \frac{\partial \langle \theta \rangle_{A,t}}{\partial z} \right|_{z=0} z + O(z^2).$$

The thermal boundary layer thickness  $\delta_\theta$  is defined (see, e.g., Chillà and Schumacher 2012) as the height at which, to first order, the temperature reaches the mean value in the well-mixed interior of the domain (i.e., zero for this problem). It follows that

$$\delta_\theta = -\frac{1}{2} \left( \left. \frac{\partial \langle \theta \rangle_{A,t}}{\partial z} \right|_{z=0} \right)^{-1}. \quad (3.18)$$

And similarly for the other boundary layer at  $z = 1$ . One then asks how many grid points must be within the thermal boundary layer. Grötzbach claimed that 3 points are sufficient for turbulent flows but did not give a theoretical argument to derive this number. Shishkina et al. (2010) presented a theoretical argument based on the (experimentally and numerically justified) assumption of laminar *Prandtl-Blasius* flow conditions in the boundary layer and were able to calculate the minimum number of grid points (e.g., 9 for  $\text{Ra} = 2 \times 10^9$  and  $\text{Pr} = 0.7$ ). The results agreed with criteria derived in previous numerical experiments. Importantly, the results of Shishkina et al. allow *a priori* determination of vertical resolution requirements, potentially bypassing the time-consuming and expensive process of iteratively running simulations, checking their convergence and updating the resolution.

### 3.5.2 Metrics sensitive to under-resolution

Performing numerical experiments for a 3D fluid layer, Grötzbach found that the RMS velocity and the *Nusselt number* were the most sensitive quantities to insufficient mean grid spacing, but even they increased “only slightly” above the values obtained from well-resolved simulations. The dimensionless Nusselt number measures the instantaneous rate of (vertical) heat transport across a horizontal plane at height  $z$ , normalised by the purely conductive rate that would exist if the fluid were at rest (Verzicco and Camussi 1999). Following Kooij et al. (2018), I define

$$\text{Nu}(z, t) \equiv \sqrt{\text{Ra Pr}} \langle w\theta \rangle_A - \left\langle \frac{\partial \theta}{\partial z} \right\rangle_A$$

where  $\langle \cdot \rangle_A$  denotes averaging over the horizontal plane at height  $z$ , and  $w = \mathbf{u} \cdot \hat{\mathbf{z}}$  is the vertical velocity. Notice that the heat transport has contributions from both advection (the first term) and conduction (the second term). Authors frequently consider the vertically-averaged value

$$\langle \text{Nu}(z, t) \rangle_z = \sqrt{\text{Ra Pr}} \langle w \theta \rangle_{A,z} + 1 \quad (3.19)$$

and/or the time-averaged value, which is independent of height due to energy conservation,

$$\langle \text{Nu}(z, t) \rangle_t = \langle \text{Nu}(z, t) \rangle_{z,t} = \sqrt{\text{Ra Pr}} \langle w \theta \rangle_{A,z,t} + 1.$$

Later work also supports the finding that the Nusselt number is sensitive to under-resolution. Even studying only steady-state convective solutions at moderate Rayleigh number, Le Quéré (1991) found that the maximum and minimum Nusselt numbers were most sensitive to changes in resolution and had the largest uncertainty among existing benchmark solutions. Other studies have used the convergence of the Nusselt number as an indicator that the spatial resolution is sufficient to produce an accurate solution (Ouertatani et al. 2008).

Stevens, Verzicco, and Lohse (2010) performed 3D simulations in a finite cylindrical cavity with the aim of reconciling the apparent disagreement between the Nusselt numbers in previous numerical studies and experimental observations. They found that agreement with experiment could be achieved, but only by using a much higher resolution than the previous studies. They offered the physical explanation that horizontally under-resolved simulations produce insufficient thermal diffusion, leading to systematic overestimation of the buoyancy of convective plumes near the side-walls of the cylinder; this results in Nusselt numbers that exceed experimentally observed values. This led them to conclude that the two criteria of Grötzbach (1983)—for mean grid spacing and for the vertical spacing near the upper and lower plates—are not independent; the definition  $h = (\Delta x \Delta y \Delta z)^{1/3}$  in (3.16) allows the horizontal spacing to remain relatively coarse near the plates, provided the vertical spacing is small. Since fine horizontal resolution is also necessary to accurately capture the dynamics of the thin plumes, they proposed that (3.16) be applied with  $h = \max(\Delta x, \Delta y, \Delta z)$  instead.

Some more recent work, however, casts doubt on the notion that the Nusselt number is sensitive to under-resolution and that its convergence is a good indicator that the flow is well-resolved. In assessing the performance of several published computational fluid dynamics codes on the Rayleigh-Bénard problem in a cylindrical cavity, Kooij et al. (2018) identified one higher-order code that reproduced the theoretically predicted scaling of Nu as a function of Ra even when the flow was deliberately under-resolved. On the other hand, the presence of numerical artefacts in the instantaneous temperature field near the bottom plate was a clear indicator of insufficient resolution.

Scheel, Emran, and Schumacher (2013) performed similar high-resolution simulations for a cylindrical cavity and also found that the Nusselt number, among other global transport properties, were “fairly insensitive to insufficient resolution, as long as the mean Kolmogorov length [was] resolved” (i.e., (3.16) was satisfied). However, they found that the horizontally averaged or local kinetic energy dissipation rate (3.17) and the corresponding thermal dissipation rate

$$\epsilon_\theta(\mathbf{x}, t) \equiv (\text{Ra Pr})^{-1/2} \sum_i \left( \frac{\partial \theta}{\partial x_i} \right)^2 \quad (3.20)$$

were much more sensitive, with their convergence requiring even stricter conditions than (3.16).

### 3.6 Choice of resolution: experiment

I have performed a series of simulations at different resolutions using the numerical model described in § 3.4, using the metrics identified in § 3.5.2 to determine appropriate resolutions for the fine and coarse models in the parametrisation study of Chapters 4 and 5. The coarsest simulation had a resolution of 256 horizontal (Fourier) modes by 64 vertical (Chebyshev) modes and the highest 2048 by 256.



Higher-resolution simulations require shorter time steps in order to be stable. The suitability of the time step is determined by the *Courant number* (Pope 2000)

$$C \equiv \frac{|u|}{\Delta x/\Delta t} + \frac{|w|}{\Delta z/\Delta t},$$

which expresses the horizontal and vertical velocities  $u$  and  $w$  in units of the grid spacings<sup>1</sup>  $\Delta x$  and  $\Delta z$  per time step  $\Delta t$ , and must be sufficiently small (typically less than 1). The time step was tuned by trial and error at an intermediate starting resolution (not shown) and the time step for other resolutions was scaled to approximately preserve  $C$ . The scaling is derived by reasoning that, if  $|u|$  and  $|w|$  are approximately equal on average and independent of resolution, then for two simulations denoted by subscripts 1 and 2, the equality of  $C_1$  and  $C_2$  implies

$$\begin{aligned} \frac{1}{\Delta x_1/\Delta t_1} + \frac{1}{\Delta z_1/\Delta t_1} &= \frac{1}{\Delta x_2/\Delta t_2} + \frac{1}{\Delta z_2/\Delta t_2} \\ \Leftrightarrow \quad \Delta t_2 &= \frac{1/\Delta x_1 + 1/\Delta z_1}{1/\Delta x_2 + 1/\Delta z_2} \Delta t_1. \end{aligned}$$

Recognising that the grid spacings are related to the numbers of grid points (or spectral modes)  $N_x$  and  $N_z$  by  $\Delta x = \Gamma/N_x$  and  $\Delta z \approx 1/N_z$  (where  $\Gamma$  is the aspect ratio of the domain), this may be further simplified to

$$\Delta t_2 = \frac{N_{x_1}/\Gamma + N_{z_1}}{N_{x_2}/\Gamma + N_{z_2}} \Delta t_1.$$

Table 3.1 summarises the resolutions of the simulations and their time steps.

$N_x$	$N_z$	$\Delta t$
256	64	$5.333 \times 10^{-3}$
512	64	$4 \times 10^{-3}$
768	96	$2.667 \times 10^{-3}$
1024	128	$2 \times 10^{-3}$
1280	160	$1.6 \times 10^{-3}$
1536	192	$1.333 \times 10^{-3}$
1792	224	$1.143 \times 10^{-3}$
2048	256	$1 \times 10^{-3}$

Table 3.1: Resolutions and time steps of the simulations used in the resolution-dependence experiment.

The simulations with resolution  $1024 \times 128$  and coarser were started from the initial condition described in Appendix B.1 and run for 1000 time units. The simulations with resolution finer than  $1024 \times 128$  were initialised by interpolating the  $1024 \times 128$  solution at time  $t = 650$  and run for a further 350 time units to reduce computational cost. All output was saved at an interval of approximately 0.2 time units. The first 750 time units of data were discarded to ensure that all simulations had reached a statistically steady state (Appendix B.2 verifies this).

For each simulation, I calculated time averages of the following quantities:

- The vertically-averaged Nusselt number  $\langle \text{Nu}(z, t) \rangle_z$ , defined by (3.19),
- The horizontally-averaged thickness of the two thermal boundary layers  $(\langle \delta_{\theta, z=0} \rangle_x + \langle \delta_{\theta, z=1} \rangle_x)/2$ , defined by (3.18),
- The RMS speed  $\sqrt{\langle \mathbf{u} \cdot \mathbf{u} \rangle_{x,z}}$ ,
- The domain-averaged kinetic energy dissipation rate  $\langle \epsilon_k \rangle_{x,z}$ , defined by (3.17), and

<sup>1</sup>Since the model uses a spectral method,  $\Delta x$  and  $\Delta z$  are taken to be the local spacings of the collocation grid in real space; see § 3.4.

- e. The domain-averaged thermal dissipation rate  $\langle \epsilon_\theta \rangle_{x,z}$ , defined by (3.20).

Time averaging was performed over the 250 time units of data that remained after discarding the first 750. I identified two related dominant sources of uncertainty in the calculated values: (i) the uncertainty due to the finite averaging time and (ii) the uncertainty arising from low-frequency fluctuations of the quantities being averaged (i.e., the fact that the results change if the finite averaging window is translated in time). I estimated (i) as half the range of the running mean,

$$\frac{1}{2} \text{range}_{\tau \in [150, 250]} \frac{1}{\tau} \int_{750}^{750+\tau} (\cdot) dt, \quad (3.21)$$

which measures the sensitivity of the result to changes in the averaging time. I estimated (ii) as half the range of the 150-time-unit rolling mean,

$$\frac{1}{2} \text{range}_{\tau \in [0, 100]} \frac{1}{150} \int_{750+\tau}^{900+\tau} (\cdot) dt, \quad (3.22)$$

which measures the sensitivity of the result to the translation of the averaging window. I took the larger of (3.21) and (3.22) as a conservative estimate of the overall uncertainty.

Figure 3.1 shows plots of the five metrics as functions of resolution, expressing the values as percentage errors relative to the highest-resolution simulation. Firstly, observe that all the metrics (except perhaps the RMS speed) appear insensitive to small changes in resolution about  $2048 \times 256$ ; I therefore choose  $2048 \times 256$  as the resolution of the fine “truth” model for the parametrisation study.

All five metrics exhibit statistically significant discrepancies across the range of resolutions tested. I choose  $256 \times 64$  as the resolution of the coarse model for the parametrisation study because it produces the largest error for all metrics while remaining numerically stable.

Notice also that the Nusselt number (Figure 3.1a) and thermal boundary layer thickness (Figure 3.1b) have both a monotonic resolution dependence and small relative uncertainties. This will make them reliable metrics for evaluating parametrisation performance in Chapter 5.

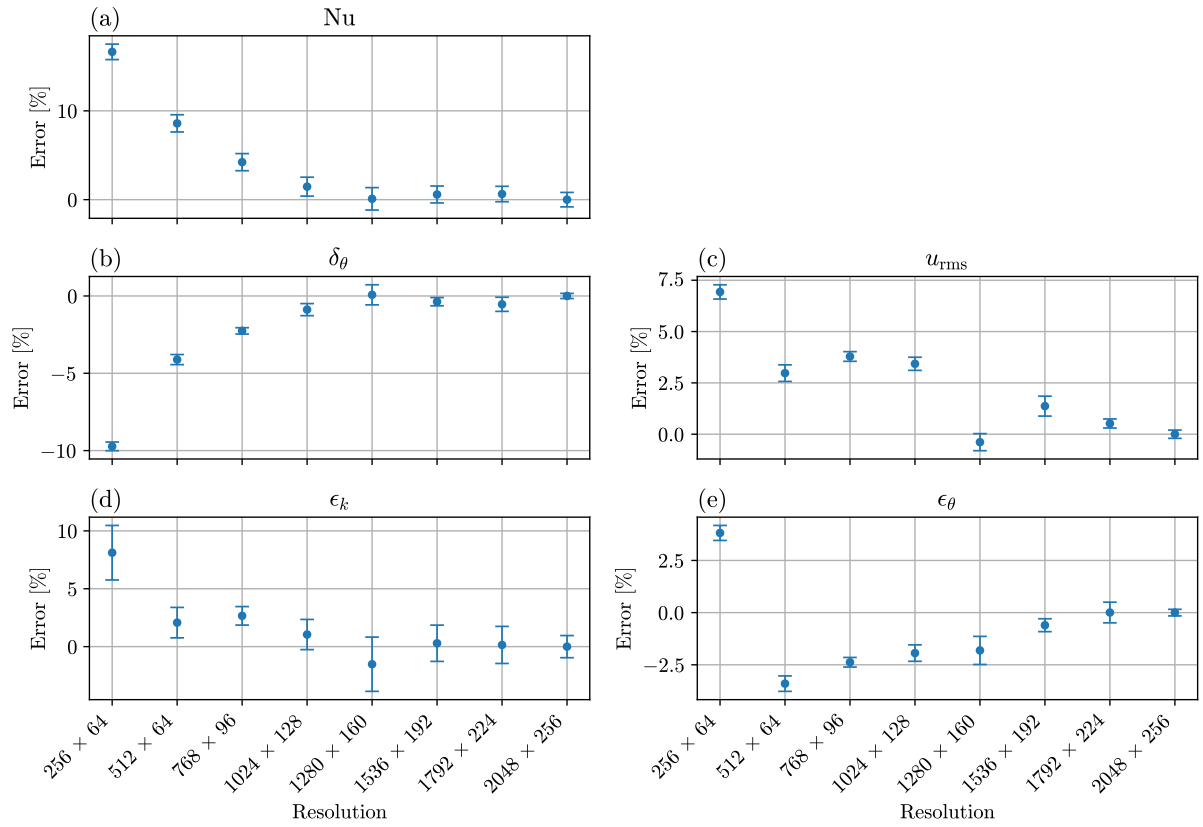


Figure 3.1: Time-averaged Nusselt number (a), thermal boundary layer thickness (b), RMS speed (c), kinetic energy dissipation rate (d) and thermal dissipation rate (e) as functions of model resolution, expressed as percentage errors relative to the values from the highest-resolution simulation (2048 × 256). Refer to the corresponding list items in the text for more detailed definitions of the quantities.

*This page intentionally left blank*

## Chapter 4

# Calculation and modelling of subgrid tendencies

TODO: introductory paragraph: why am I doing this?

### 4.1 Calculation of subgrid tendencies

The workflow used to calculate the subgrid tendencies is described below and illustrated as a flowchart in [Figure 4.1](#), which uses the same numbering to show the order of the steps.

1. The fine model was integrated for [TODO: ] time units. Every 3 time units, the model state was saved, and then saved again one time step later. This resulted in a dataset of [TODO: ] pairs of snapshots separated by 3 time units. An interval of 3 time units was chosen as it was the approximate decorrelation time of the model variables (see [Appendix B.3](#)); using a shorter interval would result in saving redundant information.
2. Each pair of snapshots was *coarse-grained*, reducing its spatial resolution to that of the coarse model. The nature of the coarse-graining operation warrants special attention and is discussed separately in [§ 4.2](#).
3. The first coarse-grained snapshot in each pair was input as an initial condition for the coarse model. The coarse model integrated for one time step only, and the resulting state—the coarse model’s prediction for the large-scale state after one time step—was saved.
4. The first coarse-grained snapshot in each pair from [Step 2](#) was subtracted from the second and the difference divided by the fine model’s time step, giving the *true coarse tendency* (i.e., the true time derivative of the large-scale state as calculated by the fine model).
5. The first coarse-grained snapshot in each pair from [Step 2](#) was subtracted from the coarse model prediction produced in [Step 3](#) and the difference divided by the coarse model’s time step, giving the tendency predicted by the coarse model (i.e., its prediction of the time derivative of the large-scale state).
6. Finally, the coarse model predicted tendencies from [Step 5](#) were subtracted from the true coarse tendencies from [Step 4](#), producing the *subgrid tendencies*. These measure the error in the coarse model’s prediction of the time derivative of the large-scale state, and would be identically zero for a perfect coarse model.

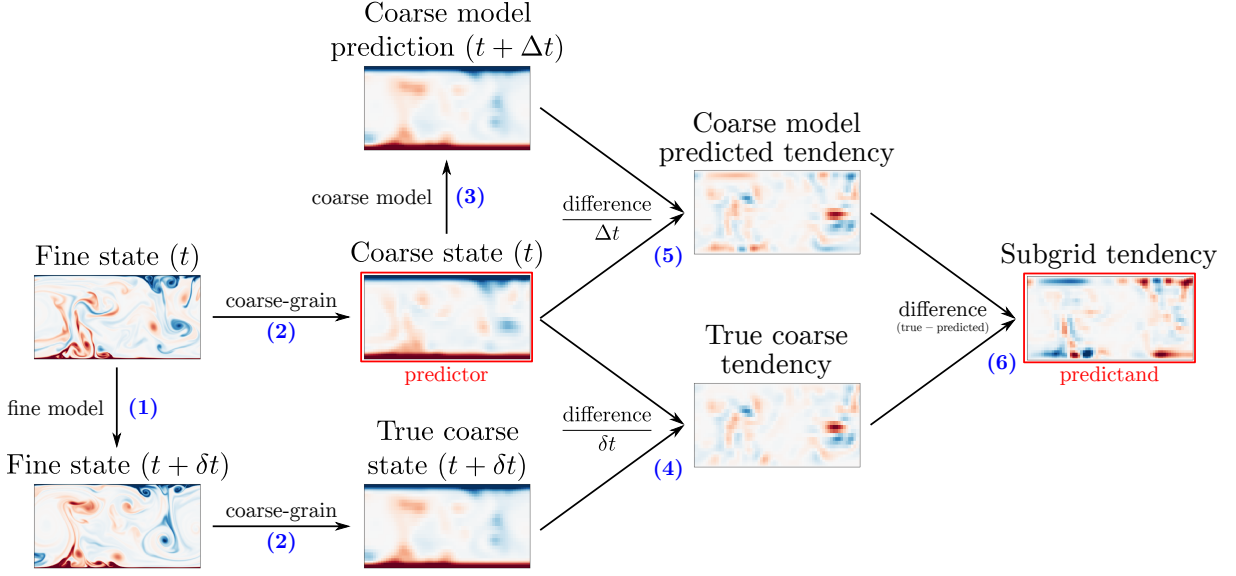


Figure 4.1: Flowchart illustrating the procedure used to calculate subgrid tendencies. The plots show an example of the workflow being applied to the temperature data but are for illustrative purposes only. The blue numbers correspond to the steps described in the text.

## 4.2 Choice of coarse-graining method

Coarse-graining is the process of reducing a gridded dataset onto a lower-resolution grid, and it is required at [Step 2](#) of the workflow described in [§ 4.1](#). It was found that the choice of coarse-graining method was a major influence on the quality of the calculated subgrid tendencies; this is the consequence of a subtle issue that may seem purely semantic at first but in fact has important practical implications.

In general, the output of a coarse (i.e., reduced-order) model is meant to approximate a certain *representation* of the output of a chosen fine model. To give three concrete examples, the coarse model might seek to reproduce (a) the values of the high-resolution fields on a sparser grid of points, or (b) the averages over a set of larger grid boxes, or (c) the first  $N$  coefficients of the discrete Fourier transform (where  $N$  is less than the number of coefficients needed to fully determine the original fields). The modeller has the freedom choose a representation, which in turn determines how the output of the coarse model should be interpreted. The choice of representation also implicitly determines a coarse-graining operation: a map from the state space of the fine model to the state space of the coarse model that isolates the necessary large-scale information and discards the rest. Referring to the previous examples, option (a) calls for an operation that simply discards, say, three out of every four or nine out of ten grid points. Option (b) calls for the grouping and averaging of the grid points that lie within each large grid box. Option (c) calls for the truncation of fine fields at the  $N$ th coefficient in Fourier space.

The key lesson that was learnt during the course of this work is that the chosen representation and coarse-graining method must be appropriate to the nature of the coarse model. In this work, the coarse model was a Dedalus solver that, in [Step 3](#) of the workflow in [§ 4.1](#), received gridded initial condition data in real space and integrated the governing equations forward in time using the same numerical method as the fine model. This gave rise to two constraints on the coarse-graining method:

1. The coarse-grained initial condition must be well-resolved on the coarse model's grid. Numerical solution algorithms for PDEs assume (e.g., by approximating derivatives as finite differences) that the solution is well-resolved on the discrete model grid, and can become unstable or produce output marred by artefacts if this condition is not met.
2. The initial condition must respect physical constraints, namely the divergence-free condition [\(3.9\)](#)

and the boundary conditions (3.10)–(3.12). A numerical algorithm cannot be expected to behave predictably when presented with unphysical initial conditions.

During the development of this study, before the above requirements were known, coarse-graining was performed by averaging the fine grid points that lay within each coarse grid box (a method known within the Earth sciences as *first-order conservative remapping* because it preserves mean values; see Jones 1999). Figure **TODO: figure comparing coarse-graining methods**, illustrating the application of first-order conservative remapping to sample temperature data, demonstrates that the result is not very well-resolved on the coarse grid; in many places, adjacent grid points have sharp differences in temperature. First-order conservative remapping is also not guaranteed to preserve boundary values or the divergence-free nature of the velocity field. Consequently, the tendencies obtained from the coarse model in Step 5 of the workflow suffered from noise and numerical artefacts that propagated to the subgrid tendencies in Step 6. With the signal of interest obscured, it was impossible to model the subgrid tendencies as functions of the coarse state.

It was evident that the coarse-graining method needed to involve a smoothing operation of some kind. However, most conventional smoothing methods, such as Gaussian filtering, also fail to preserve the boundary and divergence-free conditions. The solution to this problem was to supply each snapshot of the fine model state as an initial condition to an appropriate system of PDEs, chosen so that the integration of the system would have a smoothing effect on the initial condition data. The key advantage of this approach is that the boundary and divergence-free conditions can be explicitly enforced.

The choice of PDEs to achieve this was inspired by classical Gaussian filtering. Recall that a Gaussian filter convolves input data with a filter kernel that takes the form of a Gaussian function, which also happens to be the Green’s function for the heat equation  $\partial\psi/\partial t = \nabla^2\psi$  in an *infinite* domain. It follows that heat equations of some form on the *finite* domain of the Rayleigh–Bénard problem (whose Green’s functions are not Gaussian) will have a similar smoothing effect while preserving the required boundary conditions. The appropriate equation for smoothing the temperature field  $\theta$  is therefore

$$\frac{\partial\theta}{\partial t} = \nabla^2\theta \quad (4.1)$$

on  $[0, \Gamma] \times [0, 1]$  with  $\theta(z = 0) = +1/2$ ,  $\theta(z = 1) = -1/2$  and  $\theta(x = 0) = \theta(x = \Gamma)$ . It is tempting to propose an equation of the same form,  $\partial\mathbf{u}/\partial t = \nabla^2\mathbf{u}$ , for the velocity field, but its solution on the finite domain would not necessarily preserve  $\nabla \cdot \mathbf{u} = 0$ . By analogy to the incompressible Navier-Stokes equations, it is necessary to introduce a “pressure” term  $-\nabla\pi$  on the right-hand side, with the additional field  $\pi(x, z)$  giving an additional degree of freedom that allows one to impose  $\nabla \cdot \mathbf{u} = 0$  without over-determining the problem. The velocity field is thus smoothed by the equations

$$\frac{\partial\mathbf{u}}{\partial t} = -\nabla\pi + \nabla^2\mathbf{u} \quad \text{and} \quad (4.2)$$

$$\nabla \cdot \mathbf{u} = 0 \quad (4.3)$$

on  $[0, \Gamma] \times [0, 1]$  with  $\mathbf{u}(z = 0) = \mathbf{u}(z = 1) = \mathbf{0}$  and  $\mathbf{u}(x = 0) = \mathbf{u}(x = \Gamma)$ . The equations (4.1)–(4.3) were integrated numerically on the same grid as the fine model using a solver that was straightforwardly implemented in Dedalus.

Once the high-resolution data had been smoothed using the method described above, all that remained was to downsample the result to the resolution of the coarse model using linear interpolation **TODO: actually using change scales function in dedalus**. To summarise, the coarse-graining workflow is as follows:

1. Input a high-resolution snapshot as the initial condition for the solver of (4.1)–(4.3).
2. Integrate the solver for  $10^{-3}$  time units (the duration that was found to produce the necessary amount of smoothing) in steps of  $2 \times 10^{-4}$  time units. **TODO: check values**
3. Downsample the final state to the resolution of the coarse model and save the result.

4. Reset the solver and return to [Step 1](#) to coarse-grain the next high-resolution snapshot.

Figure [TODO: figure comparing coarse-graining methods](#) demonstrates that the improved coarse-graining method produces smooth results that are well-resolved on the coarse grid, in contrast to first-order conservative remapping.

### 4.3 Analysis of subgrid tendencies

### 4.4 Modelling of subgrid tendencies



## Chapter 5

# Evaluation of the parametrised model

*This page intentionally left blank*

## Chapter 6

## Conclusion

*This page intentionally left blank*

# Bibliography

- Alcala, J. and I. Timofeyev (2021). “Subgrid-scale parametrization of unresolved scales in forced Burgers equation using generative adversarial networks (GAN)”. *Theor. Comput. Fluid Dyn.* **35**(6). DOI: [10.1007/s00162-021-00581-z](https://doi.org/10.1007/s00162-021-00581-z).
- Andrews, T., J. M. Gregory, M. J. Webb, and K. E. Taylor (2012). “Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere-ocean climate models”. *Geophys. Res. Lett.* **39**(9). DOI: [10.1029/2012GL051607](https://doi.org/10.1029/2012GL051607).
- Arnold, H. M., I. M. Moroz, and T. N. Palmer (2013). “Stochastic parametrizations and model uncertainty in the Lorenz ’96 system”. *Phil. Trans. R. Soc. A* **371**(1991). DOI: [10.1098/rsta.2011.0479](https://doi.org/10.1098/rsta.2011.0479).
- Berner, J., U. Achatz, L. Batté, L. Bengtsson, A. de la Cámara, H. M. Christensen, M. Colangeli, D. R. B. Coleman, D. Crommelin, S. I. Dolaptchiev, C. L. E. Franzke, P. Friederichs, P. Imkeller, H. Järvinen, S. Juricke, V. Kitsios, F. Lott, V. Lucarini, S. Mahajan, T. N. Palmer, C. Penland, M. Sakradzija, J.-S. von Storch, A. Weisheimer, M. Weniger, P. D. Williams, and J.-I. Yano (2017). “Stochastic parameterization: toward a new view of weather and climate models”. *Bull. Am. Meteorol. Soc.* **98**(3). DOI: [10.1175/BAMS-D-15-00268.1](https://doi.org/10.1175/BAMS-D-15-00268.1).
- Beucler, T., I. Ebert-Uphoff, S. Rasp, M. Pritchard, and P. Gentine (2022). “Machine learning for clouds and climate”. *ESS Open Archive*. Preprint. DOI: [10.1002/essoar.10506925.1](https://doi.org/10.1002/essoar.10506925.1).
- Bhouri, M. A. and P. Gentine (2023). “Memory-based parameterization with differentiable solver: application to Lorenz ’96”. *Chaos* **33**(7). DOI: [10.1063/5.0131929](https://doi.org/10.1063/5.0131929).
- Brajard, J., A. Carrassi, M. Bocquet, and L. Bertino (2021). “Combining data assimilation and machine learning to infer unresolved scale parametrization”. *Phil. Trans. R. Soc. A* **379**(2194). DOI: [10.1098/rsta.2020.0086](https://doi.org/10.1098/rsta.2020.0086).
- Burns, K. J., G. M. Vasil, J. S. Oishi, D. Lecoanet, and B. P. Brown (2020). “Dedalus: a flexible framework for numerical simulations with spectral methods”. *Phys. Rev. Res.* **2**(2). DOI: [10.1103/PhysRevResearch.2.023068](https://doi.org/10.1103/PhysRevResearch.2.023068).
- Chandrasekhar, S. (1961). *Hydrodynamic and hydromagnetic stability*. Oxford: Clarendon Press. ISBN: 9780486319209.
- Chillà, F. and J. Schumacher (2012). “New perspectives in turbulent Rayleigh-Bénard convection”. *Eur. Phys. J. E* **35**(7). DOI: [10.1140/epje/i2012-12058-1](https://doi.org/10.1140/epje/i2012-12058-1).
- Chorin, A. J. and F. Lu (2015). “Discrete approach to stochastic parametrization and dimension reduction in nonlinear dynamics”. *PNAS* **112**(32). DOI: [10.1073/pnas.1512080112](https://doi.org/10.1073/pnas.1512080112).
- Christensen, H. M., I. M. Moroz, and T. N. Palmer (2015). “Simulating weather regimes: impact of stochastic and perturbed parameter schemes in a simple atmospheric model”. *Clim. Dyn.* **44**(7). DOI: [10.1007/s00382-014-2239-9](https://doi.org/10.1007/s00382-014-2239-9).
- Christensen, H. and L. Zanna (2022). “Parametrization in weather and climate models”. *Oxford Research Encyclopedia of Climate Science*. Oxford University Press. DOI: [10.1093/acrefore/9780190228620.013.826](https://doi.org/10.1093/acrefore/9780190228620.013.826).
- Christensen, H. M. (2020). “Constraining stochastic parametrisation schemes using high-resolution simulations”. *Q. J. R. Meteorol. Soc.* **146**(727). DOI: [10.1002/qj.3717](https://doi.org/10.1002/qj.3717).
- Christensen, H. M. and J. Berner (2019). “From reliable weather forecasts to skilful climate response: a dynamical systems approach”. *Q. J. R. Meteorol. Soc.* **145**(720). DOI: [10.1002/qj.3476](https://doi.org/10.1002/qj.3476).
- Crommelin, D. and E. Vanden-Eijnden (2008). “Subgrid-scale parameterization with conditional Markov chains”. *J. Atmos. Sci.* **65**(8). DOI: [10.1175/2008JAS2566.1](https://doi.org/10.1175/2008JAS2566.1).

- Demaeyer, J. and S. Vannitsem (2018). “Stochastic parameterization of subgrid-scale processes: a review of recent physically based approaches”. *Advances in Nonlinear Geosciences*. Ed. by A. A. Tsonis. Cham: Springer International Publishing, pp. 55–85. ISBN: 978-3-319-58895-7. DOI: [10.1007/978-3-319-58895-7\\_3](https://doi.org/10.1007/978-3-319-58895-7_3).
- Donner, L. J. and V. T. Phillips (2003). “Boundary layer control on convective available potential energy: implications for cumulus parameterization”. *J. Geophys. Res.: Atmos.* **108**(D22). DOI: [10.1029/2003JD003773](https://doi.org/10.1029/2003JD003773).
- Franzke, C. L. E., T. J. O’Kane, J. Berner, P. D. Williams, and V. Lucarini (2015). “Stochastic climate theory and modeling”. *WIREs Clim. Change* **6**(1). DOI: [10.1002/wcc.318](https://doi.org/10.1002/wcc.318).
- Gagne II, D. J., H. M. Christensen, A. C. Subramanian, and A. H. Monahan (2020). “Machine learning for stochastic parameterization: generative adversarial networks in the Lorenz ’96 model”. *J. Adv. Model. Earth Syst.* **12**(3). DOI: [10.1029/2019MS001896](https://doi.org/10.1029/2019MS001896).
- Gentine, P., M. Pritchard, S. Rasp, G. Reinaudi, and G. Yacalis (2018). “Could machine learning break the convection parameterization deadlock?” *Geophys. Res. Lett.* **45**(11). DOI: [10.1029/2018GL078202](https://doi.org/10.1029/2018GL078202).
- Grötzbach, G. (1983). “Spatial resolution requirements for direct numerical simulation of the Rayleigh-Bénard convection”. *J. Comput. Phys.* **49**(2). DOI: [10.1016/0021-9991\(83\)90125-0](https://doi.org/10.1016/0021-9991(83)90125-0).
- Hasselmann, K. (1976). “Stochastic climate models: Part I. Theory”. *Tellus* **28**(6). DOI: [10.3402/tellusa.v28i6.11316](https://doi.org/10.3402/tellusa.v28i6.11316).
- Irrgang, C., N. Boers, M. Sonnewald, E. A. Barnes, C. Kadow, J. Staneva, and J. Saynisch-Wagner (2021). “Towards neural Earth system modelling by integrating artificial intelligence in Earth system science”. *Nat. Mach. Intell.* **3**(8). DOI: [10.1038/s42256-021-00374-3](https://doi.org/10.1038/s42256-021-00374-3).
- Jones, P. W. (1999). “First- and second-order conservative remapping schemes for grids in spherical coordinates”. *Mon. Weather Rev.* **127**(9). DOI: [10.1175/1520-0493\(1999\)127<2204:FASOCR>2.0.CO;2](https://doi.org/10.1175/1520-0493(1999)127<2204:FASOCR>2.0.CO;2).
- Kooij, G. L., M. A. Botchev, E. M. A. Frederix, B. J. Geurts, S. Horn, D. Lohse, E. P. van der Poel, O. Shishkina, R. J. A. M. Stevens, and R. Verzicco (2018). “Comparison of computational codes for direct numerical simulations of turbulent Rayleigh-Bénard convection”. *Comput. Fluids* **166**. DOI: [10.1016/j.compfluid.2018.01.010](https://doi.org/10.1016/j.compfluid.2018.01.010).
- Kwasniok, F. (2012). “Data-based stochastic subgrid-scale parametrization: an approach using cluster-weighted modelling”. *Phil. Trans. R. Soc. A* **370**(1662). DOI: [10.1098/rsta.2011.0384](https://doi.org/10.1098/rsta.2011.0384).
- Lamb, D. (2003). “Cloud microphysics”. *Encyclopedia of Atmospheric Sciences*. Ed. by J. R. Holton. Oxford: Academic Press. ISBN: 978-0-12-227090-1. DOI: [10.1016/B0-12-227090-8/00111-1](https://doi.org/10.1016/B0-12-227090-8/00111-1).
- Le Quéré, P. (1991). “Accurate solutions to the square thermally driven cavity at high Rayleigh number”. *Comput. Fluids* **20**(1). DOI: [10.1016/0045-7930\(91\)90025-D](https://doi.org/10.1016/0045-7930(91)90025-D).
- Lord Rayleigh (1916). “On convection currents in a horizontal layer of fluid, when the higher temperature is on the under side”. *Philos. Mag.* **32**(192). DOI: [10.1080/14786441608635602](https://doi.org/10.1080/14786441608635602).
- Lorenz, E. N. (1995). “Predictability: a problem partly solved”. Seminar on Predictability (Shinfield Park, Reading). ECMWF. URL: <https://www.ecmwf.int/en/elibrary/75462-predictability-problem-partly-solved>.
- Manabe, S., J. Smagorinsky, and R. F. Strickler (1965). “Simulated climatology of a general circulation model with a hydrologic cycle”. *Mon. Weather Rev.* **93**(12). DOI: [10.1175/1520-0493\(1965\)093<0769:SC0AGC>2.3.CO;2](https://doi.org/10.1175/1520-0493(1965)093<0769:SC0AGC>2.3.CO;2).
- McFarlane, N. (2011). “Parameterizations: representing key processes in climate models without resolving them”. *WIREs Clim. Change* **2**(4). DOI: [10.1002/wcc.122](https://doi.org/10.1002/wcc.122).
- Monin, A. S. and A. M. Yaglom (2007). *Statistical fluid mechanics: mechanics of turbulence*. Ed. by J. L. Lumley. Vol. 1. Mineola, New York: Dover Publications. ISBN: 0-486-45883-0.
- Ouertatani, N., N. Ben Cheikh, B. Ben Beya, and T. Lili (2008). “Numerical simulation of two-dimensional Rayleigh-Bénard convection in an enclosure”. *C.R. Mec.* **336**(5). DOI: [10.1016/j.crme.2008.02.004](https://doi.org/10.1016/j.crme.2008.02.004).
- Palmer, T. N. (2001). “A nonlinear dynamical perspective on model error: a proposal for non-local stochastic-dynamic parametrization in weather and climate prediction models”. *Q. J. R. Meteorolog. Soc.* **127**(572). DOI: [10.1002/qj.49712757202](https://doi.org/10.1002/qj.49712757202).
- (2019). “Stochastic weather and climate models”. *Nat. Rev. Phys.* **1**(7). DOI: [10.1038/s42254-019-0062-2](https://doi.org/10.1038/s42254-019-0062-2).

- Palmer, T., G. Shutts, R. Hagedorn, F. Doblas-Reyes, T. Jung, and M. Leutbecher (2005). “Representing model uncertainty in weather and climate prediction”. *Annu. Rev. Earth Planet. Sci.* **33**(1). DOI: [10.1146/annurev.earth.33.092203.122552](https://doi.org/10.1146/annurev.earth.33.092203.122552).
- Pope, S. B. (2000). *Turbulent flows*. Cambridge: Cambridge University Press. ISBN: 9780521598866. DOI: [10.1017/CB09780511840531](https://doi.org/10.1017/CB09780511840531).
- Rio, C., A. D. Del Genio, and F. Hourdin (2019). “Ongoing breakthroughs in convective parameterization”. *Curr. Clim. Change Rep.* **5**(2). DOI: [10.1007/s40641-019-00127-w](https://doi.org/10.1007/s40641-019-00127-w).
- Russell, F. P., P. D. Düben, X. Niu, W. Luk, and T. N. Palmer (2017). “Exploiting the chaotic behaviour of atmospheric models with reconfigurable architectures”. *Comput. Phys. Commun.* **221**. DOI: [10.1016/j.cpc.2017.08.011](https://doi.org/10.1016/j.cpc.2017.08.011).
- Scheel, J. D., M. S. Emran, and J. Schumacher (2013). “Resolving the fine-scale structure in turbulent Rayleigh-Bénard convection”. *New J. Phys.* **15**(11). DOI: [10.1088/1367-2630/15/11/113063](https://doi.org/10.1088/1367-2630/15/11/113063).
- Shishkina, O., R. J. A. M. Stevens, S. Grossmann, and D. Lohse (2010). “Boundary layer structure in turbulent thermal convection and its consequences for the required numerical resolution”. *New J. Phys.* **12**(7). DOI: [10.1088/1367-2630/12/7/075022](https://doi.org/10.1088/1367-2630/12/7/075022).
- Stevens, B. and S. Bony (2013). “What are climate models missing?” *Science* **340**(6136). DOI: [10.1126/science.1237554](https://doi.org/10.1126/science.1237554).
- Stevens, R. J. A. M., R. Verzicco, and D. Lohse (2010). “Radial boundary layer structure and Nusselt number in Rayleigh-Bénard convection”. *J. Fluid Mech.* **643**. DOI: [10.1017/S0022112009992461](https://doi.org/10.1017/S0022112009992461).
- Verzicco, R. and R. Camussi (1999). “Prandtl number effects in convective turbulence”. *J. Fluid Mech.* **383**. DOI: [10.1017/S0022112098003619](https://doi.org/10.1017/S0022112098003619).
- Wilks, D. S. (2005). “Effects of stochastic parametrizations in the Lorenz ’96 system”. *Q. J. R. Meteorolog. Soc.* **131**(606). DOI: [10.1256/qj.04.03](https://doi.org/10.1256/qj.04.03).
- (2011). *Statistical methods in the atmospheric sciences*. 3rd ed. International Geophysics. Amsterdam, Boston: Elsevier/Academic Press. ISBN: 0123850231.
- Wouters, J. and V. Lucarini (2012). “Disentangling multi-level systems: averaging, correlations and memory”. *J. Stat. Mech.: Theory Exp.* **2012**(03). DOI: [10.1088/1742-5468/2012/03/P03003](https://doi.org/10.1088/1742-5468/2012/03/P03003).
- (2013). “Multi-level dynamical systems: connecting the Ruelle response theory and the Mori-Zwanzig approach”. *J. Stat. Phys.* **151**(5). DOI: [10.1007/s10955-013-0726-8](https://doi.org/10.1007/s10955-013-0726-8).
- Yuval, J., P. A. O’Gorman, and C. N. Hill (2021). “Use of neural networks for stable, accurate and physically consistent parameterization of subgrid atmospheric processes with good performance at reduced precision”. *Geophys. Res. Lett.* **48**(6). DOI: [10.1029/2020GL091363](https://doi.org/10.1029/2020GL091363).
- Zacharuk, M., S. I. Dolaptchiev, U. Achatz, and I. Timofeyev (2018). “Stochastic subgrid-scale parametrization for one-dimensional shallow-water dynamics using stochastic mode reduction”. *Q. J. R. Meteorolog. Soc.* **144**(715). DOI: [10.1002/qj.3396](https://doi.org/10.1002/qj.3396).
- Zhang, G. J. (2002). “Convective quasi-equilibrium in midlatitude continental environment and its effect on convective parameterization”. *J. Geophys. Res.: Atmos.* **107**(D14). DOI: [10.1029/2001JD001005](https://doi.org/10.1029/2001JD001005).

*This page intentionally left blank*



## Appendix A

# Description of codebase and data

### A.1 Reproducibility

*This page intentionally left blank*

## Appendix B

# Details of numerical experiments

### B.1 Initial condition

The initial velocity field is specified in terms of the streamfunction (whose level curves are streamlines of the flow)

$$\psi_0(x, z) = 0.1 \sin(\pi x)(1 - (2z - 1)^2)^2,$$

from which

$$u_0(x, z) = -\frac{\partial \psi_0}{\partial z} \quad \text{and} \quad w_0(x, z) = \frac{\partial \psi_0}{\partial x}.$$

It is easily verified that these satisfy the required boundary conditions. The reason for using the streamfunction is that it guarantees a divergence-free velocity field ( $\partial u / \partial x + \partial w / \partial z = -\partial^2 \psi / \partial x \partial z + \partial^2 \psi / \partial z \partial x \equiv 0$ ).

The initial temperature field is

$$\theta_0(x, z) = \frac{1}{2}(1 - 2z)^9$$

plus a random perturbation at each grid point, drawn from the normal distribution with mean zero and height-dependent standard deviation  $(1 - (2z - 1)^2) \times 10^{-2}$ . The random perturbation is necessary to break the symmetry that the initial condition would otherwise have.

Figure B.1 shows the initial streamfunction, velocity and temperature fields. There are eight equally-sized counter-rotating convection cells.

### B.2 Model spin-up time

It is critical to ensure that the simulations have reached a statistically steady state (to “spin” them up) in order to accurately calculate long-term statistics. To that end, I have calculated the 250-time-unit rolling means of the Nusselt number, thermal boundary layer thickness, RMS speed, kinetic energy dissipation rate and thermal dissipation rate (see § 3.6 for the definitions) using the data from the  $1024 \times 128$  simulation described in § 3.6. These are plotted in Figure B.2. The RMS speed and kinetic energy dissipation rate take longer than the other variables to reach a steady state and exhibit larger low-frequency oscillations. Nonetheless, I determine that the simulation has reached a sufficiently steady state at  $t = 750$ .

In the resolution-dependence experiment of § 3.6, the simulations with resolution higher than  $1024 \times 128$  were initialised by interpolating the  $1024 \times 128$  solution at time  $t = 650$ . These also needed to reach a statistically steady state. Figure B.3 shows the 150-time-unit rolling means of the same quantities as Figure B.2 for the  $2048 \times 256$  simulation. I determine that the means reach a sufficiently steady state

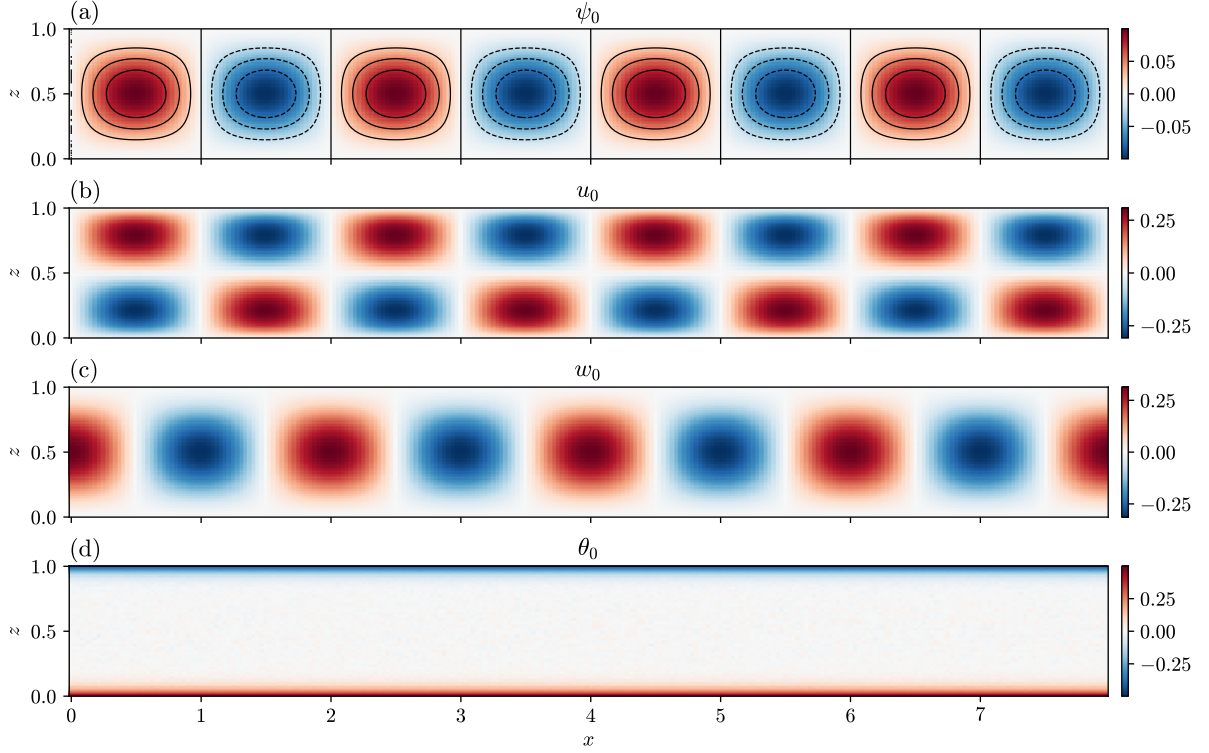


Figure B.1: Initial streamfunction (a), horizontal velocity (b), vertical velocity (c) and temperature (d).

when the right edge of the rolling window is at  $t = 900$ ; it is therefore appropriate to use the data from  $t = 900 - 150 = 750$  onwards.

### B.3 Training dataset snapshot frequency

When building the parametrisation training dataset for [Chapter 4](#), the amount of storage space demanded by the high-resolution model output made it desirable to avoid saving redundant information. Saving the model state every time step or every 0.2 time units (as in [§ 3.6](#)) would waste space and reduce the diversity of the training dataset, because the state exhibits substantial autocorrelation at these intervals. [Figure B.4](#) shows the spatially averaged temporal autocorrelation functions of  $u$ ,  $w$  and  $\theta$ , computed for the  $1024 \times 128$  simulation that is described in [§ 3.6](#) (again discarding the first 750 time units of data for spin-up). Observing that all three variables reach the first correlation minimum at a lag of approximately 3 time units, I choose to save the model output at this interval.

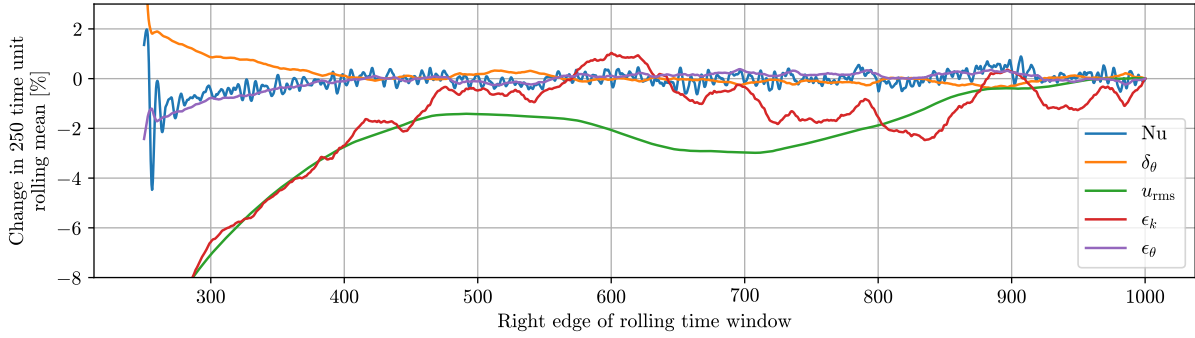


Figure B.2: 250-time-unit rolling means of the Nusselt number  $Nu$ , thermal boundary layer thickness  $\delta_\theta$ , RMS speed  $u_{rms}$ , kinetic energy dissipation rate  $\epsilon_k$  and thermal dissipation rate  $\epsilon_\theta$  for the  $1024 \times 128$  simulation described in § 3.6. The horizontal coordinate is the position of the *right* edge of the rolling window. Each quantity is expressed as a percentage deviation relative to its value when the right edge of the window is at  $t = 1000$ .

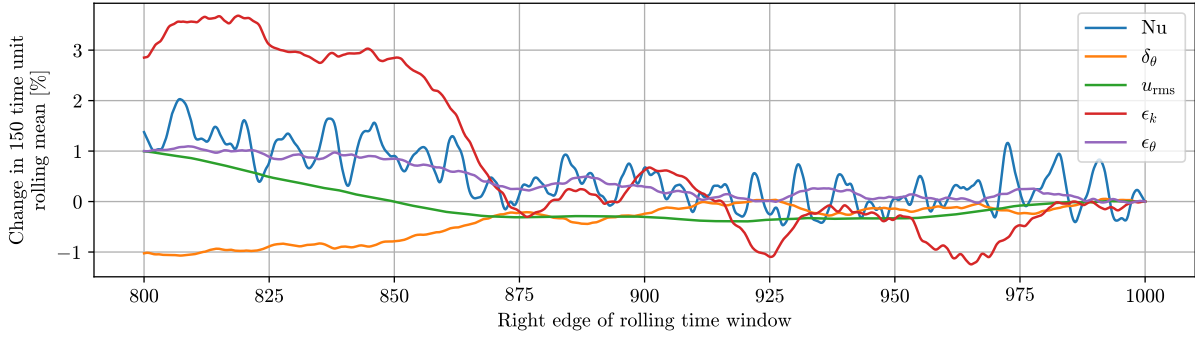


Figure B.3: Similar to Figure B.2, but showing 150-time-unit rolling means for the  $2048 \times 256$  simulation, which was initialised by interpolating the  $1024 \times 128$  solution at time  $t = 650$ .

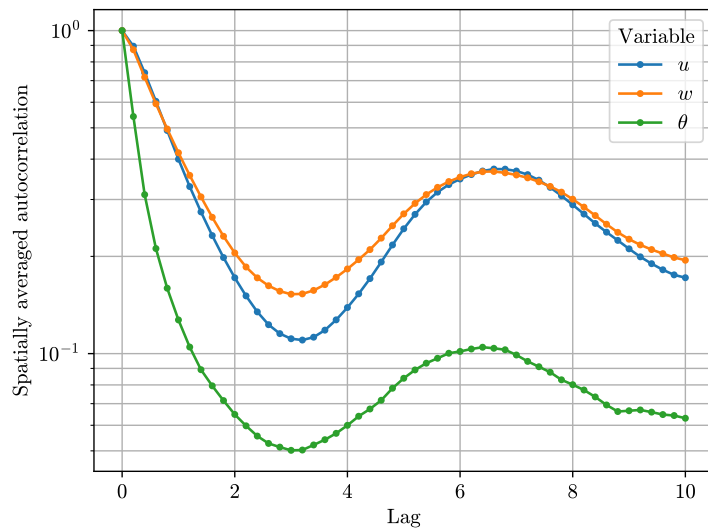


Figure B.4: Spatially averaged autocorrelation functions of the three prognostic variables in the  $1024 \times 128$  simulation that is described in § 3.6.