

# Towards transparent data-driven parametrisations for weather and climate modelling

## Honours Research Project Proposal

Thomas D. Schanzer

Supervisor: Prof. Steven Sherwood

School of Physics

Climate Change Research Centre and ARC Centre of Excellence for Climate Extremes  
University of New South Wales, Sydney, Australia

### Abstract

Constrained by computational resources, numerical models of Earth's atmosphere and oceans are unable to resolve important fine-scale dynamical features. The effect of the unresolved behaviour on the coarse state of the models must be estimated in a process known as parametrisation, but deficiencies in traditional parametrisation schemes inhibit accurate forecasting and climate prediction. The proposal for this Honours project is to help bridge the gap between parametrisation methods that have shown promise in simple toy models and the ultimate goal of generalising them to produce rigorous, accurate, robust and transparent schemes in full-scale Earth models. More specifically, the project aims to parametrise the subgrid-scale dynamics of 2D Rayleigh-Bénard convection, an intermediate-complexity fluid problem, using promising methods from toy model research. By evaluating the performance of these schemes, the project will identify their strengths and weaknesses and determine how well they generalise to more complex models.

## Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Background . . . . .	1
1.2	Brief literature review . . . . .	2
1.3	Proposed direction for research . . . . .	3
<b>2</b>	<b>Proposed aims and methods</b>	<b>4</b>
2.1	Chosen dynamical system . . . . .	4
2.2	Experiments . . . . .	5
2.3	Analysis . . . . .	6
2.4	Potential pitfalls and contingency plans . . . . .	6

# 1 Introduction

## 1.1 Background

The Earth system (including atmosphere, ocean and land) is distinguished both in its complexity and its influence on all terrestrial life. If the wellbeing of humanity in particular is to be preserved, it is difficult to overstate the importance of understanding and predicting this system's behaviour—both short-term weather and long-term climate. This understanding and predictive skill informs important decisions and government policies that have the potential to reduce our vulnerability to extreme events (e.g., floods, droughts, fires, tropical cyclones) and long-term climate change, and limit negative human impacts (e.g., greenhouse gas emissions) to sustainable levels.

A significant part of our understanding and predictive skill is derived from numerical modelling of the Earth system. Like many other models, the aim is to predict the time evolution of an initial state (e.g., pressure, temperature, wind velocity in the atmosphere) given a set of boundary conditions and external forcings. Unfortunately, there are many obstacles to accurate modelling. Arguably the most fundamental of these is chaos: even with a perfect model and unlimited computing resources, arbitrarily small differences in initial conditions grow exponentially. This constrains short-term predictability. Second, the Earth system comprises a vast number of interacting components, such as water in all three phases, solar radiation and clouds (just to name a few which are relevant to atmospheric modelling). In other words, the system is high-dimensional. Third, the dynamics occur on a wide spectrum of spatial and temporal scales. These range from large-scale, slowly-evolving motions such as ocean gyres and the atmospheric Hadley circulation, to synoptic-scale, transient weather systems and ocean eddies, to small-scale wind gusts, tornadoes, water waves and turbulence. All scales and their cross-interactions influence the overall dynamics.

Atmosphere and ocean models solve the differential equations that govern fluid flow with a finite spatial and temporal resolution and are therefore only able to resolve behaviour whose scale is of the same order of magnitude as the resolution or larger. The achievable resolution is constrained by the capabilities, availability and cost of modern computing resources. A typical atmospheric global climate model has a spatial resolution on the order of  $1^\circ$  latitude/longitude, roughly corresponding to the size of the entire Greater Sydney area from Katoomba to Bondi. Sub-grid scale features, such as individual clouds, cannot be explicitly resolved, but to ignore them completely would introduce unacceptable biases (inaccuracies, relative to observations) in the model output. The same is true for the numerous components of the Earth system that influence the fluid dynamics but are not directly predicted by the fluid equations, such as solar radiation and land interactions (which include moisture and heat fluxes from vegetation, soil and water bodies).

The process of using the information available in the model to estimate the effect of these unresolved processes on the coarse-scale variables is known as *parametrisation*, which will be the topic of the project. A general mathematical picture of parametrisation can be developed by considering a system with some degrees of freedom “hidden” from the observer, not resolvable by the model. In general, the evolution of the resolvable and unresolvable degrees of freedom is governed by a set of coupled PDEs with the generic form

$$\partial_t(\text{resolved variable}) = S_r(\text{resolved variables}) + C_r(\text{unresolved variables}) \quad (1a)$$

$$\partial_t(\text{unresolved variable}) = S_u(\text{unresolved variables}) + C_u(\text{resolved variables}) \quad (1b)$$

where  $S_r, S_u, C_r, C_u$  are (possibly nonlinear) operators containing spatial derivatives. The  $S$  terms are self-interactions: the influences of the resolved and unresolved variables on themselves. The  $C$  terms are couplings: the influence of the resolved behaviour on the unresolved behaviour and vice versa.

A model cannot solve (1) because it has no information about the unresolved variables. The goal of parametrisation is to derive a closed (i.e., well-posed) system involving only the resolved variables that still yields an accurate solution for those variables. The model is thus described by

$$\partial_t(\text{resolved variable}) = S_r(\text{resolved variables}) + P(\text{resolved variables}), \quad (2)$$

where  $P$  is the *parametrised tendency*: an estimate of the unseen coupling term  $C_r$  in (1a) based solely on the values of the resolved variables. The study of parametrisation is essentially concerned with the construction of the optimal  $P$ .

## 1.2 Brief literature review

Traditional parametrisations are often based on heavily simplified conceptual models of the processes in question. For example, the well-known Community Atmosphere Model v5 (Neale et al. 2010) parametrises atmospheric convection by considering an ensemble of rising updraft plumes, making assumptions about their mass flux and initiation conditions (Zhang and McFarlane 1995). It is now known that many existing parametrisation schemes have biases that do not reflect reality. For example, many precipitation schemes suffer from the “drizzle problem”: they overestimate the frequency of precipitation and underestimate its intensity, failing to capture its variance and extreme values (Shamekh et al. 2022). These biases limit the certainty of future climate projections and our ability to understand how extreme precipitation, for example, might change in the future.

A significant amount of work has been done to make parametrisation schemes more rigorous and complete. A key development has been the introduction of *data-driven* techniques, which construct parametrisations by fitting statistical models to a reliable dataset rather than primarily using physically-motivated conceptual models. One of the most intensely studied data-driven techniques in contemporary research is machine learning, typically employing algorithms to “learn” how the sub-grid scale behaviour relates to the resolved variables from high-resolution simulation data. For example, Shamekh et al. (2022) trained an autoencoder on high-resolution precipitable water data to produce a measure of convective organisation (the aggregation of convective clouds). By training a neural network to predict this organisation metric from the coarse-grid variables, they were able to capture the statistical distribution of precipitation intensity, especially extremes, much more accurately.

While parametrisation schemes are often tested in full-scale atmospheric models—*in situ*—the sheer scale of these models and the amount of data they produce can make it difficult to understand where their strengths and weaknesses lie. In particular, machine learning techniques, which have no basis in physical intuition, are notoriously opaque (they are often called “black boxes”). For example, Shamekh et al. (2022) had to perform a rather detailed analysis to prove that the output of their autoencoder was indeed a measure of convective organisation. This has prompted some researchers to use very simple toy models, rather than full atmospheric models, as test beds for the construction and analysis of new, more transparent data-driven parametrisation techniques. Another advantage of a simple model is that it can be solved with high numerical precision, providing a reliable “truth” dataset for data-driven techniques. In contrast, data-driven parametrisation for full atmospheric models requires either observational data, which has inherent uncertainties, or output from high-resolution simulations, which are not perfect representations of the atmosphere either.

Two important parametrisation elements have been developed with the aid of toy models. The first is stochasticity: the introduction of some randomness in the parametrised tendencies to reflect the fact that the large-scale state does not uniquely determine the fine-scale state. The second is memory: the dependence of the parametrised tendency on its own value at previous times, or the values of the resolved variables at previous times.

One of the seminal and most widely cited toy model parametrisation studies is that of Wilks (2005). Wilks used the so-called Lorenz ’96 model (Lorenz 1995), which was devised by renowned meteorologist Edward Lorenz. It matches the form of (1) by construction, having a set of slowly-varying “resolved” variables coupled to a larger set of rapidly-varying “unresolved” variables. Wilks first numerically solved the full coupled system as “truth” and used this data to retrospectively calculate the unresolved tendencies ( $C_r$  in (1a)). He then performed a simple polynomial regression of the unresolved tendencies against the resolved variables, as shown in Figure 1 (a reproduction of his Figure 2).

Wilks evaluated three parametrisation schemes. First, the polynomial regression alone was used as a simple deterministic parametrisation (i.e., the regression model was used as  $P$  in (2)). The second

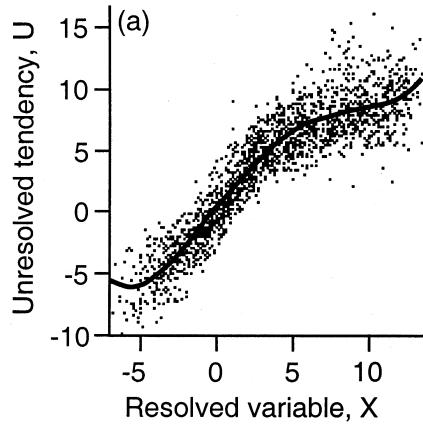


Figure 1: A reproduction of Figure 2 by Wilks (2005), showing a scatter plot of the unresolved tendencies against the values of the resolved variables in the Lorenz '96 model, with the polynomial regression model plotted as a black line.

scheme added independent Gaussian (“white”) noise to the predictions of the regression model in order to reflect the range of possible unresolved states for a given resolved state (which manifests as a scatter around the curve in Figure 1). The third scheme made the noise non-independent by introducing non-zero *temporal autocorrelation*; that is, the value of the noise at time step  $n$  is drawn from a Gaussian distribution whose mean depends on the value of the noise at time step  $n - 1$ . This type of noise is “red”, and such a scheme is technically known as a first order autoregressive or AR(1) model. Wilks found that the third scheme (with both stochasticity and memory) was the most accurate in reproducing the long-term average distributions of the resolved variables.

In the years following Wilks’ paper, other more sophisticated parametrisation schemes have been tested. Crommelin and Vanden-Eijnden (2008), for example, implemented a stochastic scheme with memory based on conditional Markov chains, also for the Lorenz '96 model. The Markov chain method assumes that the unresolved tendencies can take only a finite set of values, or “states”, with transitions between the states governed by a stochastic process. The authors found that this scheme outperformed Wilks’ AR(1) scheme in several respects. In summary, there is clear agreement in the literature that memory and stochasticity should be included in parametrisations.

### 1.3 Proposed direction for research

Despite the successes of toy model parametrisation studies, a common shortcoming—and one that many authors have identified—is that they provide limited information about how efficiently and effectively the techniques would scale in real weather/climate models (e.g., Wilks 2005, Crommelin and Vanden-Eijnden 2008). This project aims to find a middle ground by testing parametrisation techniques that have been shown to work in toy models on a more realistic model that solves a real fluid problem.

Another consideration that seems to be unaddressed in the toy model literature is the possibility for the dynamical system being modelled (e.g., the atmosphere) to have *inherent* stochasticity or memory, even before parametrisation. These phenomena could arise if some degrees of freedom are completely unknown to us, or if our physical understanding of the system is incomplete (as is the case for the atmosphere). The project will therefore also seek to determine the level of predictability that is achievable with different parametrisation schemes under these circumstances.

## 2 Proposed aims and methods

### 2.1 Chosen dynamical system

This project will study parametrisation in the context of turbulent two-dimensional Rayleigh-Bénard convection, which is the flow of a fluid between two infinite flat plates of constant temperature, the bottom plate made warmer than the top plate. This may be regarded as a simpler, more computationally tractable analogue of true atmospheric convection whose fine-scale turbulent behaviour will still be challenging to parametrise. While the (Navier-Stokes) equations that govern the flow are omitted here, I will use a representation where the dynamical variables are the vorticity  $\omega$  (curl of the velocity), temperature  $T$  and streamfunction  $\psi$  (a scalar function whose level curves are streamlines of the flow). [Figure 2](#) shows a typical temperature plot for this type of flow.

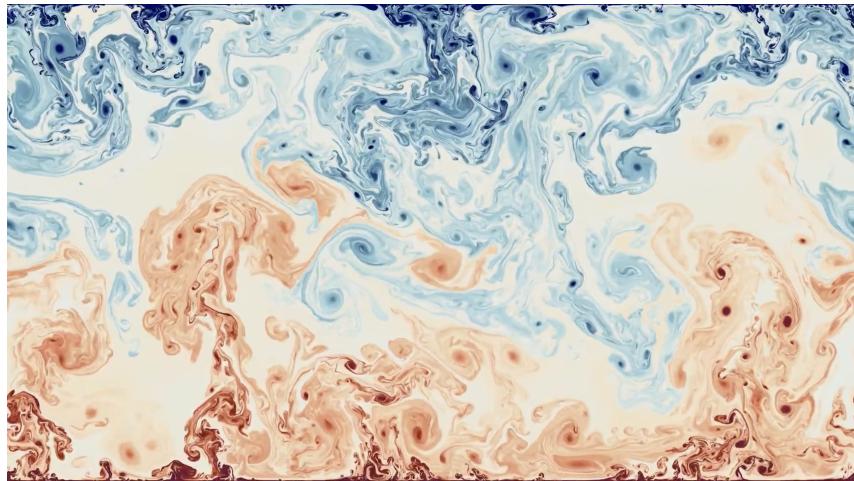


Figure 2: Temperature plot for a simulation of 2D Rayleigh-Bénard convection (red is hot, blue is cold). Courtesy of Stephan Lenz (2018), published on YouTube under the Creative Commons CC BY licence (<https://youtu.be/BJKiuwpdprQ>).

In its usual form, the system of PDEs for the flow does not distinguish resolved dynamics from unresolved dynamics like (1). Rather, the dynamical variables  $\omega, T, \psi$  exhibit a continuous spectrum of spatial scales, from large-scale overturning circulations to small eddies. Fortunately, Zacharuk et al. (2018) demonstrated a method for decomposing the shallow water equations, a system of PDEs with one spatial dimension, into two coupled sets of equations: one set for coarse-scale resolved behaviour and one for fine-scale unresolved behaviour. I have generalised the method for 2D Rayleigh-Bénard convection, allowing the application of the same parametrisation schemes that have been used for toy models like Lorenz '96 in the literature.

The decomposition assumes that numerical solutions are to be obtained using the finite difference method. The coarse temperature variable (for example), denoted  $\bar{T}$ , is defined by partitioning the gridded domain into  $n$ -by- $n$  tiles (for some integer  $n$ ) and averaging the  $n^2$  temperature values within each tile. This is equivalent to top-hat filtering the temperature field. The corresponding fine variable  $T'$  is the temperature anomaly at each grid point relative to the average value in the tile. Similar definitions are adopted for the vorticity  $\omega$  and streamfunction  $\psi$ . [Figure 3](#) shows  $n = 4$  as an example; the dashed red lines partition the domain into 4-by-4 tiles, with the coarse averages defined on the red grid and the fine anomalies defined on the black grid.

The end result, I have shown, is that the coarse and fine variables satisfy a coupled system of difference equations that can be expressed in the same form as (1). The goal will then be to derive an approximate model with the form of (2) for the coarse dynamics, testing several methods for constructing  $P$  that have been established for simpler toy models.

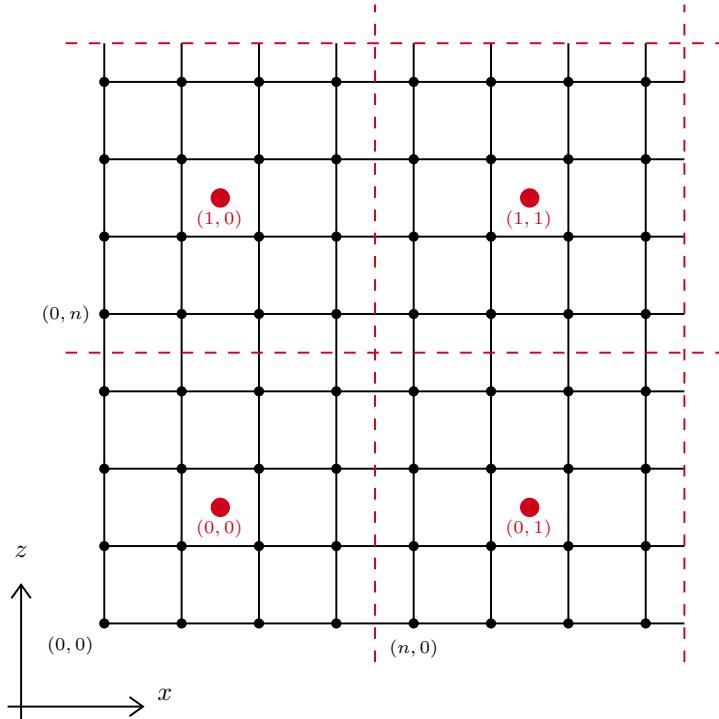


Figure 3: The grids for the coarse-scale “resolved” variables (red) and the fine-scale “unresolved” variables (black).

## 2.2 Experiments

Before parametrisation testing can begin, the first step will be to implement a numerical solution algorithm for the coupled difference equations that have already been derived. Python, standard in the geosciences community, will be the programming language of choice. Resource-intensive computations will be done on the Gadi supercomputer at the National Computational Infrastructure in Canberra if necessary. The next step will be to integrate the full coupled system of coarse and fine variables to obtain “truth” datasets. Part of this data (e.g., half) will be set aside as a benchmark against which the parametrised solutions will be evaluated. The remainder will be used as training data to construct the parametrisations.

The project aims to test a variety of schemes that have been established in the literature, especially those that have been tested on toy models but whose scalability to real models is unclear. These will include the three methods of Wilks (2005): deterministic, stochastic without memory and stochastic with memory. I will also test a judicious selection of more advanced methods, such as conditional Markov chains (following Crommelin and Vanden-Eijnden (2008)) or even machine learning. Of course, control runs with no parametrisation (that ignore the existence of the fine variables entirely) will be used as baselines to measure the degree of improvement that is achievable. The project will also aim to determine the optimal settings for these schemes, such as the magnitude of the noise terms in stochastic schemes and the number of previous time steps considered by memory-based schemes, with the objective of balancing accuracy and computational cost.

The project will additionally test the schemes in the case where the underlying Rayleigh-Bénard system has inherent stochasticity and/or memory. Stochasticity could be implemented via a random forcing of the fine variables (i.e., an additional term in (1b)). Memory could be created by introducing a dependence of the fine variables on their values at previous time steps (another term in (1b)).

## 2.3 Analysis

The approximate solutions derived from the parametrised models will be evaluated against the testing data (previously set aside from the training data) using appropriate metrics. The choice of metric will depend on whether accurate short-term forecasting or long-term “climate” prediction is prioritised (both options will be considered as they are equally important in real-world modelling). A simple metric for forecast accuracy is the root-mean-square error between the parametrised solution and the truth. Long-term climate prediction performance can be assessed by comparing the moments (mean, variance, skewness, etc.) of the long-term distributions of the coarse variables. The closeness of the distribution functions themselves may also be measured using the Kolmogorov-Smirnov statistic. All the aforementioned metrics were used by Wilks (2005) for the Lorenz ’96 model.

The envisioned end result is a rigorous and quantitative assessment of the performance of each parametrisation scheme and how well it generalises from simple toy models to real fluid models. This would enable identification of strengths and weaknesses for each scheme, leading to recommendations on scheme choice and settings for different modelling scenarios (short-term forecasting vs. climate prediction, computational resource constraints, etc.). Given the realism of the Rayleigh-Bénard system, this work could potentially inform choices in parametrisation development in full atmosphere/ocean models.

## 2.4 Potential pitfalls and contingency plans

The success of the project depends primarily on the proper functioning of the proposed numerical methods. Table 1 summarises the possible pitfalls and suggests possible solutions.

## References

- Crommelin, D. and E. Vanden-Eijnden (2008). “Subgrid-Scale Parameterization with Conditional Markov Chains”. *J. Atmos. Sci.* **65**(8), 2661–2675. doi: [10.1175/2008JAS2566.1](https://doi.org/10.1175/2008JAS2566.1).
- Lorenz, E. N. (1995). “Predictability: a problem partly solved”. Seminar on Predictability, 4-8 September 1995. ECMWF, pp. 1–18.
- Neale, R. B., C.-C. Chen, A. Gettelman, P. H. Lauritzen, S. Park, D. L. Williamson, A. J. Conley, R. Garcia, D. Kinnison, J.-F. Lamarque, D. Marsh, M. Mills, A. K. Smith, S. Tilmes, F. Vitt, H. Morrison, P. Cameron-Smith, W. D. Collins, M. J. Iacono, R. C. Easter, S. J. Ghan, X. Liu, P. J. Rasch, and M. A. Taylor (2010). *Description of the NCAR Community Atmosphere Model (CAM 5.0)*. NCAR/TN-486+STR. NCAR Technical Note.
- Shamekh, S., K. D. Lamb, Y. Huang, and P. Gentine (2022). “Implicit learning of convective organization explains precipitation stochasticity”. Preprint. doi: [10.1002/essoar.10512517.1](https://doi.org/10.1002/essoar.10512517.1).
- Wilks, D. S. (2005). “Effects of stochastic parametrizations in the Lorenz ’96 system”. *Q. J. R. Meteorol. Soc.* **131**(606), 389–407. doi: [10.1256/qj.04.03](https://doi.org/10.1256/qj.04.03).
- 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. Meteorol. Soc.* **144**(715), 1975–1990. doi: [10.1002/qj.3396](https://doi.org/10.1002/qj.3396).
- Zhang, G. J. and N. A. McFarlane (1995). “Sensitivity of climate simulations to the parameterization of cumulus convection in the Canadian Climate Centre general circulation model”. *Atmos. Ocean* **33**(3), 407–446. doi: [10.1080/07055900.1995.9649539](https://doi.org/10.1080/07055900.1995.9649539).

---

Pitfall	Solution
Delays caused by code bugs	<ul style="list-style-type: none"> <li>• Use of existing Python packages wherever possible</li> <li>• Expert consultation (e.g., the Computational Modelling Systems team at the ARC Centre of Excellence for Climate extremes)</li> <li>• Use of simpler alternative numerical methods</li> </ul>
Difficulties in implementing parametrisation schemes for the 2D Rayleigh-Bénard problem	<ul style="list-style-type: none"> <li>• Use of existing code, obtained from the authors of existing papers, if possible</li> <li>• Expert consultation (e.g., American mathematician Scott Hottovy visiting UNSW in April, authors of existing papers)</li> <li>• Use of simpler schemes</li> <li>• Reverting to a simpler dynamical system (e.g., a 1D PDE or even the Lorenz '96 model)</li> </ul>
Model crashes and numerical instability	<ul style="list-style-type: none"> <li>• Expert consultation</li> <li>• Further reading on stability criteria</li> <li>• Use of simpler numerical methods</li> </ul>
Excessive computational cost	<ul style="list-style-type: none"> <li>• Parallel computing (e.g., using the <code>dask</code> and <code>multiprocessing</code> packages in Python)</li> <li>• Reduction of domain size and simulation length</li> <li>• Reduction of spatial and temporal resolution</li> <li>• Use of more efficient numerical methods (further reading required)</li> </ul>

---

Table 1: Potential pitfalls for the project and proposed solutions.