A probabilistic framework for learning non-intrusive corrections to long-time climate simulations from short-time training data

B. Barthel Sorensen[1], L. Zepeda-Núñez[2,3],I. Lopez-Gomez[2], Z. Y. Wan[2], R. Carver[2], F. Sha[2], and T. P. Sapsis[1]

[1]Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]Google Research, 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA
[3]University of Wisconsin-Madison, 480 Lincoln Drive, Madison, WI 53706, USA

Key Points:

- We present a probabilistic framework for debiasing coarse-resolution climate simulations using machine learning
- The method accurately predicts the risk of events with return periods far longer than the training period
- The method leverages probabilistic machine learning architectures to provide built-in uncertainty quantification

Corresponding author: B. Barthel Sorensen, `bbarthel@mit.edu`

Abstract

Despite advances in high performance computing, accurate numerical simulations of global atmospheric dynamics remain a challenge. The resolution required to fully resolve the vast range scales as well as the strong coupling with – often not fully-understood – physics renders such simulations computationally infeasible over time horizons relevant for long-term climate risk assessment. While data-driven parameterizations have shown some promise of alleviating these obstacles, the scarcity of high-quality training data and their lack of long-term stability typically hinders their ability to capture the risk of rare extreme events. In this work we present a general strategy for training variational (probabilistic) neural network (NN) models to non-intrusively correct under-resolved long-time simulations of turbulent climate systems. The approach is based on the paradigm introduced by Barthel Sorensen et al. (2024) which involves training a post-processing correction operator on under-resolved simulations nudged towards a high-fidelity reference. Our variational framework enables us to learn the dynamics of the underlying system from very little training data and thus drastically improve the extrapolation capabilities of the previous deterministic state-of-the art – even when the statistics of that training data are far from converged. We investigate and compare three recently introduced variational network architectures and illustrate the benefits of our approach on an anisotropic quasi-geostrophic flow. For this prototype model our approach is abe to not only accurately capture global statistics, but also the anistropic regional variation and the statistics of multiple extreme event metrics – demonstrating significant improvement over previously introduced deterministic architectures.

Plain Language Summary

We present a probabilistic framework to build and train machine learned (ML) correction operators to improve the predicted statistics of low-resolution climate simulations. The proposed methodology is specifically focused on enabling long-time climate predictions using operators trained on short-time data. We illustrate our approach, which acts on existing data in a post-processing manner, on a prototype climate model, for which we are able to accurately quantify the regionally varying statistics as well as rare-event statistics over the previous state-of-the-art. The simple model we consider here allows us to demonstrate our method on very long simulation, but our method can be readily applied to output from full-complexity climate models.

1 Introduction

As the Earth's climate changes, we are faced with deep uncertainty about extreme weather events whose frequency and magnitude are expected to increase (Lehmann et al., 2015; Meehl & Tebaldi, 2004; Perkins-Kirkpatrick & Lewis, 2020). Due to their potential for catastrophic consequences, it is crucial to accurately quantify their long-term risk and assess their impact on communities (Fischer et al., 2021; Raymond et al., 2020; Robinson et al., 2021). In this context, "extreme events" are generally defined as high amplitude anomalies of high-impact variables, such as near-surface temperature and precipitation (Lucarini et al., 2016; Sapsis, 2021), to which human activities are highly sensitive. For instance, heatwaves can have devastating effects on an unprepared population, particularly when compounded with other events such as low rainfall (Bevacqua et al., 2023; Raymond et al., 2020; Robinson et al., 2021; Zscheischler et al., 2018).

From a statistical point of view, certain observables being susceptible to extreme events implies that their probability density functions (pdfs) have "heavy tails", i.e. they decay slowly and high amplitude events retain small but non-negligible probability. Accurately quantifying the risk of such events is subject to two main requirements: first, high-fidelity simulations that can capture the dynamics of interest, which require a high-resolution mesh in space and time; and second, sufficiently large samples to capture rare events in the tail of the distribution. The latter can be obtained either through long-term simulations or through large simulation ensembles (Deser et al., 2012). However, due to the high-dimensional, chaotic, and multi-scale nature of Earth's atmosphere, large ensembles of high-resolution simulations are computationally intractable over multi-decade or multi-century time horizons. As an example, the highest resolution climate models currently proposed fall short of fully resolving all the spatial scales of atmospheric turbulence by a factor of 10^{17} degrees of freedom (Schneider et al., 2023). These shortfalls are further compounded by the need to simulate centuries-long trajectories for climate risk assessment.

Alternative surrogate machine learning (ML) techniques are becoming increasingly attractive as a computationally efficient way to simulate the Earth's atmosphere (Pathak et al., 2022a). Alas, purely data-driven models present their own set of challenges. In contrast to dynamical models, they are often unstable when run over long time-horizons, and they struggle to extrapolate beyond the distribution defined by the scarce training data. This becomes problematic as we seek to quantify climate risks over the coming centuries with only several decades of observational data available for training. Although methodologies have been proposed to circumvent these issues by exploiting properties of the underlying dynamical system (Kochkov et al., 2023; Mathews et al., 2021), most of them require an explicit notion of ergodicity (Z. Li, Liu-Schiaffini, et al., 2022; Jiang et al., 2023; Platt et al., 2023; Schiff et al., 2024), or scale poorly as the state dimension increases (Pathak et al., 2017; Bollt, 2021; Hara & Kokubu, 2022), posing challenges for their use in climate-related applications.

These limitations have spurred a complementary line of research in which hybrid strategies are explored (Schneider, Lan, et al., 2017; Schneider et al., 2023; Eyring et al., 2024; Lam et al., 2022; Bi et al., 2023; Kochkov et al., 2023). Such methods seek to inherit the desirable properties of both numerical and ML models, while attenuating their drawbacks. A group of methods in this category focuses on correcting the dynamics *on-the-fly* by *intrusively* modifying classical numerical models (Arcomano et al., 2022; Clark et al., 2022; Sanderse et al., 2024; Boral et al., 2023). The underlying dynamical model provides a strong inductive bias, which reduces the training data requirements compared to purely-data driven models, and helps capture many of the dynamical properties of the system (Kochkov et al., 2021; Dresdner et al., 2022). In the context of climate modeling, recent approaches seek to learn state-dependent closure terms for the effect of the unresolved sub-grid-scale processes on the resolved scales. Such approaches have been shown to be effective in both reducing overall bias (Watt-Meyer et al., 2021; Guillaumin

& Zanna, 2021) as well as capturing unresolved processes (Arcomano et al., 2023). Furthermore, they have been demonstrated on a range of systems ranging from idealized aquaplanet configurations (Yuval & O'Gorman, 2020; Rasp et al., 2018; Brenowitz & Bretherton, 2019; Yuval et al., 2021; Iglesias-Suarez et al., 2024) to more realistic global climate models (Bora et al., 2023; Bretherton et al., 2022).

However, these intrusive ML corrections have several drawbacks. Their implementation requires integration into the original dynamical model, which can be a complex process (J. McGibbon et al., 2021). Furthermore, these closure terms are typically learned offline, without interaction with the dynamical system, since very few dynamical models meet the fast differentiation requirements for integrated online learning (Kochkov et al., 2023). In addition, although advances have been made in stabilizing such hybrid models, long-term instability can still be an issue (H. Zhang et al., 2021; Wikner et al., 2022; Yuval et al., 2021). Gradient-free ensemble Kalman methods have recently been proposed that enable online learning in hybrid systems (Lopez-Gomez et al., 2022; Christopoulos et al., 2024). These methods can learn from long-term statistics to guarantee stability, but their application is limited to relatively sparse ML corrections.

Another group of hybrid methods focuses on machine learning *non-intrusive* corrections, meant to be applied as a post-processing step. Since there is no interaction with the numerical solver, these methods are long-term stable by design. Post-processing methods apply a machine learned map to biased trajectories of the dynamical system such that the statistics of the output match those of the training data. The need to train on statistics rather than trajectories is necessitated by the chaotic nature of the underlying system and the absence of paired (or aligned) data available for training. Such techniques have been applied to coarse resolution weather and climate simulations in the context of statistical debiasing (Blanchard et al., 2022; J. J. McGibbon et al., 2023; L. Li et al., 2024) and downscaling (Vandal et al., 2017; Wan, Baptista, Chen, et al., 2023; Wilby et al., 1998). In the context of debiasing, multiple methods have been explored in the literature including generative models based on optimal transport theory (Arbabi & Sapsis, 2022), temporal-convolutional-network (TCN) and LSTM networks (Blanchard et al., 2022), generative adversarial networks (GAN) (J. J. McGibbon et al., 2023), unsupervised image-to-image networks (UNIT) (Fulton et al., 2023), and diffusion models (L. Li et al., 2024). However, the requirement to reproduce the statistics of the training data greatly limits the potential of such methods to generalize to longer trajectories than those observed in training.

To tackle such limitation, we propose post-processing methodology to debias coarse-resolution climate simulations that is able to correct statistics of rare extreme events even when these have return periods far longer than the period spanned by the training dataset. Our proposed methodology seeks to extend the application of trajectory-based post-processing methods to long-time simulations through the use of probabilistic neural network models trained on specific paired sets of training data. Specifically, our framework leverages a recently developed methodology to generate paired climate trajectories (Barthel Sorensen et al., 2024; S. Zhang et al., 2024) that avoids common pitfalls of training ML algorithms for chaotic systems. These paired trajectories are then used to learn a probabilistic post-processing operator using variational inference methods.

Variational inference methods seek to approximate a distribution using its samples by solving an optimization problem where the distribution itself is parameterized by a neural network. In this case, we leverage Variational Auto-Encoders (VAEs) (Kingma & Welling, 2022) coupled with Long-Short-Term-Memory based recurrent neural network (RNN) architectures (Hochreiter & Schmidhuber, 1997) and ensemble learning (Opitz & Maclin, 1999). VAEs compress the system's state into a probabilistic latent representation whose distribution is learned variationally. RNNs map one trajectory to another by processing snapshots sequentially using a latent representation of the current and previous states. By replacing the latent representation in RNNs by a probabilistic one learned

variationally one obtains a map from a trajectory to a distributions of *plausible* trajectories. Furthermore, we train a small ensemble of such networks using the same data and different random seeds. Thus, the final algorithm defines a map from a trajectory to a *composite distribution of trajectories*, which captures the uncertainty of the system more accurately, in contrast to deterministic models that tend to learn the expectation.

Our variational extension greatly increases the generalization and extrapolation capabilities of deterministic models used in previous work (Barthel Sorensen et al., 2024). This allows us to accurately predict the probability of tail-risk events with longer return periods than the training period, and which are therefore likely to be missing entirely from the training data. Furthermore, we illustrate the advantages of our framework on a range of metrics not previously considered, including two-point correlations, regional variation, and extreme event statistics. We also conduct a systematic comparison of several variational architectures to serve as a guide to researchers looking to implement our framework. In summary, our approach bypasses the three main difficulties encountered by many ML-based surrogates for chaotic systems, namely: long time inference stability, generalizability, and training stability. Our approach is stable for indefinitely long time horizons by construction, sample efficient, easy to implement, and empirically able to extrapolate statistically relevant properties. In this work we apply our methodology on an anisotropic 2D quasi-geostrophic flow, which, albeit simple, captures many of the core difficulties of models with more complex physics. Crucially, it can be simulated over very long time horizons at reasonable computational cost. This last property allows us to study the behavior of very long trajectories, which is infeasible with the time-horizon of current climate datasets.

The remainder of this article is organized as follows. In §2 we outline the mathematical formulation of problem under investigation and in §3 we introduce the specific prototypical climate model to be analyzed. §4 summarizes the specific machine learning architectures we employ, and our results are presented in §5. We conclude with a discussion of the implications of our results in §6.

2 Mathematical Framework

We consider a discretized representation of an ergodic chaotic dynamical system

$$\partial_t \mathbf{q} = F(\mathbf{q}), \quad \mathbf{q} \in \mathbb{R}^N, \tag{1}$$

with initial conditions $\mathbf{q_0}$ following a pre-defined distribution μ_0, which in turn induces a distribution of trajectories. Here we loosely define a chaotic system as one whose trajectories are highly *sensitive to perturbations of initial conditions.* Specifically, chaotic systems are characterized by having a positive Lyapunov exponent: small discrepancies in the initial conditions are exaggerated exponentially over time (Strogatz, 2018). In defining the system (1) we assume N is large enough that the statistics of the solution $\mathbf{q}$ do not change with increasing N – we refer to such a system as being *"fully-resolved"*. Correspondingly, we also consider an *"under-resolved"* discretization of the same dynamical system, described by

$$\partial_t \mathbf{v} = f(\mathbf{v}), \quad \mathbf{v} \in \mathbb{R}^n, \tag{2}$$

where $n < N$, and, crucially, the statistics of $\mathbf{v}$ depend on n. Finally, we define the projection of the fully-resolved solution onto the coarse grid via the projection operator $\mathbf{P}$

$$\mathbf{u} = \mathbf{P}\mathbf{q}, \quad \mathbf{u} \in \mathbb{R}^n. \tag{3}$$

Moving forward, $\mathbf{u}$ will be referred to as the reference data (RD) and $\mathbf{v}$ will be referred to as the coarse data (CR). We also consider the discretization in time of the solutions of (1) and (2) to snapshots sampled equi-spaced in time, resulting in the sequences $\{\mathbf{v}_j\}_{j=1}^T$ and $\{\mathbf{u}_j\}_{j=1}^T$, where $\mathbf{u}_j = \mathbf{P}\mathbf{q}_j$.

The objective of this work is to learn a *parametric correction operator*

$$\mathcal{G}_\theta : \mathbb{R}^{n\times T} \to \mathcal{P}(\mathbb{R}^{n\times T}), \tag{4}$$

where T is the length of the trajectories, $\mathcal{P}(\mathbb{R}^{n\times T})$ is the push-forward map by $\mathbf{P}$ of a distribution of trajectories of system (1), and θ are the parameters of the map. Thus, $\mathcal{G}_\theta$ maps trajectories from the distribution of the under-resolved (coarse) system (2) to distributions of trajectories of the projected fully-resolved (reference) system (1). We are focused on the statistical evaluation of long term climate risks, and thus the aim of (4) is not to approximate any specific reference trajectory on a snapshot-by-snapshot basis, but rather to generate plausible trajectories which reflect the statistics of the reference data.

We highlight that the operator $\mathcal{G}_\theta$ maps trajectories from n-dimensional state space to n-dimensional state space, and is not intended to recover the fine scales unresolved by the coarse model. Therefore, all results presented in this work should be understood as being defined on the coarse grid.

2.1 Training on Nudged Simulations

The primary obstacle to learning a map $\mathcal{G}_\theta$ is that the systems associated to $\mathbf{v}$ and $\mathbf{u}$ are chaotic, and therefore there is no natural pairing between trajectories (Wan, Baptista, Boral, et al., 2023). One could learn a map between any arbitrary pair of trajectories, but such map will be highly specific to that particular ordering, and in general will not generalize to unseen data. In addition, for the sake of generalization the mapping must directly encode the spatiotemporal dynamics of the system (1), not just the statistics of the specific trajectories used in training. This additional constraint stems from the downstream application: practical long-term (multi-century) climate forecasting will require training correction operators on the few decades of available high quality data whose statistics are not converged – especially for rare events whose characteristic return period is on the order of centuries. If $\mathcal{G}_\theta$ is trained to simply generate trajectories drawn from the distribution defined by the training data such extrapolation is

often impossible without additional strong inductive biases, which themselves are usually not well defined.

To overcome these challenges, we employ the framework introduced by Barthel Sorensen et al. (2024) in which the correction operator is trained on trajectory pairs consisting of a fully-resolved reference trajectory and an under-resolved trajectory *nudged* towards that reference trajectory. We briefly summarize the mathematical rationale of the approach below, and refer the interested reader to Barthel Sorensen et al. (2024) for a more detailed presentation.

Consider the deviation between the under- and fully- resolved representations of the dynamical system

$$\boldsymbol{\delta} \equiv \mathbf{v} - \mathbf{u}, \quad \boldsymbol{\delta} \in \mathbb{R}^n, \tag{5}$$

which is governed by the system

$$\partial_t \boldsymbol{\delta} = f(\boldsymbol{\delta} + \mathbf{P}\mathbf{q}) - \mathbf{P}F(\mathbf{q}). \tag{6}$$

Due to the chaotic nature of the system, $\boldsymbol{\delta}$ will grow exponentially. This is known as chaotic divergence and makes a map between any two arbitrary realizations of $\mathbf{v}$ and $\mathbf{u}$ meaningless. This divergence can be constrained through the introduction of a small damping term on the right hand side of (5) resulting in

$$\partial_t \boldsymbol{\delta}_\tau = f(\boldsymbol{\delta}_\tau + \mathbf{P}\mathbf{q}) - \mathbf{P}F(\mathbf{q}) - \frac{1}{\tau}\boldsymbol{\delta}_\tau, \tag{7}$$

which when expressed in terms of the original variables takes the form

$$\partial_t \mathbf{v}_\tau = f(\mathbf{v}_\tau) - \frac{1}{\tau}(\mathbf{v}_\tau - \mathbf{u}), \quad \mathbf{v}_\tau \in \mathbb{R}^n. \tag{8}$$

If the reference solution $\mathbf{u}$ is known, the system (8) is said to be *nudged* towards $\mathbf{u}$ – an approach which originates in the field of data assimilation, where it has been used to improve the predictive capabilities of weather models (Huang et al., 2021; Miguez-Macho et al., 2005; Storch et al., 2000; Sun et al., 2019) . The forcing term on the right hand side of (8) is known as the nudging tendency, and the user-defined constant τ represents a time scale over which this forcing acts. The nudging tendency will have a negligible effect when $(\mathbf{v}_\tau - \mathbf{u})$ is small and an $O(1)$ effect on the dynamics only when the deviation $(\mathbf{v}_\tau - \mathbf{u})$ grows to be $O(\tau)$. Through a multiscale analysis, Barthel Sorensen et al. (2024) showed that nudging is equivalent to forcing the dynamics evolving on time scales slower than τ to follow the slow dynamics of the reference trajectory $\mathbf{u}$, while the faster dynamics are free to evolve according to the unforced coarse dynamics (2).

Training on the pair of trajectories $\mathbf{v}_\tau$ and $\mathbf{u}$ allows the correction operator $\mathcal{G}_\theta$ to learn the fast dynamics of the fully-resolved system which are most affected by the lack of resolution, while being minimally corrupted by the chaotic divergence of the large-scale slow dynamics. The aim therein is to learn a map which reliably maps trajectories in the distribution induced by the coarse dynamics (2) to the distribution induced by the reference (fully-resolved) dynamics (1). However, the inclusion of the nudging tendency in (8) introduces artificial dissipation, which causes the spectrum of the nudged solution $\mathbf{v}_\tau$ to differ from that of the free running solution $\mathbf{v}$. To address this, we define the spectrally corrected nudged solution

$$\mathbf{v}'_\tau = \mathcal{F}^{-1}[a_k \hat{\mathbf{v}}_{\tau,k}], \tag{9}$$

where $\hat{\mathbf{v}} \equiv \mathcal{F}[\mathbf{v}]$ is the spatial Fourier transform and a_k is the spectral ratio defined as

$$a_k \equiv \sqrt{\int_0^T |\hat{\mathbf{v}}_k|^2 dt \left(\int_0^T |\hat{\mathbf{v}}_{\tau,k}|^2 dt \right)^{-1}}. \tag{10}$$

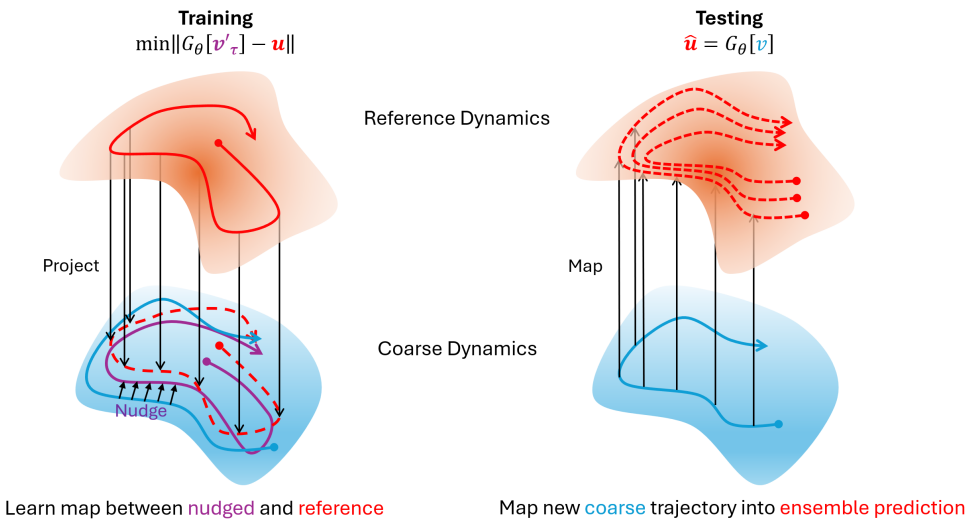


Figure 1: Diagram of the nudging-based machine learning framework.

We note that several other strategies to address such spectral inconsistencies have been proposed such as 4DVar (Dimet & Talagrand, 1986; Mons et al., 2016; Wang et al., 2019) or ensemble variational methods (Liu et al., 2008; Mons et al., 2016; Buchta & Zaki, 2021). We utilize the simple spectral correction due to its ease of implementation and the fact that it does not require iterative simulation of the governing equations as some of these other methods. In practice the training data consists of 3 trajectories, the reference data $\mathbf{u}$, the spectrally-corrected nudged coarse data $\mathbf{v}'_\tau$, and a free running coarse trajectory $\mathbf{v}$ used for the spectral correction (9). We then formulate the general supervised learning problem

$$\min_\theta \int_0^T \|\mathcal{G}_\theta[\mathbf{v}'_\tau] - \mathbf{u}\|^2 \, dt, \tag{11}$$

where $\mathbf{v}'_\tau$ and $\mathbf{u}$ are understood to be discrete trajectories. By formulating the learning in terms of trajectories – and not just statistics – the learned map directly encodes the temporal dynamics of the system. This allows for the possibility of the learned map to extrapolate to trajectories which are much longer than the training data which would be impossible if $\mathcal{G}_\theta$ was trained only to reproduce the statistics of the data seen in training (Blanchard et al., 2022). A diagram of the general learning framework is shown in Figure 1.

3 Quasi-Geostrophic Model

Similarly to (Barthel Sorensen et al., 2024), we consider a two-layer quasi geostrophic model as prototypical climate model. The model is defined on 2D Cartesian grid, $(x, y) \in [0, 2\pi]^2$, and takes the form

$$\frac{\partial q_j}{\partial t} + \left(U_j + \hat{\mathbf{k}} \times \nabla\psi_j\right) \cdot \nabla q_j + \left(\beta + k_d^2 U_j\right) \frac{\partial \psi_j}{\partial x} = -\delta_{2,j} r \nabla^2 \psi_j - \nu \nabla^8 q_j, \tag{12}$$

where $j = 1, 2$ corresponds to the upper and lower layers. The dependent variable appears in two forms: $q_j(x, y, t)$ and $\psi_j(x, y, t)$, which are the potential vorticity and stream function respectively. Without loss of generality, all results in this work will be presented in the form of the stream functions ψ_j.

The system is parameterized by the bottom-drag coefficient r, the beta-plane approximation parameter β, and the deformation frequency k_d^2. In this work we fix $[r, \beta, k_d^2] = [0.1, 2.0, 4.0]$ – values consistent with mid-latitude atmospheric flow. The imposed zonal mean flow is given by $U_j = -1^{(j+1)}U$, with $U = 0.2$.

To quantify the effectiveness of our methodology to anisotropic problems we introduce topography on the bottom surface. The topography profile $h_b(x, y)$ is introduced through the definition of the potential vorticity

$$q_j = \nabla^2 \psi_j + \frac{k_d^2}{2}\left(\psi_{3-j} - \psi_j\right) + \frac{f_0}{h_2} h_b(x, y)\delta_{j,2}. \tag{13}$$

Here f_0 is the inertial frequency which we set to 1, h_2 is the thickness of the lower layer, and $\delta_{j,2}$, indicates that the topography term is only included in the definition of the lower layer potential vorticity q_2. We consider a topography profile consisting of seven randomly spaced Gaussians with equal variance

$$h_b(x, y) = A \sum_{j=1}^{7} e^{-\frac{(x-a_j)^2 + (y-b_j)^2}{\sigma^2}}, \tag{14}$$

where the coordinates $[a_j, b_j]$ and variance σ^2 represent the centers and width of the Gaussian "mountains". The specific values were chosen to ensure that the profile would not violate the periodic boundary conditions. An illustration of the topography profile is shown in Figure 2c.

Equations (12) and (13) are solved using a spectral method in space and then integrated using a 4^{th} order Runge-Kutta scheme in time. We consider 128×128 and 24× 24 grid to represent the specific fully- and under- resolved systems (1) and (2), respectively. For each case, we run a *single* simulation for 35,000 time units, the first 1,000 time units are used for training, and the remaining 34,000 are used for testing. One additional nudged simulation over 1000 time units is performed to generate the training data (9) needed to construct the supervised learning problem (11).

Figure 2a shows the zonally averaged flow field as an illustrative example. Note the difference in amplitude between the RD and CR solutions. Figure 2b shows the spatial variation of the normalized variance of the stream function data defined as

$$\tilde{\sigma}(x, y) = \frac{\sigma(x, y) - \overline{\sigma}}{\overline{\sigma}}, \tag{15}$$

where the variance is computed over the temporal dimension (34,000 time units) and $\overline{\sigma}$ denotes a spatial average. This highlights both the anisotropy present in the flow as well as the non-trivial differences in the spatiotemporal features of the RD and CR data sets. Finally, we reemphasize that the RD dataset represents the high resolution solution projected onto the coarse grid, and thus all data and results shown in this work are defined on the coarse 24×24 grid.

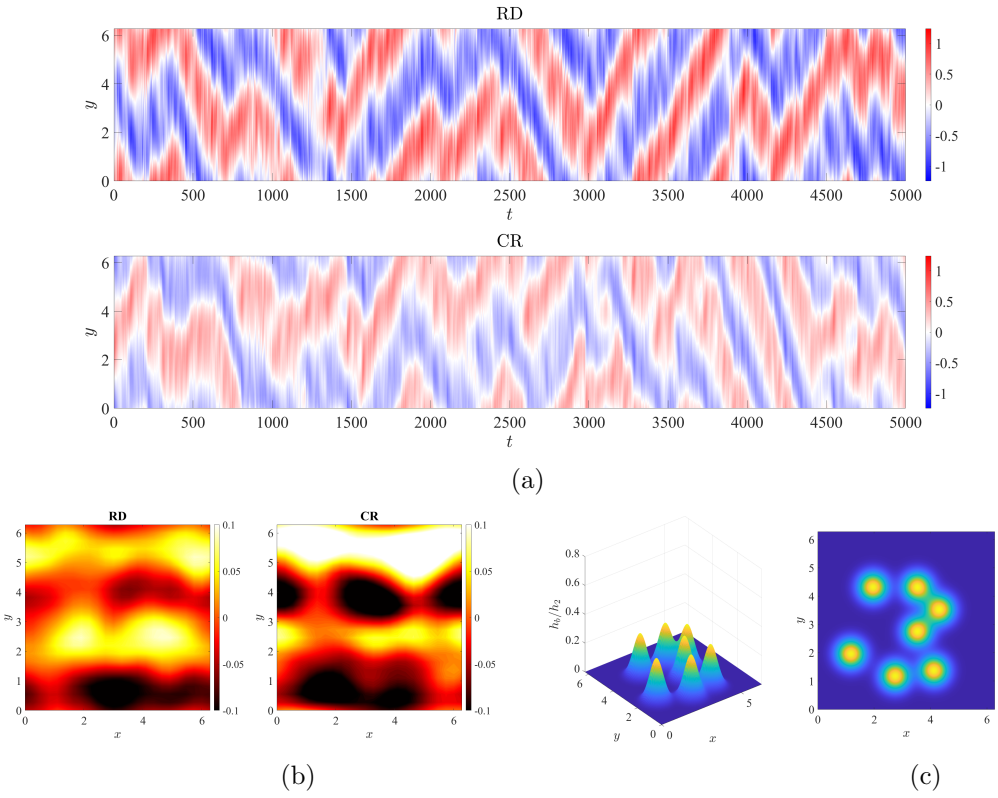


Figure 2: Zonal average (a) and normalized covariance (b) of the lower layer stream function ψ_2 of the RD and CR data sets. Illustration of the bottom topography profile (c).

4 Machine Learning Architecture

Here we provide a brief description of the neural network architectures and uncertainty quantification strategies investigated in this work. We reemphasise that the aim of the current approach is to train correction operators which are effective when applied to unseen trajectories which are significantly (perhaps orders of magnitude) longer than the training trajectories. To this end, we investigate three probabilistic extensions of the previously validated Long Short Term Memory (LSTM)-based network (Barthel Sorensen et al., 2024), all based on the principle of VAEs (Kingma & Welling, 2022). To illustrate the number of possible, and often subtle, interactions between the VAE and LSTM we begin with a brief outline of a basic RNN and then explain how each of the three architectures under investigation builds upon this baseline. At a high level, the VAE introduces a probabilistic latent space which in theory allows the network to learn embeddings of the limited training data in a manner which is cognizant of and robust to the limitations of that data. The primary variation we investigate here is whether this latent space is implemented *"upstream"* or *"downstream"* of the LSTM unit in the computational graph of the network as a whole. Much of the discussion in §5 and §6 focuses on the advantages and disadvantages of each and how these may be exploited or mitigated respectively.

4.1 Recurrent Neural Networks

One of the most widely used class of ML architectures for modeling temporal sequences such as the climate systems which motivate our research is the RNN (Graves et al., 2007; Sutskever et al., 2008; Graves et al., 2013; Graves, 2014; Sutskever et al., 2014; Cho, van Merrienboer, Gulcehre, et al., 2014). An RNN layer transforms the input sequence $\mathbf{x} = [\mathbf{x}_1, ..., \mathbf{x}_T], \mathbf{x}_t \in \mathbb{R}^n$ into an output $\mathbf{y} = [\mathbf{y}_1, ..., \mathbf{y}_T], \mathbf{y}_t \in \mathbb{R}^m$, via a hidden state $\mathbf{h} = [\mathbf{h}_1, ..., \mathbf{h}_T], \mathbf{h}_t \in \mathbb{R}^d$ according to the following recursive push-forward equations

$$\mathbf{h}_t = f_h\left(\mathbf{U}\mathbf{x}_t + \mathbf{W}\mathbf{h}_{t-1} + \mathbf{b}\right), \tag{16}$$

$$\mathbf{y}_t = f_o\left(\mathbf{V}\mathbf{h}_t + \mathbf{c}\right), \tag{17}$$

where $\mathbf{U}, \mathbf{W}, \mathbf{V}, \mathbf{b}, \mathbf{c}$ represent the trainable parameters, and both f_h and f_o are the generally nonlinear activation functions. A graphical representation of the basic RNN unit is given in Figure 3a. This basic formulation is generally augmented using gating mechanisms which alleviate the problem of vanishing gradients (Pascanu et al., 2013) during training which arise due to exponentially small weights assigned to long term dependencies. Specifically, all of the network architectures explored in this work are built on LSTM unit (Hochreiter & Schmidhuber, 1997), as LSTM based architectures have generally demonstrated superior ability to capture long time dependencies as compared to other designs such as the Gated Recurrent Unit (GRU) (Cho, van Merrienboer, Bahdanau, & Bengio, 2014).

4.2 Variational Auto Encoder

The ML correction operator will generally encounter many events which were rarely or not at all seen in training. One architecture that has been proposed to enable such generalization (for non-sequential data) is the Variational Auto-Encoder (VAE) (Kingma & Welling, 2022). A standard Auto-Encoder (AE) is a type of data compression architecture which projects the input data $\mathbf{x}$ onto a reduced order latent space $\mathbf{z}$ and then expands it back to an approximation of the original input data $\tilde{\mathbf{x}}$. The AE is then generally trained to minimize the reconstruction error: $\|\mathbf{x}-\tilde{\mathbf{x}}\|$. The VAE replaces the deterministic latent space in the standard AE with a probabilistic latent space, where for each forward pass the latent space representation is sampled from a distribution, which for ease of parameterization, is generally assumed to be Gaussian $\mathbf{z} \sim \mathcal{N}\left(\boldsymbol{\mu}_z, \boldsymbol{\sigma}_z\right)$. From

an implementation point of view, this implies that each embedding is now not just a single number but a mean and a variance. This extension to a latent space of distributions regularizes or smooths out the latent space ensuring that that structures which are similar in physical space will have similar embeddings – a property which is not guaranteed in a deterministic encoder-decoder network. This built in uncertainty improves the extrapolation capabilities of the network by increasing the likelihood that *never-before-seen* structures will be encoded into latent space representations which are similar to the embedding of similar structures which *were seen* in training, and thereby increasing the likelihood of an accurate decoding.

However, for this framework to be useful some regularization constraints are required on the latent space distribution. For example, without constraints, the network is liable to over-fit to the training data and converge to a latent space whose mean values are distant from one another and whose covariances vanish thereby negating the benefit of the probabilistic framework entirely. This regularization is achieved through an addition to the loss function which penalizes deviations of the distribution $p(z) \sim \mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\sigma}_z)$ from a standard Normal distribution: $\mathcal{N}(\mathbf{0}, \mathbf{I})$. We note that while other priors are possible, these were not pursued in this work.

4.3 Probabilistic Recurrent Neural Networks

The probabilistic treatment of sequential temporal data requires the combination of the RNN and VAE frameworks. Such hybrid architectures are also known as Deep State Space Models (DSSMs) (Gedon et al., 2021), however to minimize unnecessary jargon we will refer to such models simply as probabilistic (as opposed to deterministic) RNNs.

Here we investigate three recently proposed probabilistic RNN architectures: the VAE-RNN (Fraccaro et al., 2016; Fraccaro, 2018), the stochastic RNN (STORN) (Bayer & Osendorfer, 2015), and the variational RNN (VRNN) (Chung et al., 2016).

4.3.1 VAE-RNN

The VAE-RNN is the simplest form of probabilistic RNN. In this case a VAE is simply appended to the output of the RNN at each time step independently – this is illustrated graphically in figure 3b. The recursive push-forward equations for the VAE-RNN are

$$\mathbf{h}_t = f_h\left(\mathbf{U}\mathbf{x}_t + \mathbf{W}\mathbf{h}_{t-1} + \mathbf{b}\right), \tag{18}$$

$$\mathbf{z}_t \sim \mathcal{N}\left(\boldsymbol{\mu}_z(\mathbf{h}_t), \boldsymbol{\sigma}_z(\mathbf{h}_t)\right), \tag{19}$$

$$\mathbf{y}_t = f_o\left(\mathbf{V}\mathbf{z}_t + \mathbf{c}\right), \tag{20}$$

where $\boldsymbol{\mu}_z^2(\mathbf{h}_t) = \mathbf{B}\mathbf{h}_t$ and $\boldsymbol{\sigma}_z(\mathbf{h}_t) = \text{softplus}\left(\mathbf{C}\mathbf{h}_t\right)$ are both themselves parameterized through the trainable weight matrices $\mathbf{A}$ and $\mathbf{B}$ and the use of the softplus activation function ensures a positive variance. A critical (and limiting) feature of the VAE-RNN architecture is that the latent space dependency is *downstream* of the recurrence relationship and thus there is no communication between time steps $\mathbf{z}_t$. The following two architectures remedy this limitation.

4.3.2 STORN

The STORN architecture does not append a VAE to the output of the RNN but instead introduces the latent space upstream of the recurrence relationship, namely as an additional input to the RNN. Specifically, it consists of the following push forward

equations

$$\mathbf{z}_t \sim \mathcal{N}\left(\boldsymbol{\mu}_z(\mathbf{x}_t), \boldsymbol{\sigma}_z(\mathbf{x}_t)\right), \tag{21}$$
$$\mathbf{h}_t = f_h\left(\mathbf{U}\mathbf{x}_t + \mathbf{W}\mathbf{h}_{t-1} + \mathbf{A}\mathbf{z}_t + \mathbf{b}\right), \tag{22}$$
$$\mathbf{y}_t = f_o\left(\mathbf{V}\mathbf{h}_t + \mathbf{c}\right), \tag{23}$$

where the latent space is parameterized in terms of the input variable $\boldsymbol{\mu}_z^2(\mathbf{x}_t) = \mathbf{B}\mathbf{x}_t$ and $\boldsymbol{\sigma}_z(\mathbf{x}_t) = \text{softplus}\left(\mathbf{C}\mathbf{x}_t\right)$. A graphical illustration of the basic STORN architecture is given in figure 3c.

4.3.3 VRNN

The VRNN architecture includes both an upstream and downstream latent space dependency, and can be interpreted as a combination of the VAE-RNN and STORN architectures. The latent space is introduced as an input to the RNN but is also appended to its output. The generative equations are

$$\mathbf{z}_t \sim \mathcal{N}\left(\boldsymbol{\mu}_z(\mathbf{x}_t), \boldsymbol{\sigma}_z(\mathbf{x}_t)\right) \tag{24}$$
$$\mathbf{h}_t = f_h\left(\mathbf{U}\mathbf{x}_t + \mathbf{W}\mathbf{h}_{t-1} + \mathbf{A}\mathbf{z}_t + \mathbf{b}\right) \tag{25}$$
$$\mathbf{y}_t = f_o\left(\mathbf{V}_1\mathbf{h}_t + \mathbf{V}_2\mathbf{z}_t + \mathbf{c}\right) \tag{26}$$

where the latent space is parameterized as in the STORN model. Chung et al. (2016) investigate both a standard Gaussian prior (VRNN-I) as well as a generally time dependent prior which is learned during the training phase. Here we consider only the VRNN-I variant, which for simplicity we refer to as VRNN. In practice, these two architectures generally demonstrate similar levels of performance (Chung et al., 2016; Gedon et al., 2021).

4.4 Ensemble Analysis

The probabilistic architectures described above help to address the uncertainty due to limited training data. However, there is also uncertainty due to the random nature of the optimization algorithm used to train the network and the highly non-convex nature of the optimization landscape. To leverage this uncertainty we employ an ensemble approach in which we train the same architecture multiple times on the same training data. This results in an ensemble of NN's: $\mathcal{G}_{\theta_j}$ and therefore an ensemble of predictions $\hat{\mathbf{u}}_j = \mathcal{G}_{\theta_j}[\mathbf{v}], \quad j = 1...N_e$ where N_e is the number of ensemble members. We then define the prediction of any statistic or observable $g(\mathbf{u})$ as the average prediction of the ensemble members

$$\bar{g} = \frac{1}{N_e}\sum_{j=1}^{N_e} g(\mathcal{G}_{\theta_j}[\mathbf{v}]). \tag{27}$$

The uncertainty is then quantified through the ensemble variance

$$\sigma_g^2 = \frac{1}{N_e - 1}\sum_{j=1}^{N_e}\left(g(\mathcal{G}_{\theta_j}[\mathbf{v}]) - \bar{g}\right)^2. \tag{28}$$

We note that due to their probabilistic nature, each forward evaluation (on the same input) of the VAE-RNN, STORN, and VRNN architectures leads to slightly different outputs. However, we have found that the variance in the long time statistics of these variable predictions is negligible. In fact, the variance quantified by (28) is dominated by the ensemble variance, and is not meaningfully affected by the probabilistic nature of the architectures. This is both expected and desirable, as even if each forward pass of the model produces a different realization, we expect each of these to be drawn from the same distribution and thus to share common long time statistics.

Model	Trainable Parameters	$\overline{D_{KL}}$	$\overline{L_1}$
RNN	168,492	0.0239	3.68
VAE-RNN	175,812	0.0026	4.35
STORN	190,212	0.0089	2.46
VRNN	259,332	0.0044	2.47

Table 1: Number of trainable parameters (degrees of freedom) and global prediction errors for each network architecture considered in this work.

In A1 we present a detailed parametric study on the effects of ensemble size and training duration for each of the four architectures described above. In general, for all architectures the effect of considering an ensemble as opposed to a single network is small but meaningful. For clarity of exposition, we focus the remainder of our discussion on results computed from an ensemble of 6 neural networks each of which is trained for 500 epochs. We found that in general increasing the ensemble size further increased the computational cost substantially while leading to only marginal improvements. All following results – for all architectures – are the ensemble mean prediction as quantified by (27).

4.5 Network Architecture and Training Details

The correction operator used in this work are based on the LSTM-based architecture already validated by Barthel Sorensen et al. (2024) on the isotropic version of the QG model i.e. without topography. This architecture consists of a single layer encoder which compresses the input to a hidden state of dimension 60, followed by an LSTM layer of the same size, and a single layer decoder that restores the output to the original size. For the probabilistic models the latent space dimension was also set to 60.

As our main aim in this paper is to exhibit the advantages of the probabilistic methods, we have left the encoder, decoder, and LSTM layers of the networks as unaltered as possible. With the exception of the VRNN architecture, the inclusion of the latent space does not meaningfully impact the number of trainable parameters which are summarized in table 1. The increase in degrees of freedom for the VRNN architecture is due to the increased size of the input to the decoder layer (26). However, we found that neither increasing the depth or width of the encoder and decoder layers, nor varying the dimension of the latent space had any significant impact on the results. Therefore, we expect that any differences in performance are not simply due to an increase in the degrees of freedom.

The loss function used to train the correction operators consists of three terms: a mean squared prediction error, a term that penalizes deviations in the conservation of a mass in the QG model, and the KL divergence term regularizing the latent space distribution – the latter being only present for the probabilistic architectures. The overall expression for the loss is given by

$$L(\theta) = \int_0^T \left\| \mathcal{G}_\theta[\mathbf{v}'_\tau] - \mathbf{u} \right\|^2 dt + \int_0^T \left\| \mathcal{G}_\theta[\mathbf{v}'_\tau] \right\| dt + \lambda D_{KL} \left(\mathcal{N}\left(\boldsymbol{\mu}_z(\theta), \boldsymbol{\sigma}_z(\theta)\right), \mathcal{N}\left(\mathbf{0}, \mathbf{I}\right) \right) \quad (29)$$

The normalization constant λ sets the strength of the regularization on the probabilistic latent space: if it is too large, the model will ignore the prediction error and drive the latent space to pure white noise, and if it is too small, the model will over fit to the data and the latent space will have no effect. Empirically, we found that for our problem a value of 10^{-4} led to the best results.

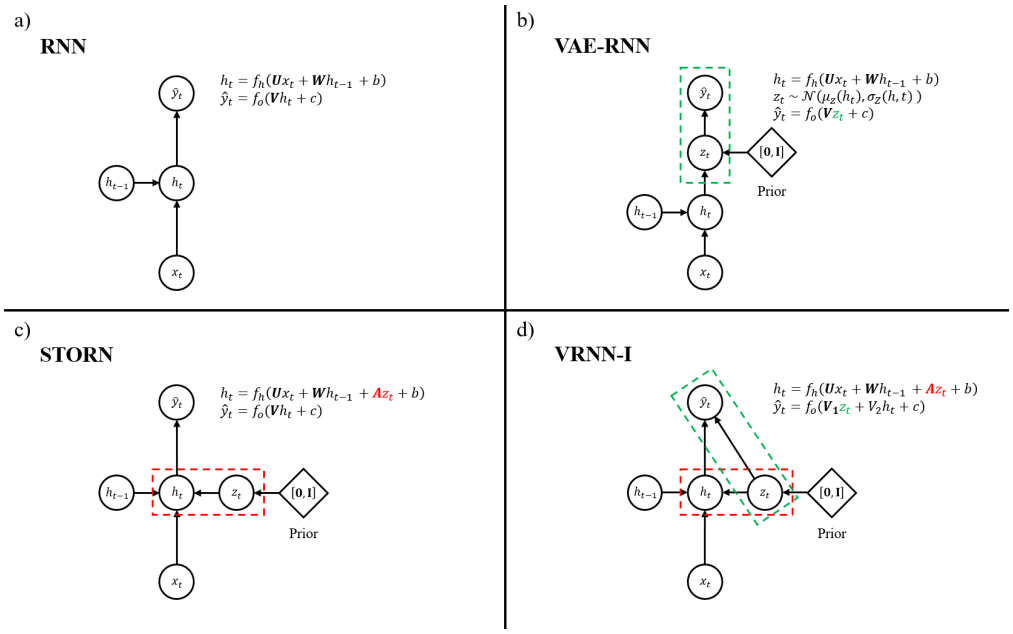


Figure 3: Graphical model and recursive evaluation equations for the four network architectures considered in this work: basic RNN (a), VAE-RNN (b), STORN (c), VRNN (d). Latent space dependencies upstream and downstream of the recurrent layer are marked red and green respectively.

5 Results

Here we showcase the results of our machine learning framework, as introduced in §2 using the network architectures described in §4, applied to the quasi-geostrophic system described in §3. All the results herein represent the ensemble mean prediction (27) of six ML correction operators applied to a single unseen realization of the flow of length 34,000 time units – 34 times the length of the training data. The focus of the discussion is the comparison of the architectures described in §4; ensemble size sensitivity is explored in A1.

We present our results in the form of probability density functions (pdfs) as well as one and two point correlations. We are interested in the ability of the correction operator to accurately quantify the probability of extreme events – particularly of those whose return period is longer than the training data. Therefore, all pdf results will be presented on both a linear and logarithmic scale. The former illustrates the bulk of the distribution, while the latter emphasizes the tails. Accordingly, we will make use of the following two error metrics to evaluate the statistical accuracy of the ML predictions. The Kullback-Liebler (KL) divergence, defined as

$$D_{KL}(p||q) \equiv \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx \tag{30}$$

and the L1 error of the log-pdf

$$L_1(p||q) \equiv \int |\log(p(x)) - \log(q(x))| dx. \tag{31}$$

This latter metric, which we will refer to as the L_1 error, is chosen specifically to emphasize deviations in the tails. These two metrics can be thought of as measures of overall and extreme event specific accuracy respectively.

5.1 Global Statistics

Results for the global pdf, log-pdf, and power spectral density of the stream function are shown in Figure 4. Here we compare the (ensemble mean) prediction of the ML corrected coarse model (shown in color) to the true statistics (solid black) and those of the uncorrected coarse model (dashed black). All three probabilistic architectures capture the true pdf better than the deterministic architecture – which while significantly improving the uncorrected simulation, still overestimates the probability of very low amplitude events and under estimates the tail statistics. The average (over ψ_1 and ψ_2) global KL-Divergence and L1 log-pdf error for each architecture is listed in table 1. In all cases, the probabilistic architectures outperform the deterministic RNN. The VAE-RNN achieves the lowest overall KL divergence, but has the highest L_1 error, meaning it captures the bulk of the distribution well but does not capture the tails accurately. In regard to capturing tail risk events, the STORN and VRNN architectures generally provide optimal results. They accurately reflect the true distribution across the full range of amplitudes, while the VAE-RNN architecture tends to mildly over-predict the tails.

To highlight the ability of our ML correction operator to extrapolate from the short training data we show in figure 5 the differences in the statistics of the long (34,000 time unit) test data and the short (1,000 time unit) training data. The training data is clearly not converged. In fact, the heavy tails are missing from the training data entirely. As shown in figure 4, the ML corrections accurately capture the tails of the underlying pdf even where there the training data does not. From this ability of the ML correction to extrapolate beyond the training data we infer that the NN is in fact learning some notion of the underlying system dynamics – a key feature in extending the proposed method to more complex system and even longer time horizons.

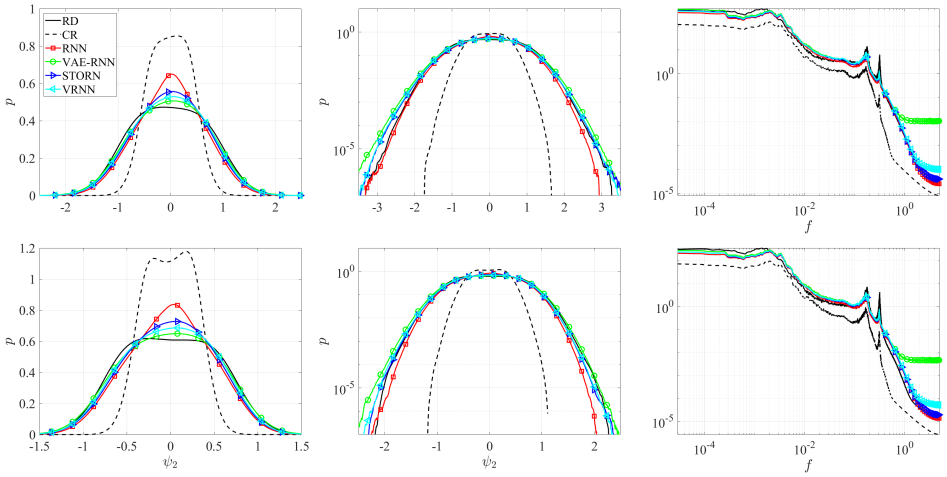


Figure 4: Global pdf, log-pdf, and PSD of ψ_1 (upper panel) and ψ_2 (lower panel). RD (solid black), CR (dashed black), RNN (red), VAE-RNN (green), STORN (blue), VRNN (teal).

In figure 4 we also plot the global power spectral density (PSD), defined as the spatial average of the temporal Fourier transform of the autocorrelation,

$$S_j(f) \equiv \int_0^{2\pi}\int_0^{2\pi}\int R_j(\tau)e^{-if\tau}d\tau dxdy, \tag{32}$$

$$R_j(\tau) \equiv \int \psi_j(t)\psi_j(t+\tau)dt. \tag{33}$$

With the exception of the VAE-RNN architecture, the ML corrections accurately reflect the true power spectrum across the full range of frequencies – including the two characteristic peaks near $f = 0.1$. The VAE-RNN architecture accurately captures the lower frequencies – those with meaningful energy content – but fails to accurately predict the energy roll off of the highest frequencies. This is an intrinsic limitation of the VAE-RNN architecture (18). For frequencies with very low energy the prediction error term in the loss function will become negligibly small, and the training loss will be dominated by the term enforcing the white noise prior placed on the latent space. For those frequencies, the latent space z will then be driven to exactly white noise, and due to the lack of communication across the time steps of z_t inherent in the solely downstream latent space interaction in (18) the output will also be dominated by white noise. This flat spectrum phenomenon is also present to a minor extent in the VRNN architecture (Fig. 4) which also has a downstream latent space dependency. However, the inclusion of the upstream dependency in the VRNN architecture enables the communication between time steps z_t which helps to additionally regularize the latent space. Finally, we again emphasize the extrapolation capabilities of our training framework evidenced by the difference between the PSD of the training data (magenta) and the test data (black).

To further probe the spatiotemporal accuracy of the ML corrected fields we compute the fraction of the domain over which the stream function exceeds a certain threshold as a function of time,

$$A_c(t)/A = \frac{1}{N_x N_y}\sum_{i,j}^{N_x,N_y} H\left(|\psi(x_i,y_j,t)| - c\right). \tag{34}$$

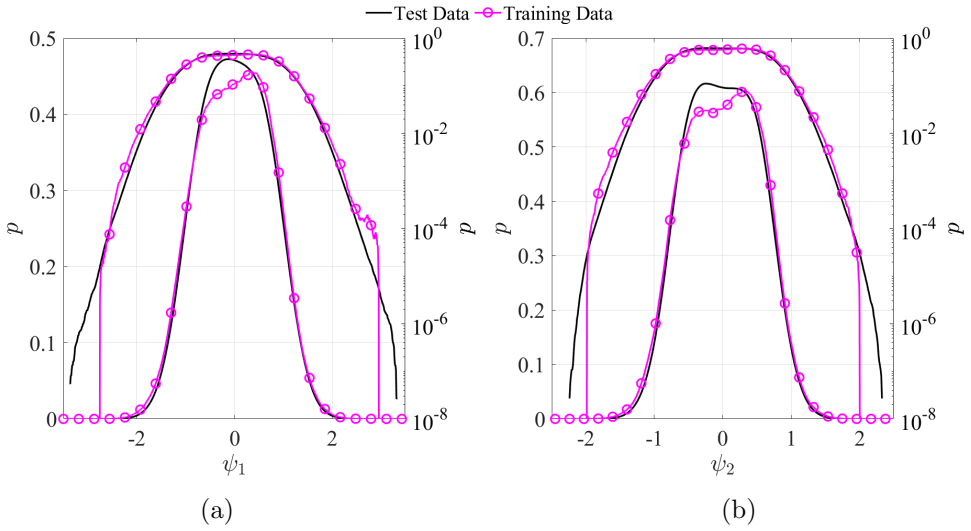


Figure 5: Ground truth reference statistics of 34,000 time unit test data (black) and 1,000 time unit training data (magenta) of ψ_1 (a) and ψ_2 (b). Each subfigure shows the pdf on linear and logarithmic scale.

Here c is the given threshold and $H(x)$ is the unit step function such that $H(x) = 1$ if $x \geq 0$ and $H(x) = 0$ if $x < 0$. This metric characterizes how reliably the ML corrections can capture the frequency and spatial extent of extremes, and is a proxy for the ability of the model to capture large-scale extreme phenomena in climate models, such as heatwaves. The probability density functions of $A_c(t)/A$ for a range of c are plotted in Figure 6. For brevity we focus on ψ_2; results for ψ_1 are included in A2. First, we note that the uncorrected (CR) solution vastly underestimates the amplitude of the true solution – missing the higher-amplitude extremes entirely. In contrast, all ML correction models are able to capture the bulk of the distribution. Compared to the RNN, the probabilistic architectures track the pdf significantly better, with the VAE-RNN demonstrating the best performance. The deterministic RNN on the other hand significantly overestimates the probability of low area ratios for the lower thresholds $c < 1$. This is consistent with the results in Figure 4, where the deterministic RNN significantly overestimates the likelihood of very low amplitudes. The probabilistic architectures also seem to demonstrate marginal improvements for higher values of c. However, in these cases the sample size is small and the pdfs – computed by Monte Carlo sampling – are clearly not fully converged.

5.2 Regional Statistics

Due to the anisotropic nature of the QG flow under consideration we are particularly interested in the regional variation of the quality of the ML correction. Therefore, in addition to the global statistics, we also analyze the statistics as a function of spatial location. For clarity of exposition we will focus here on the results in the lower layer, ψ_2. The corresponding results for the upper layer, ψ_1 – which are qualitatively similar – are summarized in A2.

5.2.1 Single-Point Statistics

We first illustrate our results in terms of single point statistics in the form of the pdf and log pdf. The regional power spectra show very little regional variation so we omit

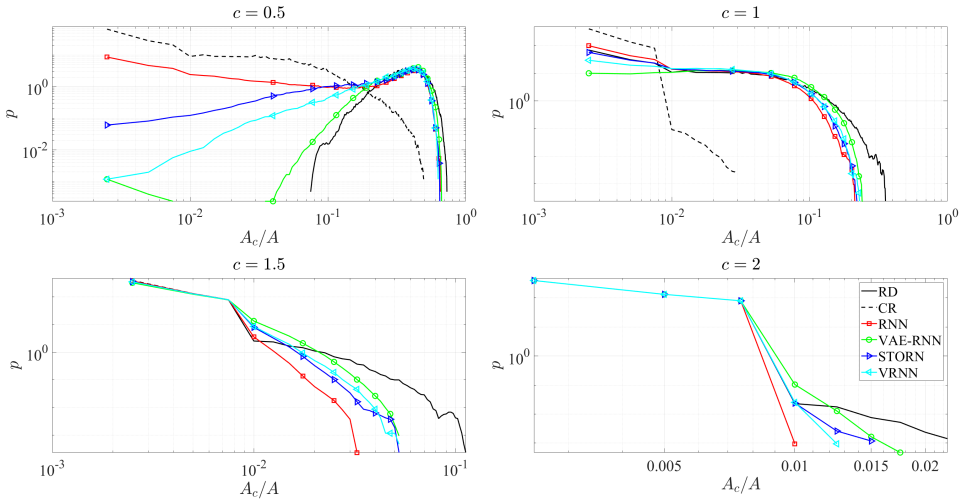


Figure 6: Pdf of fraction of domain over which $|\psi_2|$ exceeds fixed threshold c for range of $c \in [0, 2]$. RD (solid black), RNN (red), VAE-RNN (green), STORN (blue), VRNN (teal).

them here. We divide the domain $[x, y] \in [0, 2\pi]$ into a 3×3 grid and compute the statistics of the stream function in each sub-region. Figure 7a and b show the pdf and log-pdf of the ψ_2 in each sub-region. The difference in pdf shape with respect to location is seen most clearly in the asymmetry of the uncorrected coarse pdfs – some are clearly bimodal, while some peak at small negative values and others peak at small positive values. As was the case with the global statistics, the probabilistic architectures demonstrate a clear improvement over the RNN in the ability to correct the local pdfs. Specifically, the latter incorrectly predicts peaks in the pdf near $\psi_2 = 0$ – a feature which is significantly ameliorated by the probabilistic models, particularly the VAE-RNN and VRNN architectures. In many cases, the overpredictions by the RNN seem to be correlated with the previously mentioned anisotropic peaks in the pdfs of the uncorrected coarse data. This suggests an increased ability of the probabilistic models to handle anisotropic data. This is perhaps due to their ability to more efficiently encode complex (anisotropic) features which had not been seen in training.

A more quantitative view of the regional distribution of the ML correction is given in Figure 8, where the KL divergence and L_1 error are shown as a function of x and y coordinates – here the pdfs and error metrics are computed at each grid point individually. All three probabilistic architectures outperform the deterministic RNN in terms of KL divergence relative to the true pdf. The VAE-RNN architecture has the highest L_1 error, while the RNN, STORN, and VRNN models show similar performance. As a reference we also plot the topography in solid black contours, and we note that the errors in the ML prediction are generally clustered immediately upstream of the topography profile. This is possibly due to an increase in complexity of the flow in this region. It is more likely that the ML correction operator will encounter vortical structures in testing that were not observed in the short training data set which may lead to higher errors.

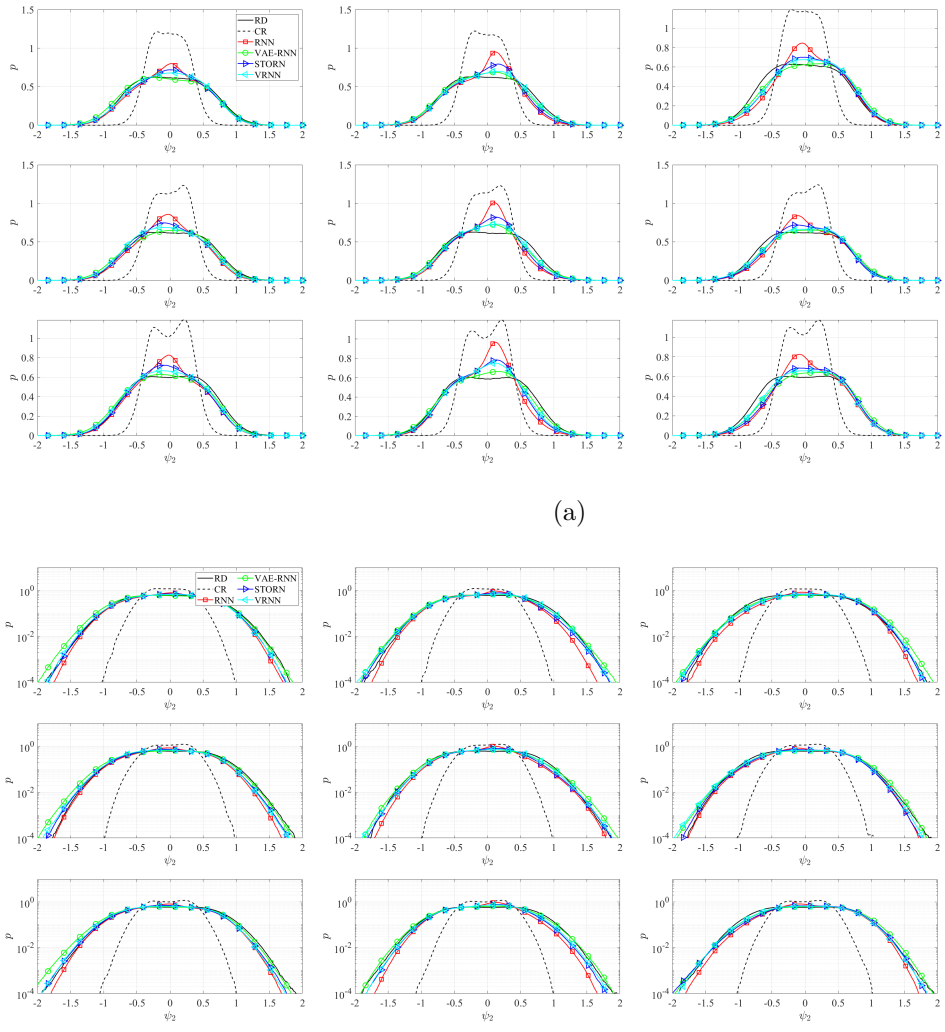


Figure 7: Regional pdf (a) and log-pdf (b) of ψ_2. RD (solid black), CR (dashed black), RNN (red), VAE-RNN (green), STORN (blue), VRNN (teal).

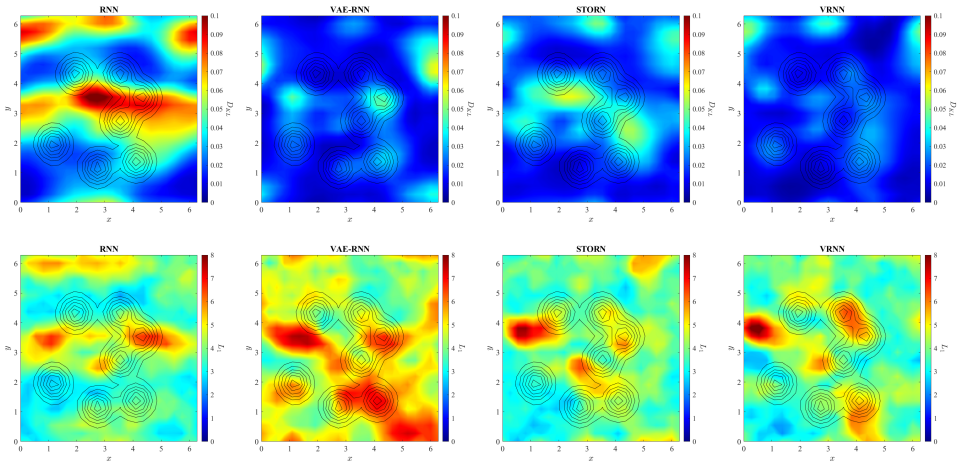


Figure 8: Spatial distribution of KL divergence (upper panel) and L_1 metric (lower panel) for ψ_2. From left to right: RNN, VAE-RNN, STORN, VRNN. The topography profile is show in black.

5.2.2 Fourier Cross-Correlations

To further investigate the spatiotemporal statistics of the corrected fields we compute the normalized cross-correlation between individual Fourier modes

$$\hat{R}_{j,m,n} \equiv \frac{\int \hat{\psi}_j(\mathbf{k}_m, t)\hat{\psi}_j(\mathbf{k}_n, t+\tau)dt}{\sqrt{\int \hat{\psi}_j^2(\mathbf{k}_m, t)dt \int \hat{\psi}_j^2(\mathbf{k}_n, t)dt}}. \tag{35}$$

We focus our discussion on the zonally constant modes, with wave number $\mathbf{k}_m = [0, k_m]$. If $m = n$, this metric is equivalent to a normalized autocorrelation, and for the case $m \neq n$ this metric can be interpreted as a phase shift between Fourier modes. The results for the three largest modes are shown in Figure 9. We find that the uncorrected coarse model correlations are already very similar to those of the high resolution reference. Therefore, the effects of the ML correction on this metric are marginal. In all cases we observe similar decorrelation profiles (top row of Fig. 9) – with the ML correction affording a marginal improvement over the uncorrected baseline. The cross-correlations between Fourier modes (top row of Fig. 9) all fluctuate near 0 for all τ, but again for all architectures we see minimal affect of the ML correction. One potential strategy to address this shortfall in the future is through network architectures which operate directly in Fourier space (Z. Li et al., 2021) – an approach which has been demonstrated to be effective in modeling turbulent flows including global weather patterns (Z. Li et al., 2021; Pathak et al., 2022b) .

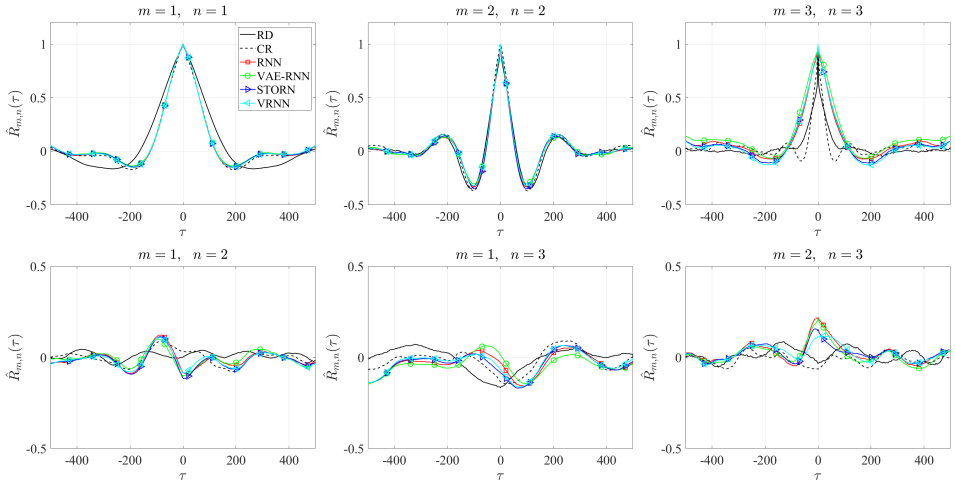


Figure 9: Normalized correlation between three largest zonally constant Fourier modes of ψ_2. RD (solid black), CR (dashed black), RNN (red), VAE-RNN (green), STORN (blue), VRNN (teal).

5.3 Spatiotemporal Features

Finally, we investigate how the spatiotemporal features of the corrected flow fields compare to the reference solution. To this end we define the zonally averaged stream function

$$\overline{\psi}_j(y,t) \equiv \frac{1}{2\pi}\int_0^{2\pi} \psi_j(x,y,t)dx, \tag{36}$$

a quantity that enables us to analyze the meridional advection of structures in the field (HovmÖller, 1949; Qi & Majda, 2020). Figure 10 compares the zonally averaged flow field of the ML corrections to the RD and CR solutions. Since CR and RD are independent trajectories, we expect the corrected flow fields to share the statistics of the reference but not agree on a snapshot-by-snapshot basis. To improve the readability of the figure we limit the time axis to $10,000$ time units. The post-processed flow fields all display characteristic spatiotemporal structures which are consistent with the reference solution, and correct the significant magnitude underestimation of the coarse-resolution field.

In the context of climate, persistent extreme weather events such as long periods of high temperature (heat waves) or low precipitation (droughts) can have outsized effects on the population (Perkins-Kirkpatrick & Lewis, 2020). In order to implement effective mitigation strategies it is crucial to accurately quantify the expected duration of such events, especially as these can occur over a wide range of time scales from days to months (heatwaves) or years (droughts). These concerns are heightened by the expectation that climate change will lead to an increase in both the frequency and severity of such events (Barriopedro et al., 2011; Geirinhas et al., 2021; Meehl & Tebaldi, 2004). For these reasons, it is critical that the ML corrected flow fields accurately reflect the frequency and duration of such extended high amplitude events. While the QG model under investigation here lacks temperature or precipitation, we aim to quantify this ability through the observable

$$\gamma_j(y,t) = \mathrm{MA}_{100}\left(|\bar{\psi}_j(y,t)|^2\right), \tag{37}$$

which we generically refer to as "energy". Here MA_T represents a moving average with a window of length T. We use the filtered energy to eliminate high frequency fluctua-

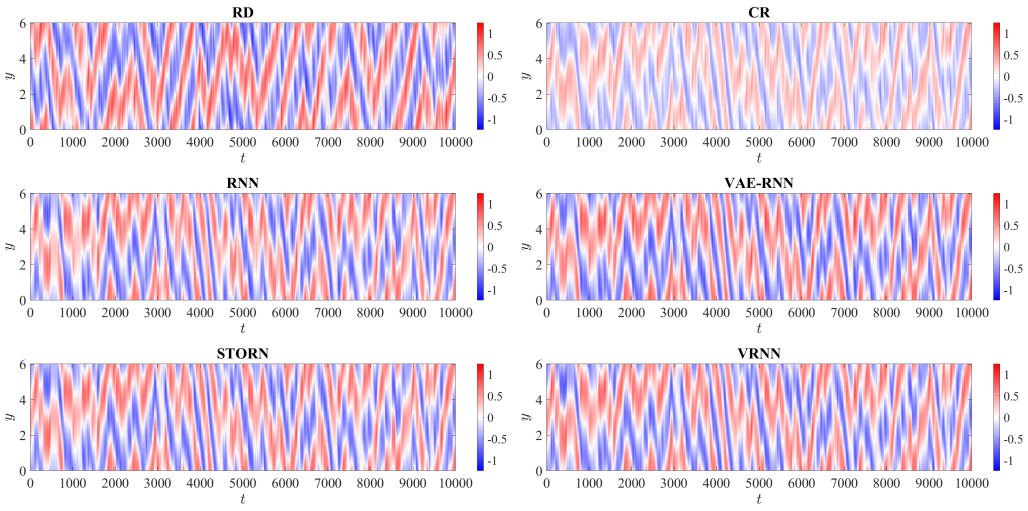


Figure 10: Zonally averaged stream function $\bar{\psi}_2(y,t)$.

tions and focus instead on large deviation events which occur over long time scales. To quantify the statistics of high amplitude excursions of $\gamma_j(y,t)$, we count and measure the duration of periods over which the energy exceeds a given threshold c. We denote the duration of each such period as τ. Figure 11a shows the total number N_c of high amplitude periods as well as their mean duration, $\bar{\tau}$, and standard deviation $\sigma\tau$ as a function of threshold c. We consider values of c ranging from 20% to 90% of the maximum value of γ observed in the reference dataset: γ_{max}. Note that the uncorrected solution (CR) fails to accurately capture any of these statistics. On the other hand, all four ML predictions accurately reflect the dependence of the high amplitude excursion statistics on the threshold c, while slightly under-predicting the total number and average duration. However, in all cases the variance of the high amplitude excursions is well predicted. Note that the ML predictions even capture the non-monotonic behaviour of the total number of excursions N_c for $0.2 < c < 0.4$. This slightly counter-intuitive behaviour indicates that the energy often fluctuates about elevated levels before decaying back down to a lower baseline. We also show in Figures 11b-h the probability density functions of the duration τ for a range of $c \in [0.2, 0.8]\gamma_{max}$ – for higher values of c there are insufficient excursions for meaningfully converged statistics. We omit the pdfs of the uncorrected (CR) solution for $c > 0.2\gamma_{max}$ as these fail to capture the true distributions entirely. Consistent with Figure 11a, we see that in general the pdfs of the ML predictions peak at slightly lower τ for values of $c/\gamma_{max} > 0.3$. However, the probabilistic architectures are in some case able to ameliorate this underprediction – as seen in 11e,f. In these cases, the inclusion of the probabilistic latent space pushes the ML prediction slightly towards higher values of τ – and thus closer to the truth.

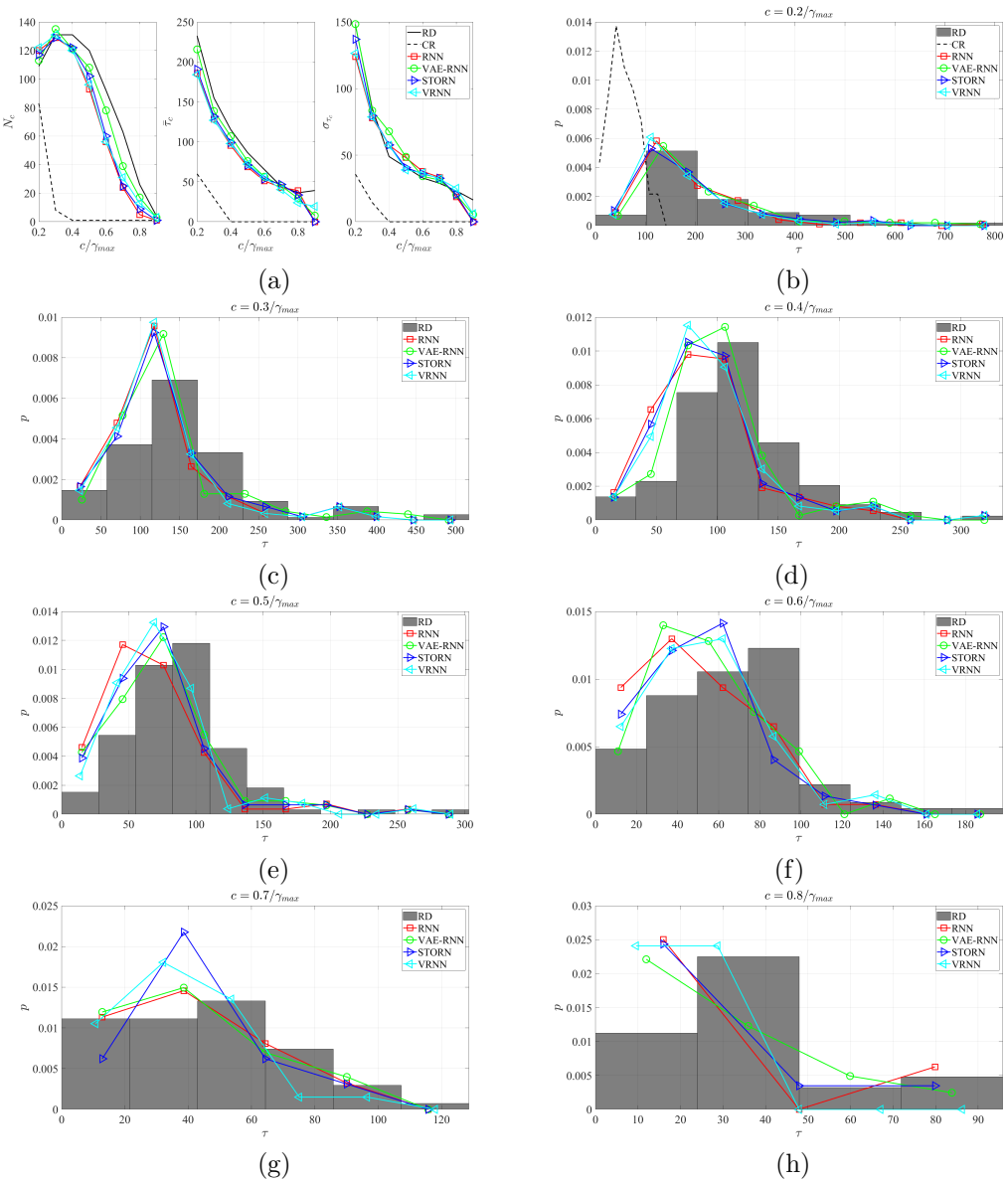


Figure 11: Frequency, expected duration, and variance of high amplitude excursions of γ_2 as a function of threshold c (a). Probability density function of τ for fixed values of c (b-h).

6 Discussion

In this work we developed a non-intrusive data-driven framework for probabilistically debiasing under-resolved long-time climate simulations. This framework, based on training a NN correction operator on nudged simulations of an under-resolved dynamical system, enables learning the intrinsic system dynamics from very short training data sets. The probabilistic extension we propose in this work allows us to significantly improve the extrapolation capabilities of the previous state-of-the-art and enable the quantification of the uncertainty therein. As a test case we considered a two-layer quasi-geostrophic flow in a periodic domain with imposed bottom topography. The topography was included to introduce anisotropy for the purposes of studying the ability of our approach to capture varying regional statistics – a feature not included in the QG example described in previous work. The ML correction operators were trained on trajectories spanning 1,000 time units and tested on 34,000 time units – the statistics of which differ significantly from those of the much shorter training data. We demonstrated the superior performance of our probabilistic framework through it's increased ability to accurately predict both global and regional statistics as well as multiple metrics quantifying the spatial and temporal distribution of rare events. The improvements over the deterministic model described in previous work were especially pronounced when analyzing the spatial variation of extreme (high amplitude) events – a crucial feature in assessing the impact of extreme weather.

One of the key innovations of this work is the variational generalization of the LSTM based network architectures used in previous studies. We investigated three recently proposed architectures (VAE-RNN, STORN, VRNN) (Fraccaro, 2018; Bayer & Osendorfer, 2015; Chung et al., 2016), which primarily differ in the way the probabilistic latent space interacts with the recurrent layer of the network. These dependencies can be categorized as being either *upstream* or *downstream* of the recurrence relation in the computational graph. While we found that all three architectures provide a benefit over the deterministic baseline, the VAE-RNN, which has a strict downstream dependence, achieved the lowest overall error as measured by the KL-divergence. However, the downstream dependency hinders the prediction of outlier events and leads to an overestimation of the high frequency spectral content. These issues are ameliorated through the introduction of an upstream latent space dependency which further regularizes the latent space by allowing for communication between time steps of the latent space encoding. Accordingly, the STORN (upstream only) and VRNN (upstream and downstream) architectures demonstrate the greatest ability to accurately capture the far tails of the true distribution as well as the energy content across the full spectral range. Additionally, we found the STORN and VRNN architectures were significantly more robust to over-fitting, with the VAE-RNN on the other hand showing significant deterioration in predicative capabilities when trained for longer than optimal. However, as the VAE-RNN is simply a VAE appended independently at each time step, these shortcomings should be weighed against its simplicity and ease of implementation. The optimal architecture design will likely depend on the specific application and we hope the analysis discussed in this work can serve as a guide to researchers employing our framework.

While our approach has demonstrated significant skill in correcting the long time statistics of the QG climate model over a range of scales, several limitations remain. Specific to our results: the accurate reconstruction of two point statistics remains a challenge. Our results show that the primary means by which our approach corrects the under-resolved trajectory is by correcting the spatiotemporal dynamics of different Fourier modes independently – while the phase shifts between these remain relatively unchanged. One potential avenue to address this issue is through the use of Fourier Neural operators (Z. Li et al., 2021; Z. Li, Peng, et al., 2022; Pathak et al., 2022b) which operate directly in Fourier space, and may therefore be more effective at correcting the small discrepancies in phase shifts between individual modes. Additionally, the ML corrected fields slightly, but systematically, underestimate the number and duration of high amplitude excursions. Ad-

ditionally, in order to facilitate long time horizon simulations we have focused in this work on a simplified flow, and quantifying the improvements of our probabilistic framework when applied to a full-scale climate model remains the topic of ongoing research.

The proposed framework also has several more fundamental limitations which must be mentioned. First and foremost, an intrinsic limitation of post-processing approaches is their inability to correct processes that are missing from the coarse-resolution model entirely. Improvements in the representation of such processes, such as cloud formation and convective precipitation, requires intrusive corrections to the coarse-resolution model via either improved subgrid-scale closures (Schneider, Teixeira, et al., 2017; Cohen et al., 2020; Lopez-Gomez et al., 2020), or localized high-resolution simulation (Randall et al., 2003; Kooperman et al., 2016). Second, the current framework implicitly assumes that the system is statistically stationary. While similar frameworks have been applied in non-stationary systems (S. Zhang et al., 2024), a correction operator trained under this assumption may fail when applied to trajectories which include strong transitory periods. Finally, the fact that the ML correction operators discussed here are intended to produce long time statistics, but are trained on very short data, implies that there is no obvious metric which can be monitored during training to prevent over-fitting (see A1). However, we have found that the upstream latent space dependencies, as in the STORN and VRNN architectures, serve to regularize the network and drastically increase the robustness of the NN's to over-training. Additionally, ensemble-based predictions help to ameliorate these concerns even further. However, we acknowledge that the efficacy of these strategies may vary from application to application, and in some cases more rigorous regularization strategies may be needed.

In conclusion, we have demonstrated that ensembles of VAE-based RNNs are effective at increasing the extrapolation and rare event quantification capabilities of the non-intrusive debiasing framework introduced by Barthel Sorensen et al. (2024). We investigated several recently developed architectures, which differ primarily in how the probabilistic latent space interacts with the recurrent layer of the neural network. We have classified these interactions as upstream or downstream, and demonstrated that while both are effective, networks with downstream interactions – especially in the absence of additional upstream interactions – are susceptible to over-fitting, and noise corruption. While our work has focused on the application to climate modeling, the general training strategy outlined in this work is applicable to any scenario in which long time statistical analysis requires computationally intractable high resolution numerical simulations.

7 Acknowledgments

This work has been supported through the Google-MIT program "Hybrid Physics and Data-Driven Methods for Statistics of Extreme Weather Events from Climate Simulations".

Open Research Section

The software and data needed to generate the results described in this work can be found at `https://zenodo.org/doi/10.5281/zenodo.13833073` (Barthel, 2024) and `https://github.com/ben-barthel/learning_dynamics`.

Appendix A

A1 Validation and Model Selection

Many machine learning applications operate in what might be referred to as a "data-rich" environment. Even if the total available data is small in an absolute sense, it is common to use a large fraction $75-90\%$ of this available data for training, with only the small remainder used to generate the presented results. The ML models considered here operate in a much more "data-poor" environment, with only 3% of the total data seen in training. One of the challenges in this regard is the lack of obvious metric for online validation. In a "data-rich" environment, a small fraction of the training set may be set aside for validation. Then, as training progresses, the validation error, i.e. the training loss evaluated on the validation set, is monitored and training is stopped when the validation error no longer decreases with each passing epoch. However, if the goal is statistical accuracy over time horizons much longer than the training data, monitoring the training loss (generally the L2 error) over a small fraction of the already limited training set does not provide meaningful insight into the eventual performance when applied to long time series data.

To this end, we conducted a parametric study of the impact of both training time per ensemble member and ensemble size. The results thereof are summarized in figure A2 which shows the global average (over ψ_1 and ψ_2) KL-divergence (30) and L_1 error (31). This parametric study revealed four crucial observations. First, the probabilistic architectures generally lead to lower Kl divergence regardless of training time or ensemble size – note the different color scales between the four subfigures in figure A2a. Second, for a given ensemble size the variational models require less training time to reach a desired level of accuracy. Third, above a certain minimum training time – approximately 500 epochs – it is more advantageous to increase the ensemble size rather than train the models for longer. Fourth and finally, for the probabilistic architectures the error in the prediction of the tails (quantified by the L_1 metric) increases if the model is trained for too many epochs. This is especially pronounced for the VAE-RNN architecture, and is in contrast to the KL-divergence – which quantifies the overall accuracy – which decreases monotonically in almost all cases. The lone exception being the small ensembles of the VAE-RNN architecture.

This deterioration of rare event prediction with increased training is likely due to the probabilistic models over-fitting to the latent space prior. The magnitude of the MSE term in the training loss is proportional to the magnitude of the model output, while the KL divergence term enforcing the latent space prior remains the same order of magnitude regardless of the output. This means that the optimization will tend to ignore errors in the tails of the output distribution in favor of driving the latent space representation of these outlier events ever closer to the pure Gaussian prior. This phenomenon is especially pronounced, in the VAE-RNN with its purely downstream latent space dependency (18). In that case – assuming a linear activation – we have $y_t \sim z_t$ and thus the model output will become increasingly corrupted by white noise. This mechanism is also present to a smaller extent in the VRNN architecture, but is largely ameliorated by the regularizing affects of the upstream latent space dependency which enables communication between time steps z_t.

To illustrate the effects of considering an ensemble of NNs we show in figure A1 the ensemble mean and one standard deviation spread of the global pdf predictions for an ensemble size of 6 – the same as the results presented in §5. To illustrate the variance in both the bulk and the tails of the distribution we plot these on a linear and log scale – for the latter we zoom in on the tail of the pdf to best illustrate the ensemble variance. There are two main conclusions to be drawn from this figure. First, the variance is modest but meaningful – most notably the tails of the pdf – indicating that an ensem-

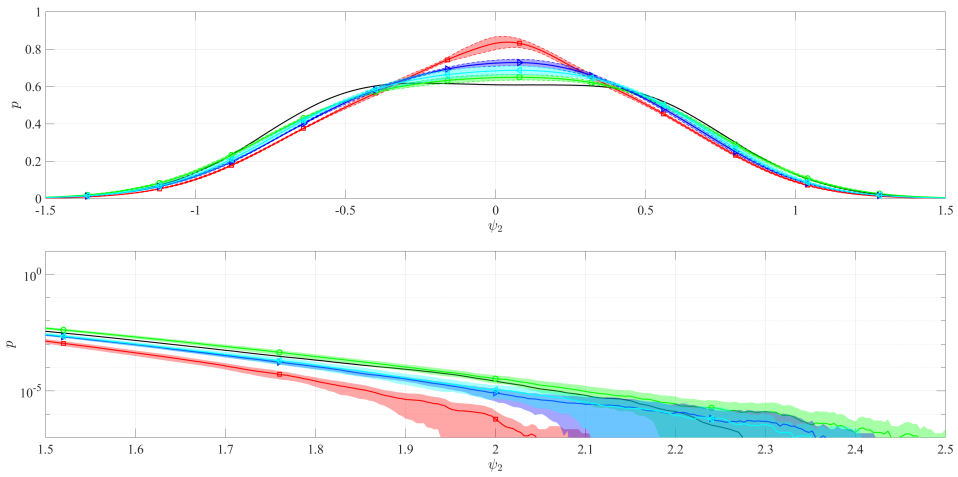
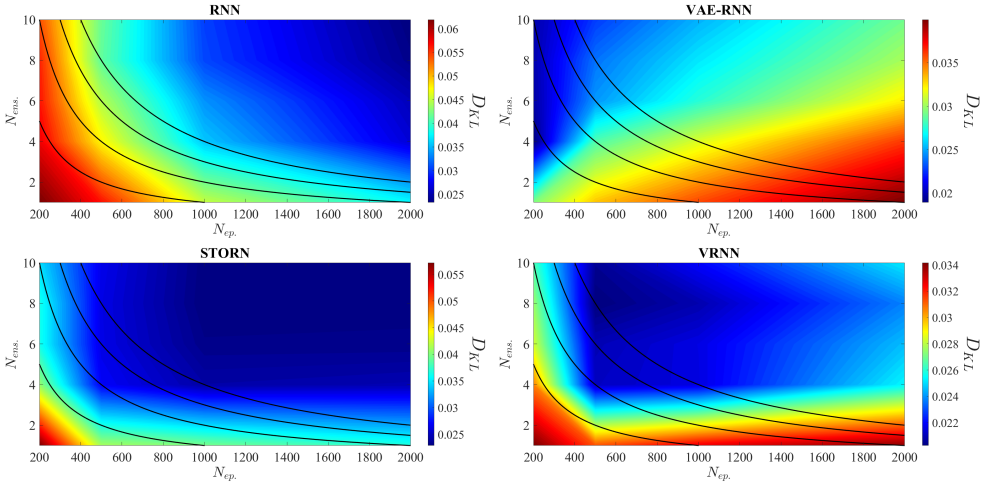
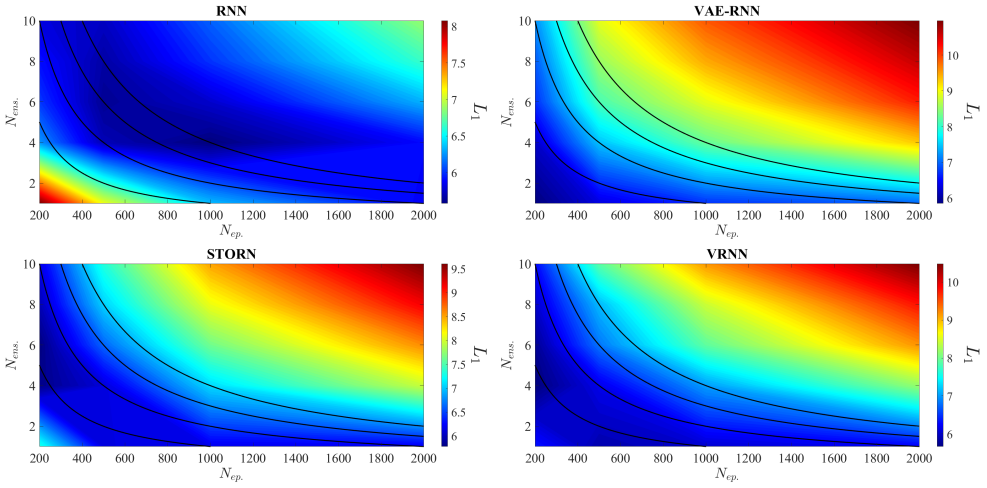


Figure A1: Global pdf of ψ_2 on a linear scale (upper panel) and log-scale (lower panel). RD (solid black), RNN (red), VAE-RNN (green), STORN (blue), VRNN (teal). Shaded area signifies ensemble mean $\pm$ 1 standard deviation.

ble analysis does improve the predictive capabilities of the ML correction. Second, the variance is very similar across all architecture types.



(a)



(b)

Figure A2: Average error in prediction as a function of ensemble size and number of epochs trained. KL divergence (a), L1 norm of log pdf error (b), integrated variance (c), and integrated log variance (d). Lines indicate 1000, 2000, 3000, and 4000 total epochs (left to right).

A2 Additional Results

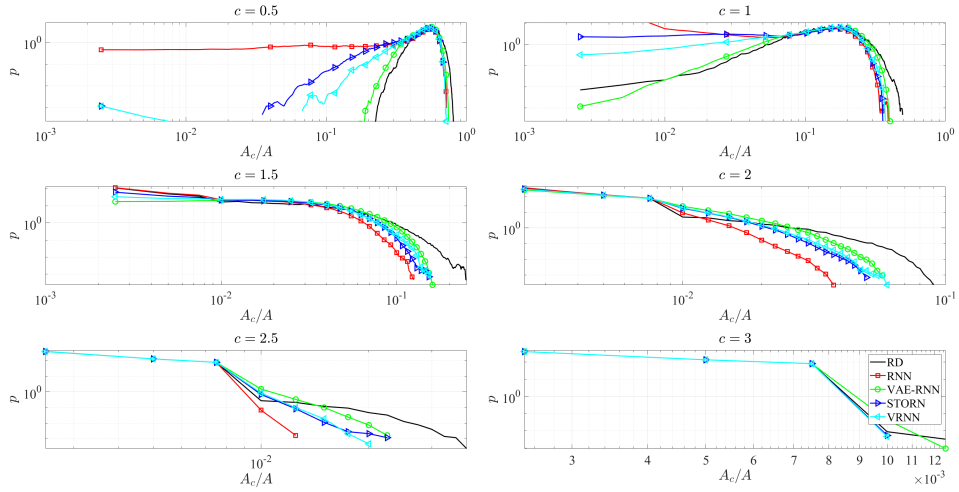


Figure A3: Pdf of fraction of domain over which $|\psi_1|$ exceeds fixed threshold c for range of $c \in [0, 2]$. RD (solid black), RNN (red), VAE-RNN (green), STORN (blue), VRNN (teal).

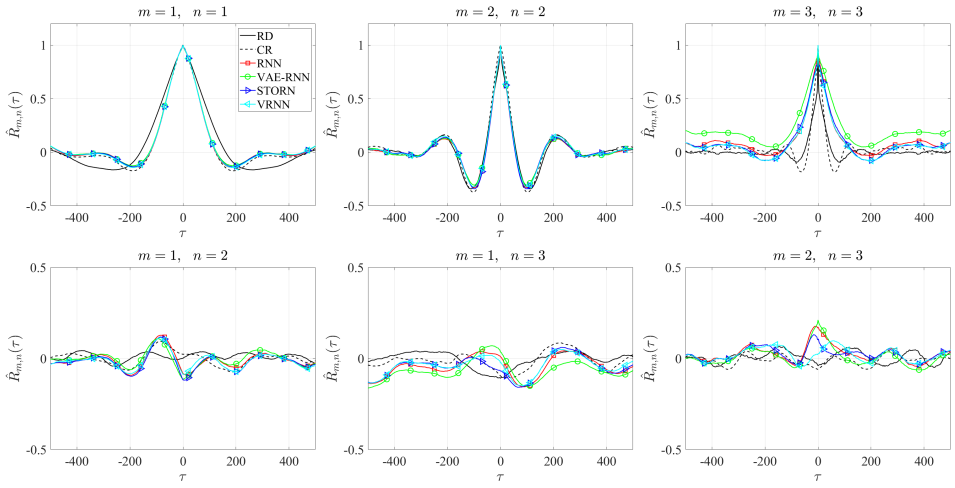
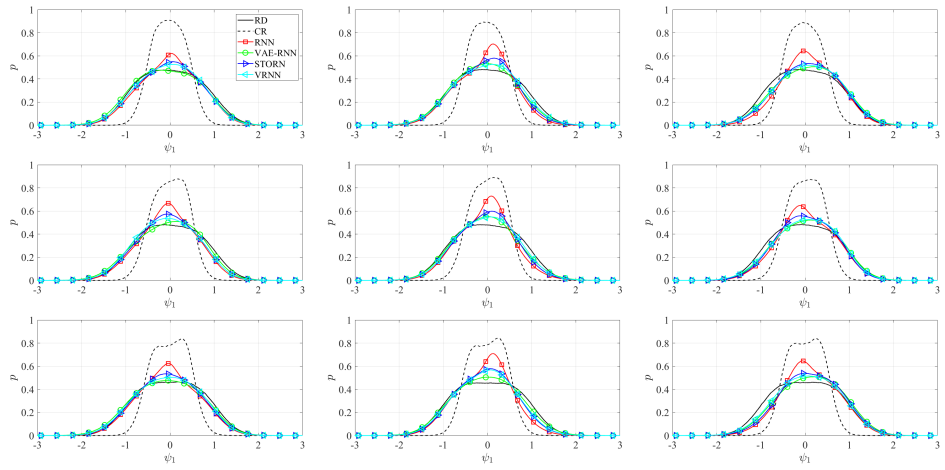
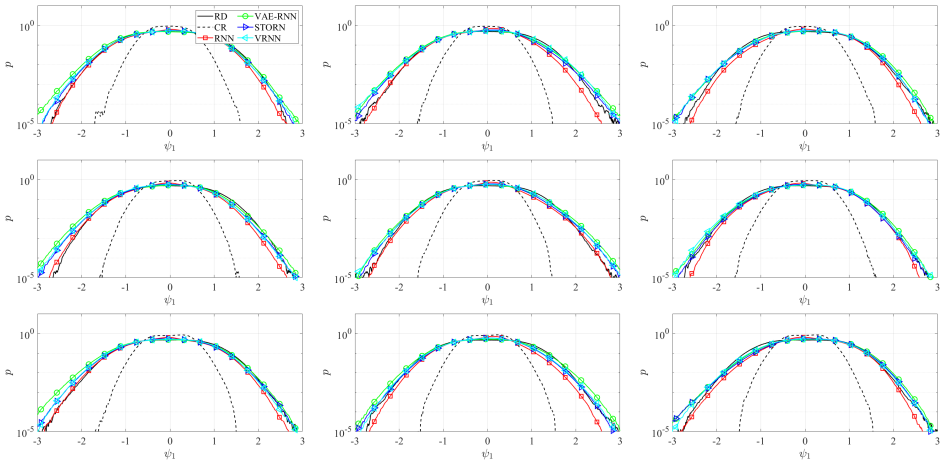


Figure A4: Normalized correlation between three largest zonally constant Fourier modes of ψ_1. RD (solid black), CR (dashed black), RNN (red), VAE-RNN (green), STORN (blue), VRNN (teal).



(a)



(b)

Figure A5: Regional pdf (a) and log-pdf (b) of ψ_1. RD (solid black), CR (dashed black), RNN (red), VAE-RNN (green), STORN (blue), VRNN (teal).

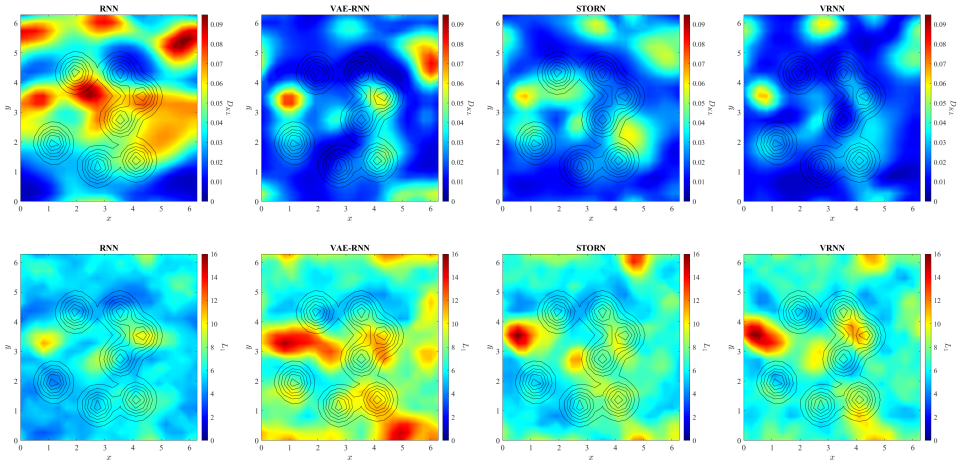


Figure A6: Spatial distribution of KL divergence (upper panel) and L_1 metric (lower panel) for ψ_1. From left to right: RNN, VAE-RNN, STORN, VRNN. The topography profile is show in black.

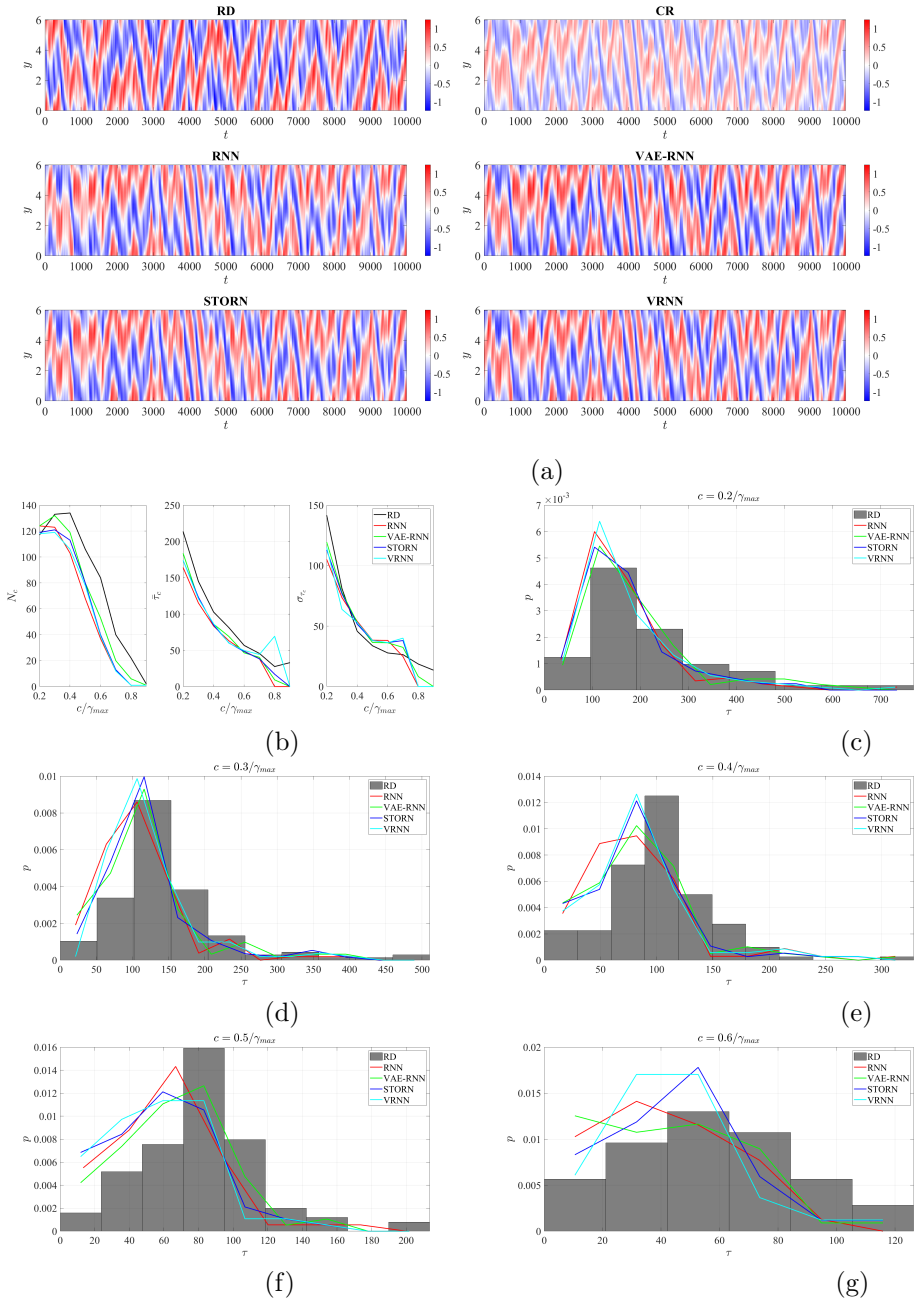


Figure A7: Zonally averaged stream function $\bar{\psi}_1$ (a). Frequency, expected duration, and variance of high amplitude excursions of γ_1 as a function of threshold c (b). Probability density function of τ for fixed values of c (c-g).

References

Arbabi, H., & Sapsis, T. (2022, June). Generative Stochastic Modeling of Strongly Nonlinear Flows with Non-Gaussian Statistics. *SIAM/ASA Journal on Uncertainty Quantification*, *10*(2), 555–583. doi: 10.1137/20M1359833

Arcomano, T., Szunyogh, I., Wikner, A., Hunt, B. R., & Ott, E. (2023). A Hybrid Atmospheric Model Incorporating Machine Learning Can Capture Dynamical Processes Not Captured by Its Physics-Based Component. *Geophysical Research Letters*, *50*(8), e2022GL102649. doi: 10.1029/2022GL102649

Arcomano, T., Szunyogh, I., Wikner, A., Pathak, J., Hunt, B. R., & Ott, E. (2022). A Hybrid Approach to Atmospheric Modeling That Combines Machine Learning With a Physics-Based Numerical Model. *Journal of Advances in Modeling Earth Systems*, *14*(3), e2021MS002712. doi: 10.1029/2021MS002712

Barriopedro, D., Fischer, E. M., Luterbacher, J., Trigo, R. M., & García-Herrera, R. (2011, April). The Hot Summer of 2010: Redrawing the Temperature Record Map of Europe. *Science*, *332*(6026), 220–224. doi: 10.1126/science.1201224

Barthel, B. (2024, September). *Software and dataset for Barthel Sorensen et. al. (2024)[Dataset].* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.13833073` doi: 10.5281/zenodo.13833073

Barthel Sorensen, B., Charalampopoulos, A., Zhang, S., Harrop, B. E., Leung, L. R., & Sapsis, T. P. (2024). A Non-Intrusive Machine Learning Framework for Debiasing Long-Time Coarse Resolution Climate Simulations and Quantifying Rare Events Statistics. *Journal of Advances in Modeling Earth Systems*, *16*(3). doi: 10.1029/2023MS004122

Bayer, J., & Osendorfer, C. (2015, March). *Learning Stochastic Recurrent Networks.* arXiv. doi: 10.48550/arXiv.1411.7610

Bevacqua, E., Suarez-Gutierrez, L., Jézéquel, A., Lehner, F., Vrac, M., Yiou, P., & Zscheischler, J. (2023, April). Advancing research on compound weather and climate events via large ensemble model simulations. *Nature Communications*, *14*(1), 2145. doi: 10.1038/s41467-023-37847-5

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, *619*(7970), 533–538.

Blanchard, A., Parashar, N., Dodov, B., Lessig, C., & Sapsis, T. (2022, December). A Multi-Scale Deep Learning Framework for Projecting Weather Extremes. In *Climate Change AI.* Climate Change AI.

Bollt, E. (2021). On explaining the surprising success of reservoir computing forecaster of chaos? the universal machine learning dynamical system with contrast to var and dmd. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *31*(1).

Bora, A., Shukla, K., Zhang, S., Harrop, B., Leung, R., & Karniadakis, G. E. (2023, February). Learning bias corrections for climate models using deep neural operators.
doi: 10.48550/arXiv.2302.03173

Boral, A., Wan, Z. Y., Zepeda-Núñez, L., Lottes, J., Wang, Q., Chen, Y.-F., ... Sha, F. (2023). Neural ideal large eddy simulation: Modeling turbulence with neural stochastic differential equations. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in neural information processing systems* (Vol. 36, pp. 69270–69283). Curran Associates, Inc. Retrieved from `https://proceedings.neurips.cc/paper_files/paper/2023/file/dabaded617b3be96c3ed161498a7d71c-Paper-Conference.pdf`

Brenowitz, N. D., & Bretherton, C. S. (2019). Spatially Extended Tests of a Neural Network Parametrization Trained by Coarse-Graining. *Journal of Advances in Modeling Earth Systems*, *11*(8), 2728–2744. doi: 10.1029/2019MS001711

Bretherton, C. S., Henn, B., Kwa, A., Brenowitz, N. D., Watt-Meyer, O., McGibbon, J., ... Harris, L. (2022). Correcting Coarse-Grid Weather and Climate

Models by Machine Learning From Global Storm-Resolving Simulations. *Journal of Advances in Modeling Earth Systems*, *14*(2), e2021MS002794. doi: 10.1029/2021MS002794

Buchta, D. A., & Zaki, T. A. (2021, June). Observation-infused simulations of high-speed boundary-layer transition. *Journal of Fluid Mechanics*, *916*, A44. doi: 10.1017/jfm.2021.172

Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014, October). On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. doi: 10.48550/arXiv.1409.1259

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014, September). *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation.* arXiv. doi: 10.48550/arXiv.1406.1078

Christopoulos, C., Lopez-Gomez, I., Beucler, T., Cohen, Y., Kawczynski, C., Dunbar, O., & Schneider, T. (2024). Online learning of entrainment closures in a hybrid machine learning parameterization. *ESS Open Archive preprint*. Retrieved from `http://dx.doi.org/10.22541/essoar.171804905.55213571/v1` doi: 10.22541/essoar.171804905.55213571/v1

Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A., & Bengio, Y. (2016, April). A Recurrent Latent Variable Model for Sequential Data. doi: 10.48550/arXiv.1506.02216

Clark, S. K., Brenowitz, N. D., Henn, B., Kwa, A., McGibbon, J., Perkins, W. A., ... Harris, L. M. (2022). Correcting a 200 km Resolution Climate Model in Multiple Climates by Machine Learning From 25 km Resolution Simulations. *Journal of Advances in Modeling Earth Systems*, *14*(9), e2022MS003219. doi: 10.1029/2022MS003219

Cohen, Y., Lopez-Gomez, I., Jaruga, A., He, J., Kaul, C. M., & Schneider, T. (2020, 9). Unified entrainment and detrainment closures for extended eddy-diffusivity mass-flux schemes. *Journal of Advances in Modeling Earth Systems*, *12*, e2020MS002162. Retrieved from `https://doi.org/10.1029/2020MS002162` (https://doi.org/10.1029/2020MS002162) doi: https://doi.org/10.1029/2020MS002162

Deser, C., Phillips, A., Bourdette, V., & Teng, H. (2012). Uncertainty in climate change projections: the role of internal variability. *Climate Dynamics*, *38*, 527-546. Retrieved from `https://doi.org/10.1007/s00382-010-0977-x` doi: 10.1007/s00382-010-0977-x

Dimet, F.-X. L., & Talagrand, O. (1986, January). Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. , *38*(2), 97. doi: 10.3402/tellusa.v38i2.11706

Dresdner, G., Kochkov, D., Norgaard, P., Zepeda-Núñez, L., Smith, J. A., Brenner, M. P., & Hoyer, S. (2022). Learning to correct spectral methods for simulating turbulent flows. *arXiv preprint arXiv:2207.00556*.

Eyring, V., Collins, W. D., Gentine, P., Barnes, E. A., Barreiro, M., Beucler, T., ... Zanna, L. (2024). Pushing the frontiers in climate modelling and analysis with machine learning. *Nature Climate Change*, *14*, 916-928. Retrieved from `https://doi.org/10.1038/s41558-024-02095-y` doi: 10.1038/s41558-024-02095-y

Fischer, E. M., Sippel, S., & Knutti, R. (2021, August). Increasing probability of record-shattering climate extremes. *Nature Climate Change*, *11*(8), 689–695. doi: 10.1038/s41558-021-01092-9

Fraccaro, M. (2018). *Deep latent variable models for sequential data* (phd). Technical University of Denmark.

Fraccaro, M., Sønderby, S. K., Paquet, U., & Winther, O. (2016, November). *Sequential Neural Models with Stochastic Layers.* arXiv. doi: 10.48550/arXiv.1605.07571

Fulton, D. J., Clarke, B. J., & Hegerl, G. C. (2023, May). Bias Correcting Climate Model Simulations Using Unpaired Image-to-Image Translation Networks. *Artificial Intelligence for the Earth Systems*, *2*(2). doi: 10.1175/AIES-D-22-0031.1

Gedon, D., Wahlström, N., Schön, T. B., & Ljung, L. (2021, June). *Deep State Space Models for Nonlinear System Identification.* arXiv. doi: 10.48550/arXiv.2003.14162

Geirinhas, J. L., Russo, A., Libonati, R., Sousa, P. M., Miralles, D. G., & Trigo, R. M. (2021, February). Recent increasing frequency of compound summer drought and heatwaves in Southeast Brazil. *Environmental Research Letters*, *16*(3), 034036. doi: 10.1088/1748-9326/abe0eb

Graves, A. (2014, June). *Generating Sequences With Recurrent Neural Networks.* arXiv. doi: 10.48550/arXiv.1308.0850

Graves, A., Fernández, S., Liwicki, M., Bunke, H., & Schmidhuber, J. (2007, January). Unconstrained On-line Handwriting Recognition with Recurrent Neural Networks. In (Vol. 20). doi: 10.1057/9780230226203.3287

Graves, A., Mohamed, A.-r., & Hinton, G. (2013, March). *Speech Recognition with Deep Recurrent Neural Networks.* arXiv. doi: 10.48550/arXiv.1303.5778

Guillaumin, A. P., & Zanna, L. (2021, 9). Stochastic-deep learning parameterization of ocean momentum forcing. *Journal of Advances in Modeling Earth Systems*, *13*, e2021MS002534. Retrieved from `https://doi.org/10.1029/2021MS002534` (https://doi.org/10.1029/2021MS002534) doi: https://doi.org/10.1029/2021MS002534

Hara, M., & Kokubu, H. (2022). Learning dynamics by reservoir computing (in memory of prof. pavol brunovskỳ). *Journal of Dynamics and Differential Equations*, 1–26.

Hochreiter, S., & Schmidhuber, J. (1997, November). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. doi: 10.1162/neco.1997.9.8.1735

HovmÖller, E. (1949). The trough-and-ridge diagram. *Tellus*, *1*(2), 62–66. Retrieved from `https://doi.org/10.3402/tellusa.v1i2.8498` doi: 10.3402/tellusa.v1i2.8498

Huang, Z., Zhong, L., Ma, Y., & Fu, Y. (2021, May). Development and evaluation of spectral nudging strategy for the simulation of summer precipitation over the Tibetan Plateau using WRF (v4.0). *Geoscientific Model Development*, *14*(5), 2827–2841. (Publisher: Copernicus GmbH) doi: 10.5194/gmd-14-2827-2021

Iglesias-Suarez, F., Gentine, P., Solino-Fernandez, B., Beucler, T., Pritchard, M., Runge, J., & Eyring, V. (2024, 2). Causally-informed deep learning to improve climate models and projections. *Journal of Geophysical Research: Atmospheres*, *129*, e2023JD039202. Retrieved from `https://doi.org/10.1029/2023JD039202` doi: https://doi.org/10.1029/2023JD039202

Jiang, R., Lu, P. Y., Orlova, E., & Willett, R. (2023). Training neural operators to preserve invariant measures of chaotic attractors. *arXiv preprint arXiv:2306.01187*.

Kingma, D. P., & Welling, M. (2022, December). Auto-Encoding Variational Bayes. doi: 10.48550/arXiv.1312.6114

Kochkov, D., Smith, J. A., Alieva, A., Wang, Q., Brenner, M. P., & Hoyer, S. (2021, May). Machine learning–accelerated computational fluid dynamics. *Proceedings of the National Academy of Sciences*, *118*(21). (Publisher: National Academy of Sciences Section: Physical Sciences) doi: 10.1073/pnas.2101784118

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., ... others (2023). Neural general circulation models. *arXiv preprint arXiv:2311.07222*.

Kooperman, G. J., Pritchard, M. S., Burt, M. A., Branson, M. D., & Randall, D. A. (2016). Robust effects of cloud superparameterization on simulated daily rainfall intensity statistics across multiple versions of the community earth

system model. *Journal of Advances in Modeling Earth Systems*, *8*(1), 140-165. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015MS000574 doi: https://doi.org/10.1002/2015MS000574

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., ... others (2022). Graphcast: Learning skillful medium-range global weather forecasting. *arXiv preprint arXiv:2212.12794*.

Lehmann, J., Coumou, D., & Frieler, K. (2015). Increased record-breaking precipitation events under global warming. *Climatic Change*, *132*, 501-515. Retrieved from https://doi.org/10.1007/s10584-015-1434-y doi: 10.1007/s10584-015-1434-y

Li, L., Carver, R., Lopez-Gomez, I., Sha, F., & Anderson, J. (2024, 6). Generative emulation of weather forecast ensembles with diffusion models. *Science Advances*, *10*, eadk4489. Retrieved from https://doi.org/10.1126/sciadv.adk4489 (doi: 10.1126/sciadv.adk4489) doi: 10.1126/sciadv.adk4489

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., & Anandkumar, A. (2021, May). Fourier Neural Operator for Parametric Partial Differential Equations.
doi: 10.48550/arXiv.2010.08895

Li, Z., Liu-Schiaffini, M., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., ... Anandkumar, A. (2022). Learning chaotic dynamics in dissipative systems. *Advances in Neural Information Processing Systems*, *35*, 16768–16781.

Li, Z., Peng, W., Yuan, Z., & Wang, J. (2022, November). Fourier neural operator approach to large eddy simulation of three-dimensional turbulence. *Theoretical and Applied Mechanics Letters*, *12*(6), 100389. doi: 10.1016/j.taml.2022.100389

Liu, C., Xiao, Q., & Wang, B. (2008, September). An Ensemble-Based Four-Dimensional Variational Data Assimilation Scheme. Part I: Technical Formulation and Preliminary Test. *Monthly Weather Review*, *136*(9), 3363–3373. doi: 10.1175/2008MWR2312.1

Lopez-Gomez, I., Christopoulos, C., Langeland Ervik, H. L., Dunbar, O. R. A., Cohen, Y., & Schneider, T. (2022). Training physics-based machine-learning parameterizations with gradient-free ensemble kalman methods. *Journal of Advances in Modeling Earth Systems*, *14*(8), e2022MS003105. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2022MS003105 (e2022MS003105 2022MS003105) doi: https://doi.org/10.1029/2022MS003105

Lopez-Gomez, I., Cohen, Y., He, J., Jaruga, A., & Schneider, T. (2020). A generalized mixing length closure for eddy-diffusivity mass-flux schemes of turbulence and convection. *Journal of Advances in Modeling Earth Systems*, *12*, e2020MS002161. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002161 (e2020MS002161 10.1029/2020MS002161) doi: https://doi.org/10.1029/2020MS002161

Lucarini, V., Faranda, D., Freitas, A., Freitas, J., Holland, M., Kuna, T., ... Vaienti, S. (2016). *Extremes and Recurrence in Dynamical Systems*. doi: 10.1002/9781118632321

Mathews, A., Francisquez, M., Hughes, J. W., Hatch, D. R., Zhu, B., & Rogers, B. N. (2021). Uncovering turbulent plasma dynamics via deep learning from partial observations. *Physical Review E*, *104*(2), 025205.

McGibbon, J., Brenowitz, N. D., Cheeseman, M., Clark, S. K., Dahm, J. P. S., Davis, E. C., ... Fuhrer, O. (2021, July). fv3gfs-wrapper: a Python wrapper of the FV3GFS atmospheric model. *Geoscientific Model Development*, *14*(7), 4401–4409. doi: 10.5194/gmd-14-4401-2021

McGibbon, J. J., Clark, S. K., Henn, B., Kwa, A., Watt-Meyer, O., Perkins, W. A., & Bretherton, C. S. (2023, July). Global Precipitation Correction Across a Range of Climates Using CycleGAN [preprint].

doi: 10.22541/essoar.168881853.36817507/v1

Meehl, G. A., & Tebaldi, C. (2004). More intense, more frequent, and longer lasting heat waves in the 21st century. *Science*, *305*(5686), 994-997. Retrieved from `https://www.science.org/doi/abs/10.1126/science.1098704` doi: 10.1126/science.1098704

Miguez-Macho, G., Stenchikov, G. L., & Robock, A. (2005, April). Regional Climate Simulations over North America: Interaction of Local Processes with Improved Large-Scale Flow. *Journal of Climate*, *18*(8), 1227–1246. (Publisher: American Meteorological Society Section: Journal of Climate) doi: 10.1175/JCLI3369.1

Mons, V., Chassaing, J. C., Gomez, T., & Sagaut, P. (2016, July). Reconstruction of unsteady viscous flows using data assimilation schemes. *Journal of Computational Physics*, *316*, 255–280. doi: 10.1016/j.jcp.2016.04.022

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of artificial intelligence research*, *11*, 169–198.

Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *International conference on machine learning* (pp. 1310–1318).

Pathak, J., Lu, Z., Hunt, B. R., Girvan, M., & Ott, E. (2017). Using machine learning to replicate chaotic attractors and calculate Lyapunov exponents from data. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *27*(12).

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., ... others (2022a). Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., ... Anandkumar, A. (2022b, February). *FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators.* arXiv. doi: 10.48550/arXiv.2202.11214

Perkins-Kirkpatrick, S. E., & Lewis, S. C. (2020). Increasing trends in regional heatwaves. *Nature Communications*, *11*, 3357. Retrieved from `https://doi.org/10.1038/s41467-020-16970-7` doi: 10.1038/s41467-020-16970-7

Platt, J. A., Penny, S. G., Smith, T. A., Chen, T.-C., & Abarbanel, H. D. (2023). Constraining chaos: Enforcing dynamical invariants in the training of recurrent neural networks. *arXiv preprint arXiv:2304.12865*.

Qi, D., & Majda, A. J. (2020, January). Using machine learning to predict extreme events in complex systems. *Proceedings of the National Academy of Sciences*, *117*(1), 52–59. (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.1917285117

Randall, D., Khairoutdinov, M., Arakawa, A., & Grabowski, W. (2003). Breaking the cloud parameterization deadlock. *Bulletin of the American Meteorological Society*, *84*(11), 1547 - 1564. Retrieved from `https://journals.ametsoc.org/view/journals/bams/84/11/bams-84-11-1547.xml` doi: 10.1175/BAMS-84-11-1547

Rasp, S., Pritchard, M. S., & Gentine, P. (2018, September). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, *115*(39), 9684–9689. doi: 10.1073/pnas.1810286115

Raymond, C., Horton, R. M., Zscheischler, J., Martius, O., AghaKouchak, A., Balch, J., ... White, K. (2020, July). Understanding and managing connected extreme events. *Nature Climate Change*, *10*(7), 611–621. doi: 10.1038/s41558-020-0790-4

Robinson, A., Lehmann, J., Barriopedro, D., Rahmstorf, S., & Coumou, D. (2021, October). Increasing heat and rainfall extremes now far outside the historical climate. *npj Climate and Atmospheric Science*, *4*(1), 1–4. doi: 10.1038/s41612-021-00202-w

Sanderse, B., Stinis, P., Maulik, R., & Ahmed, S. E. (2024). Scientific machine learning for closure models in multiscale problems: A review. *arXiv preprint arXiv:2403.02913*.

Sapsis, T. P. (2021). Statistics of Extreme Events in Fluid Flows and Waves. *Annual Review of Fluid Mechanics*, *53*(1), 85–111. (_eprint: https://doi.org/10.1146/annurev-fluid-030420-032810) doi: 10.1146/annurev-fluid-030420-032810

Schiff, Y., Wan, Z. Y., Parker, J. B., Hoyer, S., Kuleshov, V., Sha, F., & Zepeda-Núñez, L. (2024). Dyslim: Dynamics stable learning by invariant measure for chaotic systems. In *Forty-first international conference on machine learning.*

Schneider, T., Behera, S., Boccaletti, G., Deser, C., Emanuel, K., Ferrari, R., ... Yamagata, T. (2023, September). Harnessing AI and computing to advance climate modelling and prediction. *Nature Climate Change*, *13*(9), 887–889. doi: 10.1038/s41558-023-01769-3

Schneider, T., Lan, S., Stuart, A., & Teixeira, J. (2017). Earth System Modeling 2.0: A Blueprint for Models That Learn From Observations and Targeted High-Resolution Simulations. *Geophysical Research Letters*, *44*(24), 12,396–12,417. doi: 10.1002/2017GL076101

Schneider, T., Teixeira, J., Bretherton, C. S., Brient, F., Pressel, K. G., Schär, C., & Siebesma, A. P. (2017, January). Climate goals and computing the future of clouds. *Nature Climate Change*, *7*(1), 3–5. doi: 10.1038/nclimate3190

Storch, H. v., Langenberg, H., & Feser, F. (2000, October). A Spectral Nudging Technique for Dynamical Downscaling Purposes. *Monthly Weather Review*, *128*(10), 3664–3673. (Publisher: American Meteorological Society Section: Monthly Weather Review) doi: 10.1175/1520-0493(2000)128⟨3664:ASNTFD⟩2.0.CO;2

Strogatz, S. H. (2018). *Nonlinear dynamics and chaos with student solutions manual: With applications to physics, biology, chemistry, and engineering.* CRC press.

Sun, J., Zhang, K., Wan, H., Ma, P.-L., Tang, Q., & ZHANG, S. (2019, December). Impact of Nudging Strategy on the Climate Representativeness and Hindcast Skill of Constrained EAMv1 Simulations. *Journal of Advances in Modeling Earth Systems*, *11*. doi: 10.1029/2019MS001831

Sutskever, I., Hinton, G. E., & Taylor, G. W. (2008). The Recurrent Temporal Restricted Boltzmann Machine. In *Advances in Neural Information Processing Systems* (Vol. 21). Curran Associates, Inc.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014, December). *Sequence to Sequence Learning with Neural Networks.* arXiv. doi: 10.48550/arXiv.1409.3215

Vandal, T., Kodra, E., Ganguly, S., Michaelis, A., Nemani, R., & Ganguly, A. R. (2017). Deepsd: Generating high resolution climate change projections through single image super-resolution. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (p. 1663–1672). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/3097983.3098004` doi: 10.1145/3097983.3098004

Wan, Z. Y., Baptista, R., Boral, A., Chen, Y.-F., Anderson, J., Sha, F., & Zepeda-Núñez, L. (2023). Debias coarsely, sample conditionally: Statistical downscaling through optimal transport and probabilistic diffusion models. In *Thirty-seventh conference on neural information processing systems.* Retrieved from `https://openreview.net/forum?id=5NxJuc0T1P`

Wan, Z. Y., Baptista, R., Chen, Y.-f., Anderson, J., Boral, A., Sha, F., & Zepeda-Núñez, L. (2023, May). *Debias Coarsely, Sample Conditionally: Statistical Downscaling through Optimal Transport and Probabilistic Diffusion Models.* arXiv. (arXiv:2305.15618 [physics]) doi: 10.48550/arXiv.2305.15618

Wang, M., Wang, Q., & Zaki, T. A. (2019, November). Discrete adjoint of fractional-step incompressible Navier-Stokes solver in curvilinear coordinates and application to data assimilation. *Journal of Computational Physics*, *396*, 427–450. doi: 10.1016/j.jcp.2019.06.065

Watt-Meyer, O., Brenowitz, N. D., Clark, S. K., Henn, B., Kwa, A., McGibbon, J., ... Bretherton, C. S. (2021). Correcting Weather and Climate Models by Machine Learning Nudged Historical Simulations. *Geophysical Research Letters*, *48*(15), e2021GL092555. doi: 10.1029/2021GL092555

Wikner, A., Harvey, J., Girvan, M., Hunt, B. R., Pomerance, A., Antonsen, T., & Ott, E. (2022, December). Stabilizing Machine Learning Prediction of Dynamics: Noise and Noise-inspired Regularization. doi: 10.48550/arXiv.2211.05262

Wilby, R. L., Wigley, T. M. L., Conway, D., Jones, P. D., Hewitson, B. C., Main, J., & Wilks, D. S. (1998). Statistical downscaling of general circulation model output: A comparison of methods. *Water Resources Research*, *34*(11), 2995-3008. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/98WR02577` doi: https://doi.org/10.1029/98WR02577

Yuval, J., O'Gorman, P. A., & Hill, C. N. (2021). Use of Neural Networks for Stable, Accurate and Physically Consistent Parameterization of Subgrid Atmospheric Processes With Good Performance at Reduced Precision. *Geophysical Research Letters*, *48*(6), e2020GL091363. doi: 10.1029/2020GL091363

Yuval, J., & O'Gorman, P. A. (2020, July). Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. *Nature Communications*, *11*(1), 3295. doi: 10.1038/s41467-020-17142-3

Zhang, H., Harlim, J., & Li, X. (2021, December). Error bounds of the invariant statistics in machine learning of ergodic Itô diffusions. *Physica D: Nonlinear Phenomena*, *427*, 133022. doi: 10.1016/j.physd.2021.133022

Zhang, S., Harrop, B., Leung, L. R., Charalampopoulos, A.-T., Barthel Sorensen, B., Xu, W., & Sapsis, T. (2024). A Machine Learning Bias Correction on Large-Scale Environment of High-Impact Weather Systems in E3SM Atmosphere Model. *Journal of Advances in Modeling Earth Systems*, *16*(8), e2023MS004138.

Zscheischler, J., Westra, S., van den Hurk, B. J. J. M., Seneviratne, S. I., Ward, P. J., Pitman, A., ... Zhang, X. (2018, June). Future climate risk from compound events. *Nature Climate Change*, *8*(6), 469–477. doi: 10.1038/s41558-018-0156-3