# The High-Frequency and Rare Events Barriers to Neural Closures of Atmospheric Dynamics

Mickaël D. Chekroun[*]
*Department of Atmospheric and Oceanic Sciences, University of California, Los Angeles, CA and Department of Earth and Planetary Sciences, Weizmann Institute of Science, Rehovot 76100, Israel*

Honghu Liu
*Department of Mathematics, Virginia Tech, Blacksburg, VA 24061, USA*

Kaushik Srinivasan
*Department of Atmospheric and Oceanic Sciences, University of California, Los Angeles, CA 90095-1565, USA*

James C. McWilliams
*Department of Atmospheric and Oceanic Sciences and Institute of Geophysics and Planetary Physics, University of California, Los Angeles, CA 90095-1565, USA*

(Dated: September 21, 2023)

Neural parameterizations and closures of climate and turbulent models have raised a lot of interest in recent years. In this short paper, we point out two fundamental problems in this endeavour, one tied to sampling issues due to rare events, and the other one tied to the high-frequency content of slow-fast solutions which constitute an intrinsic barrier to neural closure of such multiscale systems. We argue that the atmospheric 1980 Lorenz model, a truncated model of the Primitive Equations—the fuel engine of climate models—serves as a remarkable metaphor to illustrate these fundamental issues.

## I. INTRODUCTION

Atmospheric and oceanic flows constrained by Earth's rotation satisfy an approximately geostrophic momentum balance on larger scales, associated with slow evolution on time scales of days, but they also exhibit fast inertia-gravity wave oscillations. The problems of identifying the slow component (e.g., for weather forecast initialization [1–4]) and of characterizing slow-fast interactions are central to geophysical fluid dynamics, and the former was first coined as a slow manifold problem by Leith [5]. The L63 model [6] famous for its chaotic strange attractor is a paradigm for the geostrophic component, while the L80 model [7] is its paradigmatic successor both for the generalization of slow balance and for slow-fast coupling.

The explosion of machine learning (ML) methods provides an unprecedented opportunity to analyze data and accelerate scientific progress. A variety of ML methods have emerged for solving dynamical systems [8–10], predicting [11] or discovering [12] them from data. For larger scale problems, much effort has been devoted lately to the learning of neural subgrid-scale parameterizations in coarse-resolution climate models [13] but yet the lack of interpretability and reliability prevents a widespread adoption so far [14, 15].

In parallel, the learning of stable neural parameterizations of small scales or neglected variables has progressed remarkably for the closure of fluid models in turbulent regimes such as the forced Navier-Stokes equations or quasi-geostrophic flow models on a $\beta$-plane; see [16–20].

Nevertheless, the case of the Primitive Equations (PE) has its mysteries yet untouched by neural parameterizations. This study is aimed at pointing out potential challenges that one may have to face for the efficient neural closures of the PE. To do so, the L80 model, a truncated model of the PE, serves as a remarkable metaphor to illustrate some fundamental issues in this task, as explained in this article.

At small Rossby numbers, the solutions to the L80 model remain entirely slow for all time (i.e. dominated by Rossby waves) whereas fast oscillations get superimposed to such slow solutions as the Rossby number is further increased. Such a spontaneous emergence of fast oscillations, tied to inertia gravity waves (IGWs) evolving on top of the slow geostrophic motion, complicate severely the closure problem [21, 22].

Regimes with a multiscale mixture of slow and fast dynamics without timescale separation are not only intimate to the L80 model. They have been observed in PE models accounting for a greater amount of multiscale interactions as conspicuously generated by fronts and jets [23, 24], and in cloud-resolving models in which large-scale convectively coupled gravity waves spontaneously develop [25]. Regions of organized convective activity in the tropics generate also gravity waves leading to a spectrum that contains notable contributions from horizontal wavelengths of 10 km through to scales beyond 1000 km [26] and such IGWs have been also identified from satellite observation of continental shallow convective cumulus forming organized mesoscale patterns over vegetated areas [27].

In fact, IGWs can be energetic on surprisingly large scales. For instance, in certain regions of the oceans too, Rocha et al. have shown in [28] that IGWs can account for roughly half of the near-surface kinetic energy at scales between 10 and 40 km. Thus, geophysical kinetic energy spectra can exhibit a band of wavenumbers within which waves and turbulence are equally energetic. A painful consequence of this length-scale overlap is that perturbation methods such as that of Wentzel, Kramers, and Brillouin (WKB) [29] do not apply to all scales [30].

In the L80 model, regimes with time-scale overlap in which the solutions exhibit a mixture of slow motion punctuated by

[*] mchekroun@atmos.ucla.edu

bursts of fast IGWs containing a large fraction of the total energy (referred to as high-low frequency (HLF) solutions) were shown to be responsible for a severe breakdown of slaving relationships [21], that occurs at large Rossby numbers.

Only recently, the generic elements for solving such hard closure problems with a lack of timescale separation, have been identified [22]. Key to its solution is the Balance Equation (BE) manifold [31, 32] as rooted in the works of Monin [33], Charney and Bolin [1, 34], and Lorenz [35]. The BE manifold has been shown to provide, even for large Rossby numbers, the slow trend motion of HLF solutions that optimally averages out the fast oscillations [21, 36], while the fast motions can be efficiently modeled by means of networks of stochastic oscillators [22].

In this article we emphasize the fundamental issues encountered by neural networks to parameterize unequivocally the slow balance motion and restore the fast oscillations. The drawbacks shown by neural parameterizations of the L80 slow motion are discussed in Sec. II. In particular, the sensitivity displayed by the learnt neural parameterizations is shown to be attributed to rare event statistics in Sec. III. Finally, the issues met by neural networks to learn both the slow and high-frequency content of the L80 model's solutions and its damaging consequences for closure, are then pointed out in Sec. IV.

## II. LEARNING SLOW NEURAL CLOSURE: SENSITIVITY

The L80 model, obtained by Lorenz in [7] as a nine-dimensional truncation of the PE onto three Fourier modes with low wavenumbers, can be written as:

$$\begin{aligned} a_i \frac{\mathrm{d}x_i}{\mathrm{d}t} &= -\nu_0 a_i^2 x_i - c(a_i - a_k)x_j y_k + c(a_i - a_j)y_j x_k \\ &\quad + a_i b_i x_j x_k - 2c^2 y_j y_k + a_i(y_i - z_i), \\ a_i \frac{\mathrm{d}y_i}{\mathrm{d}t} &= -a_k b_k x_j y_k - a_j b_j y_j x_k + c(a_k - a_j)y_j y_k \\ &\quad - a_i x_i - \nu_0 a_i^2 y_i, \\ \frac{\mathrm{d}z_i}{\mathrm{d}t} &= g_0 a_i x_i - b_k x_j(z_k - h_k) - b_j(z_j - h_j)x_k \\ &\quad + c y_j(z_k - h_k) - c(z_j - h_j)y_k - \kappa_0 a_i z_i + F_i, \end{aligned} \tag{1}$$

whose model parameters are described in [7, 21].

The above equations are written for each cyclic permutation of the set of indices $(1, 2, 3)$, namely, for $(i, j, k)$ in $\{(1, 2, 3), (2, 3, 1), (3, 1, 2)\}$. The model variables $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ are amplitudes for the divergent velocity potential, streamfunction, and dynamic height, respectively.

In this model, the square root of the constant forcing $F_1$ can be interpreted as the Rossby number; see [32] and [21, Eq. (2.4)]. Transitions to chaos occur as the Rossby number $Ro$ is increased [21, 32]. As mentioned above, at small Rossby numbers, the solutions to the L80 model are dominated by Rossby waves and thus remain entirely slow for all time. As identified in [21], when the Rossby number is further increased beyond a critical Rossby number $Ro^*$, fast IGW oscillations emerge spontaneously and are superimposed on the slow component of the solutions. For such regimes, the aforementioned BE manifold on which balanced solutions lie [21, 31, 32] is no longer able to parameterize fully the L80 dynamics since a substantial portion of it, associated with the IGWs, evolves transversally to the BE manifold [22, Fig. 3]. These regimes with energetic bursts of IGWs lie beyond the parameter range explored by Lorenz in his original 1980 article [7] and beyond other regimes with exponential smallness of IGW amplitudes as studied in subsequent Lorenz 86 models [37–40] and the full primitive equations [41] at smaller Rossby numbers [42].

The HLF solutions considered in this study are obtained for such a critical parameter regime where $Ro > Ro^*$. They correspond to those of [22, Fig. 7]); see Appendix A for details. We first analyze the ability of neural parameterizations to learn the slow motion of the L80 dynamics in the HLF regime. To do so, we preprocess the target variables $\boldsymbol{x}$ and $\boldsymbol{z}$ to be parameterized by applying a low-pass filter in order to extract the slow motion. In that respect, a simple moving average is adopted with a window size equal to $T_{GW}$, the dominant period of the gravity waves. The results are shown in Fig. 1A for the $z_3$-variable for which we observe that the low-pass filtered solution almost coincides this way with the BE parameterization $\boldsymbol{z}_{BE}(t) = G(\boldsymbol{y}(t))$ with $\boldsymbol{y}(t)$ denoting the $\boldsymbol{y}$-component of the HLF solution to the L80 model.

Exploiting the structure of the L80 model, the BE manifold as built in two consecutive steps—first parameterizing $\boldsymbol{z}$ as a function $G$ of $\boldsymbol{y}$ and then $\boldsymbol{x}$ as a function of $\boldsymbol{y}$ and $G(\boldsymbol{y})$—was shown to provide very good closure skills for a wide range of parameter regimes [32]; see Appendix B and [21] for details. To benefit from this a priori knowledge and to favor comparison with the BE manifold, we thus parallel this BE manifold construction to learn our NN parameterizations. In that respect, we first learn a neural parameterization of $\boldsymbol{z}$ in terms of $\boldsymbol{y}$ and then learn a neural parameterization of $\boldsymbol{x}$ that is conditioned on the former. In that respect, a multilayer perceptron (MLP), $\boldsymbol{\mathcal{Z}_\theta}$, is learnt with $\boldsymbol{y}$ as input and the filtered $\boldsymbol{z}$-variable as output (see (2)). Once this MLP is learnt, a subsequent MLP, $\boldsymbol{\mathcal{X}_\theta}$, is learnt with $(\boldsymbol{y}, \boldsymbol{\mathcal{Z}_\theta}(\boldsymbol{y}))$ as input and the filtered $\boldsymbol{x}$-variable as output.

The structure of our MLPs is standard. Each neural parameterization, e.g. $\boldsymbol{z}$ in terms of $\boldsymbol{y}$, is sought by means of an MLP with $L$ hidden layers of $p$ neurons each. It boils down to find

$$\boldsymbol{\mathcal{Z}_\theta}(\boldsymbol{y}) = \mathcal{N}_{\text{out}} \circ \mathcal{N}_L \circ \cdots \circ \mathcal{N}_1 \circ \mathcal{N}_{\text{in}}(\boldsymbol{y}), \tag{2}$$

in which $\mathcal{N}_{\text{in}}$ (resp. $\mathcal{N}_{\text{out}}$) constitute the input (resp. output) layer, while $\mathcal{N}_k$ is a mapping from $\mathbb{R}^p$ (the space of neurons) onto itself, given by $\mathcal{N}_k(\xi) = \Psi_k(\mathbf{W}_k \xi + \mathbf{b}_k)$ ($\xi$ in $\mathbb{R}^p$) where $\Psi_k$ is a $p$-dimensional elementwise function, i.e. a function that applies a (scalar) activation function to each of its inputs individually, and the $\mathbf{W}_k$ and $\mathbf{b}_k$ denote respectively the weight matrices and bias vectors to be learnt. In (2), the subscript $\boldsymbol{\theta}$ denotes the collection of these parameters. In this work, the nonlinear activation function is a simple $\tanh$ function, and the input and output layers consist just of linear normalization and reversal operations. It turns out that NNs with one hidden layer and 5 neurons are sufficient to obtain loss functions with a small residual; see Table I.

Based on our approach paralleling the BE manifold construction, we learn our neural parameterizations for the L80 model, through the following consecutive minimizations.

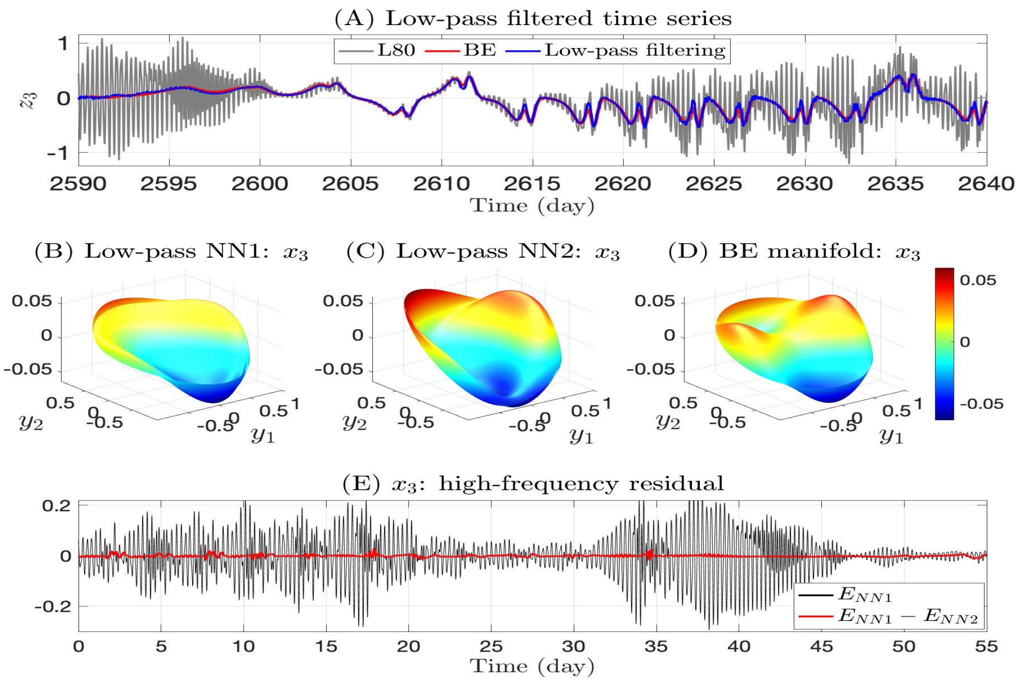

FIG. 1: **Panel A**: Illustration, for the $z_3$-variable, of the BE manifold's ability in capturing the L80 model's slow motion. See [32] and Appendix B for a derivation. **Panels (B) and (C):** Neural parameterizations $\boldsymbol{\mathcal{X}}_{\boldsymbol\theta}^3$ for the $x_3$-variable, as learnt through random selection (NN$_1$)/predefined selection (NN$_2$). Visualized here as mappings from $(y_1, y_2)$ onto the unit sphere in $\mathbb{R}^3$. **Panels (D):** Same visualization adopted for the BE manifold. **Panels (E)**: High-frequency residual $E_{NN_1}(t)$ for $x_3$ (black) given by (5) and its difference with $E_{NN_2}(t)$ (red).

TABLE I: **Loss function evaluations for two neural networks**. The loss functions (3) for $\boldsymbol z$ and (4) for $\boldsymbol x$, are minimized using two neural networks, NN$_1$ and NN$_2$ providing each a parameterization $(\boldsymbol{\mathcal{Z}}_{\boldsymbol\theta}, \boldsymbol{\mathcal{X}}_{\boldsymbol\theta})$, differing only in the way the training, validation, and testing sets are selected. In each case, the aspect ratios between these sets are the same.

| Epochs | 10 | 50 | 100 | 300 | 500 | 1000 |
|---|---|---|---|---|---|---|
| NN$_1$ loss for $\boldsymbol z$ (random) ($\times 10^{-3}$) | 11.17 | 9.26 | 9.26 | 9.26 | 9.26 | 9.26 |
| NN$_2$ loss for $\boldsymbol z$ (predefined) ($\times 10^{-3}$) | 13.70 | 10.66 | 9.28 | 9.05 | 9.05 | 9.05 |
| NN$_1$ loss for $\boldsymbol x$ (random) ($\times 10^{-4}$) | 1.76 | 1.38 | 1.35 | 1.33 | 1.32 | 1.32 |
| NN$_2$ loss for $\boldsymbol x$ (predefined) ($\times 10^{-4}$) | 1.62 | 1.37 | 1.33 | 1.31 | 1.31 | 1.31 |

First, given a discrete set of time instants $t_j$, one minimizes

$$\mathcal{L}_{\boldsymbol\theta}(\boldsymbol z;\boldsymbol y)=\sum_j \left\|\boldsymbol z_{t_j}-\boldsymbol{\mathcal{Z}}_{\boldsymbol\theta}(\boldsymbol y_{t_j})\right\|^2, \tag{3}$$

in which $\boldsymbol z$ is filtered (in time) while $\boldsymbol y$ is not, followed by the minimization of

$$\mathcal{L}_{\boldsymbol\theta}(\boldsymbol x;(\boldsymbol y,\boldsymbol{\mathcal{Z}}_{\boldsymbol\theta_1^*}(\boldsymbol y)))=\sum_j \left\|\boldsymbol x_{t_j}-\boldsymbol{\mathcal{X}}_{\boldsymbol\theta}(\boldsymbol y_{t_j},\boldsymbol{\mathcal{Z}}_{\boldsymbol\theta_1^*}(\boldsymbol y_{t_j}))\right\|^2, \tag{4}$$

with $\boldsymbol x$ filtered and where $\boldsymbol{\mathcal{Z}}_{\boldsymbol\theta_1^*}$ denotes the optimal parameterization obtained after minimization of (3). Note that the use of the unfiltered $\boldsymbol y$-component of the HLF solution albeit containing fast oscillations, is key for the discovery of a proper parameterization of the slow motion. For instance, if one replaces it with a filtered version of $\boldsymbol y$ such as shown in Fig. 2A by the blue curve for $y_3$, the resulting closure fails to capture the lobe dynamics by producing an unrealistic quasiperiodic behavior not even reminiscent to a quasiperiodic behavior that would lie nearby in the parameter space as documented in [32]; see red curves in Fig. 2.

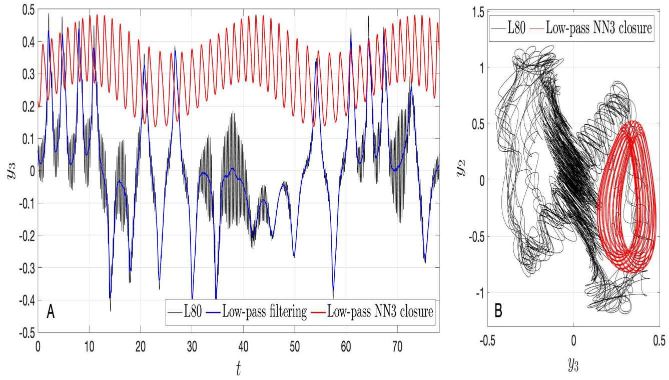

FIG. 2: **False quasiperiodicity produced by a slow neural closure**. Here, the slow neural closure Eq. (6) is driven by $\boldsymbol{\mathcal{Z}}_{\boldsymbol\theta}$ and $\boldsymbol{\mathcal{X}}_{\boldsymbol\theta}$ that are trained using a low-pass filtered version of $\boldsymbol y(t)$ (blue curve in Panel (A)) unlike in Eq. (6) where the slow neural closures are trained using $\boldsymbol y(t)$, unfiltered.

To assess whether a neural parameterization is successful in capturing the slow motion, we evaluate also the following *high-frequency (HF) residual*

$$E^j_{NN}(t) = x_j(t) - \boldsymbol{\mathcal{X}}^j_{\boldsymbol{\theta}_2^*}(\boldsymbol{y}(t), \boldsymbol{\mathcal{Z}}_{\boldsymbol{\theta}_1^*}(\boldsymbol{y}(t))), \qquad (5)$$

in which the $x_j(t)$ and $\boldsymbol{y}(t)$ are *both* unfiltered. For an NN with small residual, $E^j_{NN}(t)$ is typically void of slow oscillations (see Fig. 1E) with mean $\langle E^j_{NN}\rangle \approx 0$ for each $1 \leq j \leq 3$. Fig. 1 shows such a situation for two NNs, NN$_1$ and NN$_2$, learnt through different modalities of selection of training, validation, and testing sets. The learnt neural parameterizations are thus able to capture, offline, the slow motion as the BE manifold does. Noteworthy though is to observe that the underlying manifolds associated with the neural parameterizations exhibit noticeable differences with the BE manifold. Denoting by $\boldsymbol{\mathcal{X}}^j_{\boldsymbol{\theta}_2^*}$ (resp. $\boldsymbol{\mathcal{Z}}^j_{\boldsymbol{\theta}_1^*}$) the $j$th component of the neural parameterization of $x_j$ (resp. $z_j$) for $1 \leq j \leq 3$, we embark into plotting its level sets on a sphere of radius $r$ in order to reveal some of the geometric attributes associated with $\boldsymbol{\mathcal{X}}^j_{\boldsymbol{\theta}_2^*}$ (resp. $\boldsymbol{\mathcal{Z}}^j_{\boldsymbol{\theta}_1^*}$), as the latter is a scalar field of $\mathbb{R}^3$. The interest of doing so is that for any given radius $r$, the level set of $\boldsymbol{\mathcal{X}}^j_{\boldsymbol{\theta}_2^*}$ (resp. $\boldsymbol{\mathcal{Z}}^j_{\boldsymbol{\theta}_1^*}$) on the sphere, $y_1^2 + y_2^2 + y_3^2 = r^2$, can be visualized as a 2D surface that maps e.g $(y_1, y_2)$ to $x_j$ (resp. $z_j$). The results are shown in Figs. 1B, C and D, for which $r = 1$. These visualizations allow us to reveal noticeable differences in the minimizers (and thus parameterization formulas) whereas the loss function evaluations differ only by $1\%$ (Table I) and the difference between their high-frequency residuals is small (see red curve in Fig. 1E).

These geometric offline differences hide more profound consequences when the neural parameterizations are used online, for closure. As explained below, the sensitivity of online predictions that are tied to sampling issues is indeed observed. In that respect, recall that a common practice to train NNs is to divide the dataset into three subsets. The first subset is the training set, which is used for computing the loss function's gradient and updating the network weights and biases.

The second subset is the validation set. It corresponds to the second dataset over which the prediction skills of the fitted model are assessed. The error on the validation set is monitored during the training process to provide an unbiased evaluation while tuning the model's hyperparameters. When the network begins to overfit the data, the error on the validation set typically begins to rise after an initial decrease. The network parameters are saved at the minimum. It gives then the ”final model” that is tested over the test set that is typically a holdout dataset not used as a validation nor a training set.

The parameterization NN$_1$ shown in Fig. 1A is learnt through a random selection while NN$_2$ through a predefined selection. In each case, ratios for training, testing, and validation are 0.7, 0.15, and 0.15, respectively. The total length of the training is 700 days. Given the same input and target data, the minimal values of the loss functions (3)-(4) for NN$_1$ and NN$_2$ are reported in Table I, across epochs. Already after 500 epochs, one observes that the loss function evaluations differ only by $1\%$ between the random or predefined selection protocol of the training, validation, and testing sets.

We now discuss the sensitivity issue of online predictions driven by such neural parameterizations close in terms of their loss function evaluation. This point is illustrated in Fig. 3. There, we show online prediction corresponding to a given slow NN-parameterization $(\boldsymbol{\mathcal{X}}_{\boldsymbol{\theta}_2^*}, \boldsymbol{\mathcal{Z}}_{\boldsymbol{\theta}_1^*})$ learnt by minimization of the loss functions (3)-(4), namely the solution to the slow neural closure

$$a_i \frac{\mathrm{d}y_i}{\mathrm{d}t} = -a_k b_k \boldsymbol{\mathcal{X}}^j_{\boldsymbol{\theta}_2^*}(\boldsymbol{y}, \boldsymbol{\mathcal{Z}}_{\boldsymbol{\theta}_1^*}(\boldsymbol{y}))y_k - a_j b_j y_j \boldsymbol{\mathcal{X}}^k_{\boldsymbol{\theta}_2^*}(\boldsymbol{y}, \boldsymbol{\mathcal{Z}}_{\boldsymbol{\theta}_1^*}(\boldsymbol{y})) + c(a_k - a_j)y_j y_k - a_i \boldsymbol{\mathcal{X}}^i_{\boldsymbol{\theta}_2^*}(\boldsymbol{y}, \boldsymbol{\mathcal{Z}}_{\boldsymbol{\theta}_1^*}(\boldsymbol{y})) - \nu_0 a_i^2 y_i, \qquad (6)$$

which is obtained by replacing the $x_\ell$ in the $\boldsymbol{y}$-equation of the L80 model by its neural parameterization, either NN$_1$ or NN$_2$.

The attractor corresponding to the slow NN$_1$-closure (with random selection) differs clearly from that of slow NN$_2$-closure (with predefined selection) in spite of convergence and closeness of the loss functions at their respective minimal value; see Fig. 3B. Both predict periodic orbits with different attributes, one self-intersecting in the $(y_2, y_3)$-plane (NN$_1$), the other without intersection point (NN$_2$).

A closer inspection at these topological differences reveals in the time domain that the slow NN$_1$-closure is able to capture more accurately the low-frequency content of certain temporal patterns exhibited by the HLF solutions of the L80 model compared to the slow NN$_2$-closure; blue vs red curves in Fig. 3A. We argue next that such a sensitivity between online solutions takes its root in the rare events tied to the irregular transitions exhibited by the HLF solutions to the L80 model that spoils the offline learning.

In contrast, at lower Rossby numbers, for regimes devoid of fast oscillations such as shown in Fig. 4D below corresponding to $F_1 = 6.97 \times 10^{-2}$ in the L80 model, neural closures of high-accuracy are easily accessible with skills comparable to those obtained with the BE manifold; see Fig. 5. As explained next, the reasons for this success lie in the absence of high-frequencies in the solutions to parameterize and in the absence of rare events in the statistics of lobe transitions.

## III. IRREGULAR TRANSITIONS, RARE EVENTS AND LEARNING CONSEQUENCES

The sensitivity in the online capture of the low-frequency content between two nearby neural parameterizations (as measured through their loss functions), calls for more understanding. Keeping in mind that the differences pointed out in Fig. 3 are only tied to the way the training, validation, and test sets are chosen, we perform a statistical analysis of key features of the L80 dynamics in the HLF regime. In particular, we focus on the irregular lobes' transitions exhibited by HLF solutions, and for comparison, we analyze the lobes' transitions in the slow

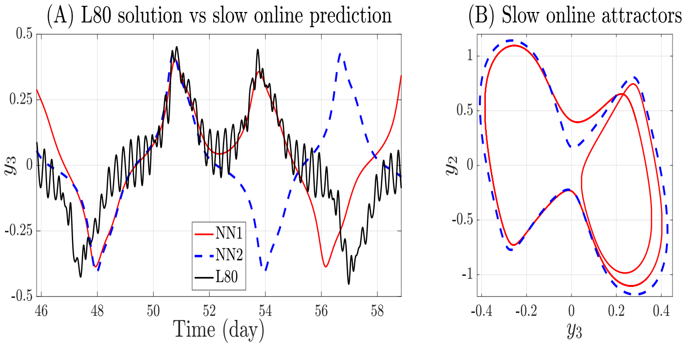

FIG. 3: **Sensitivity of the slow neural closures.** Here, $\text{NN}_1$ and $\text{NN}_2$ differ only in their training modalities. $\text{NN}_1$ is learnt from random selection of the training, validation, and testing sets, and $\text{NN}_2$ from a predefined selection with the same aspect ratios; see Text. The corresponding loss functions differ by 1% (see Table I), while the dynamical differences of the online predictions are substantial.

chaotic regime of Fig. 4D for which neural parameterizations do not suffer from sensitivity and have no difficulty in learning the proper closure. Indeed, as shown in Fig. 5 for the slow chaotic regime of Fig. 4D, neural closures of high-accuracy are easily accessible with skills comparable to those obtained with the BE manifold.

To further probe into the statistics of lobes' transitions in the slow chaotic and HLF regimes, we perform, for each regime, a high-resolution and long-term integration of the L80 model corresponding to a 500-yr simulation ($\sim 730,000 \times T_{GW}$) with a time-stepping of 0.75 min.

In either case, the L80 attractor exhibits two lobes as visible in e.g. the $(y_2, y_3)$-projection; see Figs. 4A and D. These lobes are essentially separated by the vertical line $y_3 = 0$. Numerical integration of the L80 model reveals that the visit of the right lobe comes with $y_3(t)$ getting greater than some threshold value $y_b$, while the visit of the left lobe comes with $y_3(t)$ getting smaller than $y_a = -y_b$. A close inspection of the solution in the HLF case reveals that the choice of $y_b = 0.2$ constitutes a good one to identify the sojourn of the dynamics within one lobe from the other. This choice leads furthermore to an interval $(-y_b, y_b)$ that provides a good bound of the bursts of fast oscillations crossing the vertical line $y_3 = 0$ in the $(y_2, y_3)$-plane (”grey” zone).

To count the transitions from one lobe to the other one thus proceeds as follows. Given our 500-yr long simulation of $y_3(t)$ we first find the local maxima and minima that are above $y_b$ and below $y_a$, respectively. No transition occurs between consecutive such local maxima or minima. A transition occurs only when a local maximum above $y_b$ is immediately followed by a local minimum below $y_a$ or vice versa. If a local maximum is immediately followed by a local minimum, the intermediate time instant at which the trajectory goes below zero is identified as the transition instant, and the other way around if a local minimum is immediately followed by a local maximum. These transition times characterized this way allow us to count the sojourn times in a lobe and display the distribution of these sojourn times shown in Figs. 4C and F.

These lobe sojourn time distributions reveal a striking difference between the HLF and slow chaotic regimes. In the HLF case, we observe indeed that the solution can stay in one lobe for a period of time that can be arbitrarily long (see solution's segment between $t = 763$ and $t = 893$ shown in blue in Fig. 4B) albeit of probability of occurrence vanishing exponentially as shown in Fig. 4C. As a comparison, the transitions between the attractor's lobes occur at a much more regular pace in the slow chaotic regime (see Fig. 4E) in which the solutions to the L80 model are void of fast oscillations. In this case, the distribution of sojourn times drops quickly below a 60-day duration barrier (Fig. 4F).

One can thus argue that such rare events with an exponential distribution are troublesome for the derivation of a reliable slow neural closure. They add diversity in the temporal distribution of the time series patterns that explain the sensitivity results shown in Fig. 3. Indeed, a random selection of the training set may contain temporal episodes that are more skewed towards one lobe than those in a predefined set, confusing this way the learning procedure.

## IV. THE HIGH-FREQUENCY BARRIER TO NEURAL CLOSURE

We address now the issue of direct parameterizations of HLF solutions containing a multiscale mixture of slow and fast motions. To do so, we learn an MLP for $\boldsymbol{x}(t)$, denoted by $\boldsymbol{\mathcal{V}}_{\boldsymbol{\theta}}$, with (the unfiltered) $\boldsymbol{y}(t)$-variable of the L80 model (1), as input, and the *unfiltered* $\boldsymbol{x}$-component, $\boldsymbol{x}(t)$, as output. Note that unlike the slow NN-parameterizations above, the parameterization $\boldsymbol{\mathcal{V}}_{\boldsymbol{\theta}}$ aims at parameterizing $\boldsymbol{x}(t)$ directly as a nonlinear mapping of $\boldsymbol{y}(t)$ without conditioning on $\boldsymbol{z}(t)$ nor filtering of any sort. The corresponding closure, called a vanilla NN-closure, consists then of Eq. (6) in which $\boldsymbol{\mathcal{X}}_{\boldsymbol{\theta}_2^*}(\boldsymbol{y}, \boldsymbol{\mathcal{Z}}_{\boldsymbol{\theta}_1^*}(\boldsymbol{y}))$ is replaced by $\boldsymbol{\mathcal{V}}_{\boldsymbol{\theta}^*}(\boldsymbol{y})$, obtained after minimization of the following $L^2$-loss function

$$\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{x}; \boldsymbol{y}) = \sum_j \left\| \boldsymbol{x}_{t_j} - \boldsymbol{\mathcal{V}}_{\boldsymbol{\theta}}(\boldsymbol{y}_{t_j}) \right\|^2, \qquad (7)$$

for which the target variable $\boldsymbol{x}(t)$ is *unfiltered*, i.e. containing a mixture of fast and slow oscillations. To address this more challenging problem we use MLPs with a larger capacity either with more neurons and/or layers.

Our experiments reveal that an NN with one hidden layer and 20 neurons turns out to provide the best closure results. As a comparison, we show in Fig. 6 the simulated time series from vanilla NN-closures in four different settings. For the one with one hidden layer and 20 neurons, the corresponding vanilla neural closure is able to apprehend to a certain extent the complexity of the temporal patterns exhibited by the HLF solution (see Fig. 7A-B), albeit completely failing to predict the high-frequency content i.e. the physics of IGWs as revealed by the power spectral density (PSD) comparison shown in Fig. 8.

The failure experienced by vanilla NN-closure in capturing the IGWs dynamics mirrors the *spectral bias problem* exhibited by NNs for function fitting [43]; feedforward NNs prioritizing the learning of the low-frequency features. We emphasize though that here the difficulty is of next order than for function fitting. The problem is indeed to properly learn offline the neglected variables along with their high-frequency

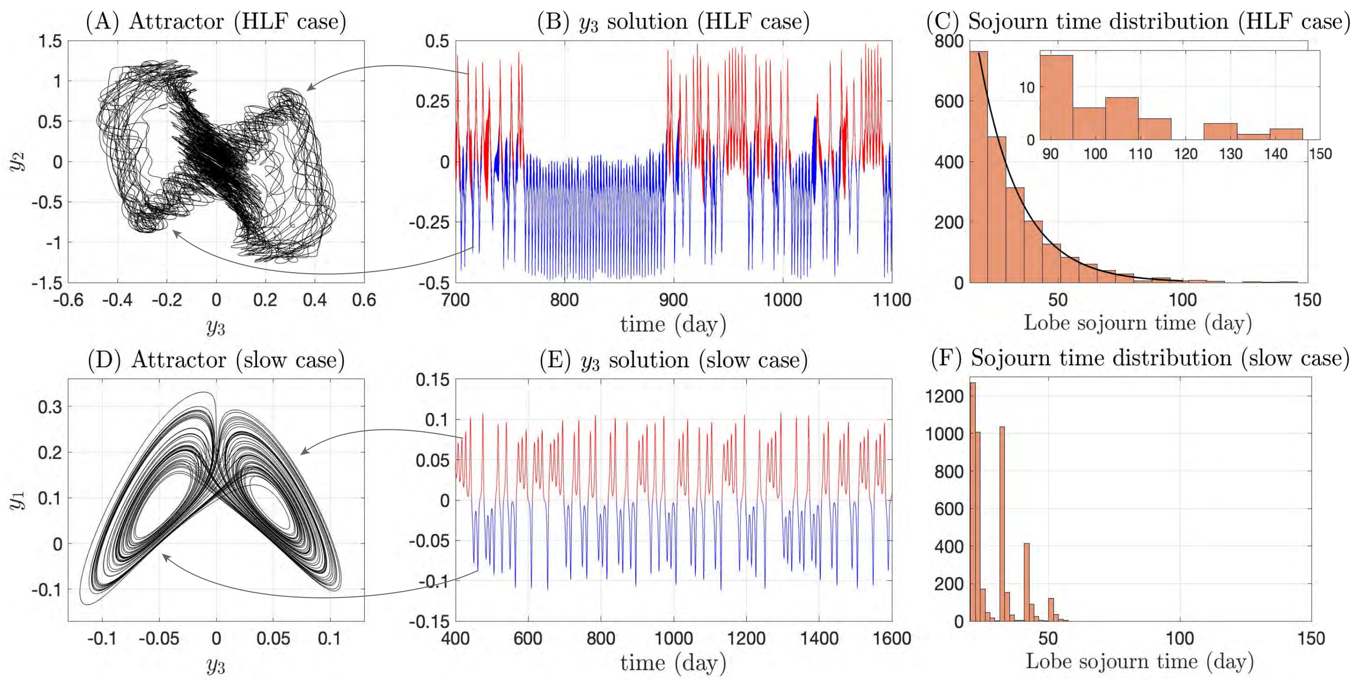

FIG. 4: **Panel A:** Attractor in the HLF case. **Panel B:** The sojourn episodes within one particular lobe are marked by different colours. Here, the parameters are those used in Lorenz's original paper [7] except $F_1 = 0.3027$ in Eq. (1). **Panel C: Lobe sojourn time distributions.** The exponential fit is calculated over 500 yr-long simulation of Eq. (1) and is shown by the black curve $f(t) = ae^{bt}$ with $a = 2292$ and $b = -6.05 \times 10^{-2}$ with $t$ in day. The inset in panel C shows a magnification of the distribution for the rare and large sojourn times. **Panels E and F:** Same as panels B and C except that $F_1 = 6.97 \times 10^{-2}$, corresponding to the slow chaotic regime shown in panel D in which the solutions are void of fast oscillations. In this regime, no rare event statistics emerge.

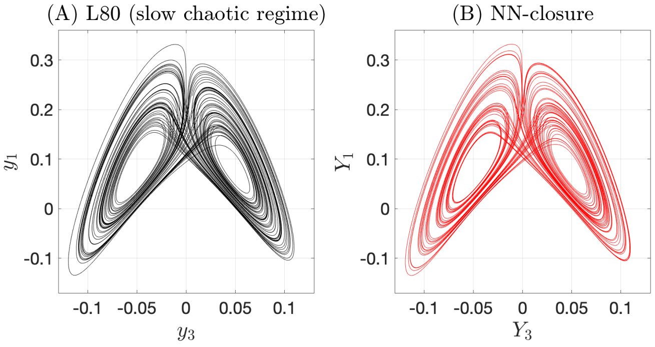

FIG. 5: **The L80 attractor vs. its NN-closure in the slow chaos regime.** Here $F_1 = 6.97 \times 10^{-2}$ in the L80 model, which corresponds to the slow chaos case shown in Fig. 4D and in [21, Fig. 7].

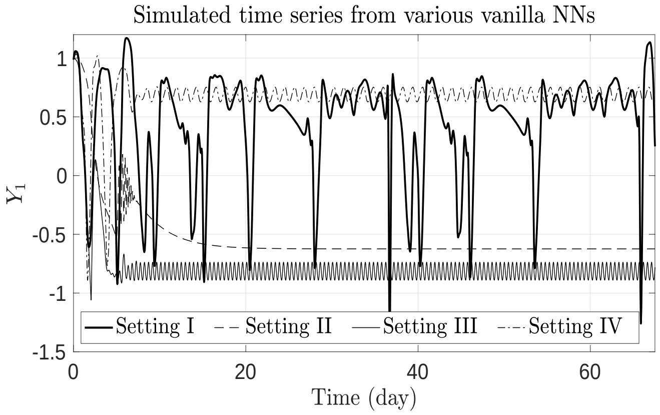

FIG. 6: **Simulated time series from vanilla NN-closures in four different settings**. **Setting I** (same as used for the results shown in Fig. 7): one hidden layer with 20 neurons (thick solid line); **Setting II**: two hidden layers with 5 neurons in each layer (dashed line); **Setting III**: two hidden layers with 10 neurons in each layer (light solid line); **Setting IV**: two hidden layers with 20 neurons in each layer (dash-dotted line). The corresponding loss function values are given in Table II.

content, so that the online solution via neural closure reproduces both the slow and fast motions of the original dynamics. Even global geometric features are misrepresented by vanilla NN-closures, with e.g. distortion of the attractor and breakdown of symmetry compared to the L80 attractor; see Fig. 7C.

Finally, it is worth mentioning that even larger vanilla NNs do not necessarily help. Increasing the number of hidden layers or neurons may drive down the loss function value, but not necessarily improve the performance of the corresponding NN closure. For instance, a vanilla NN, $\boldsymbol{\mathcal{V}}_{\boldsymbol{\theta}}$, with 5 hidden layers and 20 neurons predicts an unrealistic periodic orbit (of small amplitude) when iterated through time-stepping online in the neural closure and tends to exaggerate the high-frequency content of the parameterized solutions offline; see Fig. 9 and Table II.

These results, tied with the multiscale nature of the L80 regime analyzed with its rare events, show that the L80 model provides a cautionary illustration for the challenges of using machine learning for deriving accurate dynamical-system closures for geophysical fluid dynamics. Because rare events and extreme statistics are likely to play a more and more dominant role in a changing climate [44–48], this study shows that in

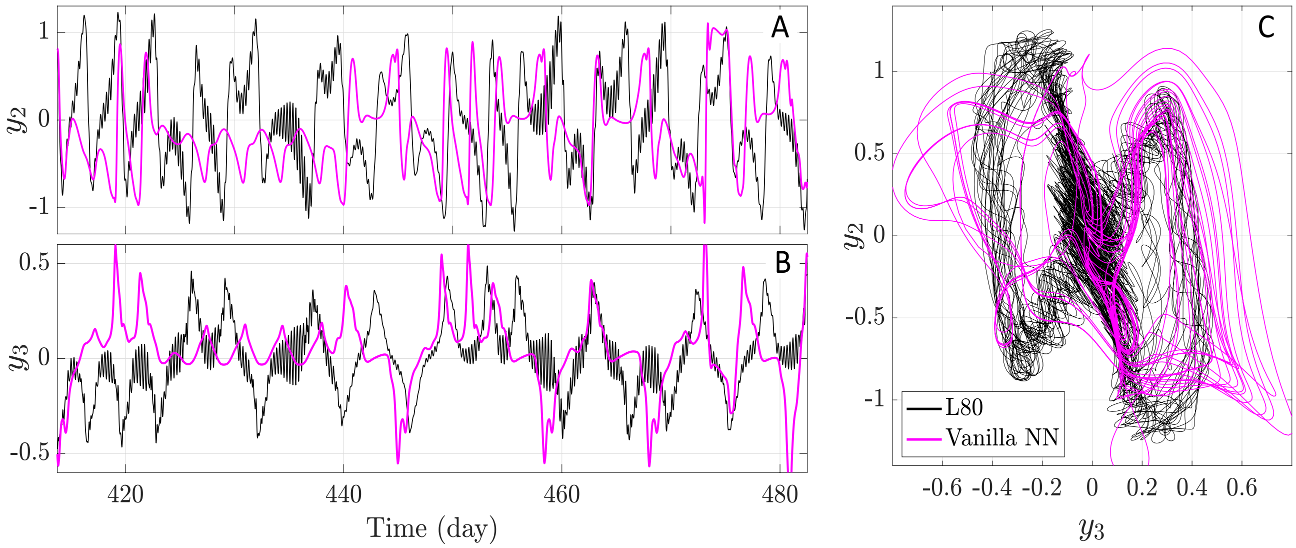

FIG. 7: **Vanilla NN-closure vs L80 dynamics**. Failure to capture the high-frequency content and symmetry of the L80 attractor.

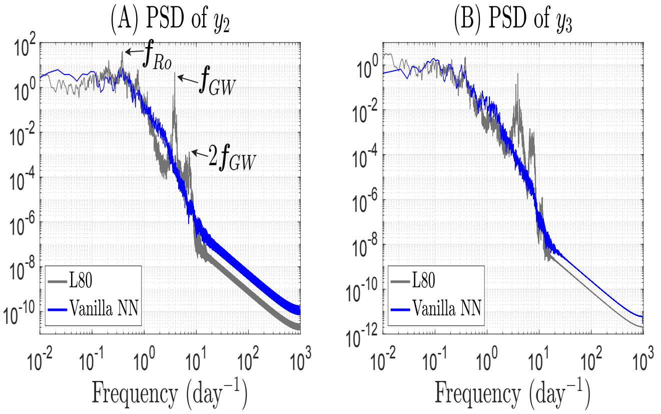

FIG. 8: **Power spectral density (PSD) comparison.** This figure compares the PSD of $y_2$ (Panel (A)) and $y_3$ (Panel (B)) as computed from the L80 model (gray) and from the vanilla NN-closure (blue) corresponding to Setting I in Fig. 6. Although the solution of this vanilla NN-closure shows a suitable capture of the spectral background of the L80 solutions, it fails in capturing the frequencies $f_{GW}$ and $f_{Ro}$ (along with their subharmonics) associated with inertia-gravity and Rossby waves, respectively.

TABLE II: **Loss function evaluations**. In this table are reported the loss values corresponding to the vanilla NN-closures shown in Fig. 6. Note that the underlying loss function is that defined in Eq. (7).

| Epochs | 10 | 50 | 100 | 300 | 500 |
|---|---|---|---|---|---|
| Setting I loss ($\times 10^{-2}$) | 2.62 | 2.54 | 2.52 | 2.49 | 2.49 |
| Setting II loss ($\times 10^{-2}$) | 2.74 | 2.67 | 2.66 | 2.64 | 2.64 |
| Setting III loss ($\times 10^{-2}$) | 2.72 | 2.45 | 2.44 | 2.43 | 2.43 |
| Setting IV loss ($\times 10^{-2}$) | 2.42 | 2.33 | 2.32 | 2.30 | 2.30 |

spite of the recent promises of neural parameterizations, still a lot of fundamental issues lie in front of us to derive reliable parameterizations accounting for rare events in a robust way. In that respect, rare events algorithms [49–51] could be of great use to simulate rare events offline in order to better sample e.g. the distribution's tails to train the neural networks.

We hope that this study sheds modestly new lights on the wealth of difficulties revealed by this other model of Lorenz, regarding its closure using data-driven methods. The L80 model of Lorenz has indeed attracted much less attention for closure than others of his models such as the Lorenz 96 model known however to be much less challenging for closure by machine learning [52]. In this perspective, the recent stochastic approach of [22] for closing the L80 model in the challenging regimes corresponding to $Ro > Ro^*$, could serve as a meaningful benchmark.

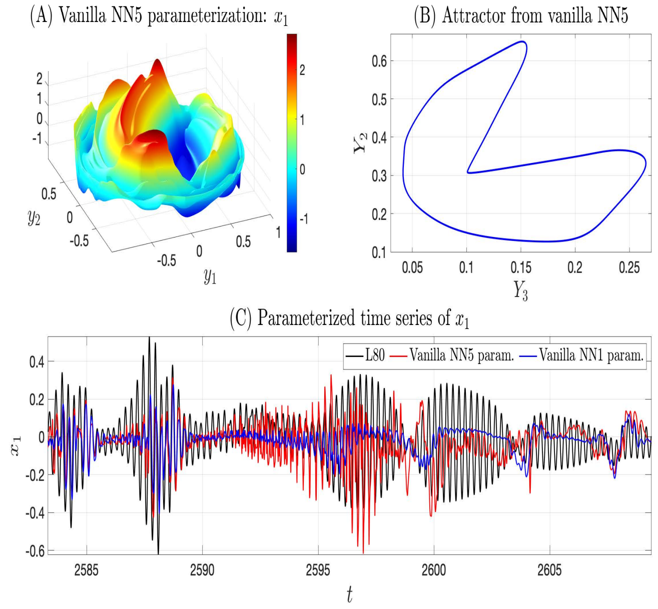

FIG. 9: **Panel A:** Neural parameterization $\boldsymbol{\mathcal{V}}_{\boldsymbol{\theta}}$ with 5 layers and 20 neurons (referred to as NN5), shown here for $x_1$, by adopting the same visualization method as for panels (B)-(D) of Fig. 1. Note the sharp gradients that reflect in this representation high-frequency attributes displayed by this parameterization. **Panel B:** Corresponding neural closure solution in the $(Y_2, Y_3)$-plane. **Panel C:** Compared to vanilla NN$_1$, i.e. the NN-parameterization (Setting I) showing the best vanilla closure skills (see Fig. 6), NN$_5$ exaggerates the high-frequency content as shown here offline.

## ACKNOWLEDGMENTS

This work has been partially supported by the Office of Naval Research (ONR) Multidisciplinary University Research Initiative (MURI) grant N00014-20-1-2023, by the National Science Foundation grant DMS-2108856, and by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 810370). We also acknowledge the computational resources provided by Advanced Research Computing at Virginia Tech.

## Appendix A: HLF solutions and the slow motion learning

The high-low frequency (HLF) solutions used in this article are those associated with [22, Fig. 7]. These solutions are obtained from the parameters used in Lorenz's original paper [7] except $F_1$ chosen to be $F_1 = 3.027 \times 10^{-1}$ as identified in [21]; see the Materials and Methods section in [22] for details.

As shown in Fig. 10, for this parameter regime, the HLF solutions contain a mixture of slow and fast oscillations in each variable $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ of the L80 model that causes serious difficulties for closure [22]. The dominant frequency of the Rossby wave content in the HLF solutions is $f_{Ro} = 0.31\,\text{day}^{-1}$ ($T_{Ro} = 3.2\,\text{days}$) and that of the inertia-gravity wave (IGW) content is $f_{GW} = 3.76\,\text{day}^{-1}$ ($T_{GW} = 6.3\,\text{hours}$).

To learn a neural parameterization of the slow motion, the weights and biases of the NNs are updated according to a Levenberg-Marquardt (LM) optimization [53]. The LM algorithm is known to be efficient for small or medium-scaled problems [54, Chap. 12], especially when the loss function is just a mean squared error, which is the case here. This algorithm is sufficient to obtain loss functions with small residuals; see Table I.

## Appendix B: The BE manifold and BE closure

For consistency, we recall from [21] the derivation of the BE manifold that serves as our parameterization baseline. Mathematically, the BE manifold aims at reducing the L80 model to a 3D system of ODE, by means of nonlinear parameterization of the variables $\boldsymbol{x} = (x_1, x_2, x_3)^{\mathrm{T}}$ and $\boldsymbol{z} = (z_1, z_2, z_3)^{\mathrm{T}}$, in terms of the variable $\boldsymbol{y} = (y_1, y_2, y_3)^{\mathrm{T}}$; see [32]. By analyzing the order of magnitudes of the different terms in the $x_i$-equations and after rescaling following [21], we arrive to the following parameterization of the $\boldsymbol{z}$-variable in terms of the rotational $\boldsymbol{y}$-variable

$$z_i = G_i(\boldsymbol{y}) = y_i - \frac{2c^2}{a_i} y_j y_k. \tag{B1}$$

Further algebraic manipulations show that under an invertibility condition of a matrix $M(\boldsymbol{y}, G(\boldsymbol{y}))$ conditioned on the $\boldsymbol{y}$-variable, one obtains (implicitly) $\boldsymbol{x}$ as a function $\Phi$ of $\boldsymbol{y}$ given by

$$\Phi(\boldsymbol{y}) = \big[M(\boldsymbol{y}, G(\boldsymbol{y}))\big]^{-1} \begin{pmatrix} d_{1,2,3}(\boldsymbol{y}, G(\boldsymbol{y})) \\ d_{2,3,1}(\boldsymbol{y}, G(\boldsymbol{y}))) \\ d_{3,1,2}(\boldsymbol{y}, G(\boldsymbol{y}))) \end{pmatrix}, \tag{B2}$$

where the $d_{i,j,k}$ are given explicitly; see [21, 32]. The function $\Phi(\boldsymbol{y}) = (\Phi_1(\boldsymbol{y}), \Phi_2(\boldsymbol{y}), \Phi_3(\boldsymbol{y}))^{\mathrm{T}}$ corresponds to the *BE manifold*, it is aimed to provide a nonlinear parameterization between $\boldsymbol{x}$ and $\boldsymbol{y}$ when the latter exists.

The BE closure is then

$$\begin{aligned} \frac{\mathrm{d}y_i}{\mathrm{d}\tau} &= -a_i^{-1} a_k b_k \Phi_j(\boldsymbol{y}) y_k - a_i^{-1} a_j b_j y_j \Phi_k(\boldsymbol{y}) \\ &\quad + c a_i^{-1} (a_k - a_j) y_j y_k - \Phi_i(\boldsymbol{y}) - \nu_0 a_i y_i, \end{aligned} \tag{B3}$$

for which $(i, j, k)$ lies in $\{(1,2,3), (2,3,1), (3,1,2)\}$.

[1] B. Bolin, Tellus **7**, 27 (1955).

[2] F. Baer and J. J. Tribbia, Monthly Weather Review **105**, 1536 (1977).

[3] B. Machenhauer, Beitr. Phys. Atmos **50** (1977).

[4] R. Daley, Reviews of Geophysics **19**, 450 (1981).

[5] C. E. Leith, J. Atmos. Sci. **37**, 958 (1980).

[6] E. N. Lorenz, J. Atmos. Sci. **20**, 130 (1963).

[7] E. N. Lorenz, J. Atmos. Sci. **37**, 1685 (1980).

[8] J. Sirignano and K. Spiliopoulos, Journal of Computational Physics **375**, 1339 (2018).

[9] M. Raissi, P. Perdikaris, and G. Karniadakis, Journal of Computational physics **378**, 686 (2019).

[10] Y. Bar-Sinai, S. Hoyer, J. Hickey, and M. Brenner, Proceedings of the National Academy of Sciences **116**, 15344 (2019).

[11] J. Pathak, B. Hunt, M. Girvan, Z. Lu, and E. Ott, Physical Review Letters **120**, 024102 (2018).

[12] S. L. Brunton, J. L. Proctor, and J. N. Kutz, Proceedings of the national academy of sciences **113**, 3932 (2016).

[13] S. Rasp, M. Pritchard, and P. Gentine, Proceedings of the National Academy of Sciences **115**, 9684 (2018).

[14] P. Gentine, M. Pritchard, S. Rasp, G. Reinaudi, and G. Yacalis, Geophysical Research Letters **45**, 5742 (2018).

[15] N. D. Brenowitz, T. Beucler, M. Pritchard, and C. Bretherton, Journal of the Atmospheric Sciences **77**, 4357 (2020).

[16] T. Bolton and L. Zanna, Journal of Advances in Modeling Earth Systems **11**, 376 (2019).

[17] R. Maulik, O. San, A. Rasheed, and P. Vedula, Journal of Fluid Mechanics **858**, 122 (2019).

[18] D. Kochkov, J. Smith, A. Alieva, Q. Wang, M. P. Brenner, and S. Hoyer, Proceedings of the National Academy of Sciences **118**, e2101784118 (2021).

[19] L. Zanna and T. Bolton, Geophysical Research Letters **47**, e2020GL088376 (2020).

[20] A. Subel, Y. Guan, A. Chattopadhyay, and P. Hassanzadeh, PNAS nexus **2**, pgad015 (2023).

[21] M. D. Chekroun, H. Liu, and J. C. McWilliams, Computers and Fluids **151**, 3 (2017).

[22] M. D. Chekroun, H. Liu, and J. C. McWilliams, Proc. Natl. Acad. Sci. USA **118**, e2113650118 (2021).

[23] R. Plougonven and C. Snyder, Journal of the atmospheric sciences **64**, 2502 (2007).

[24] I. Polichtchouk and R. Scott, Quarterly Journal of the Royal Meteorological Society **146**, 1516 (2020).

[25] S. Tulich, D. Randall, and B. Mapes, J. Atmos. Sci. **64**, 1210

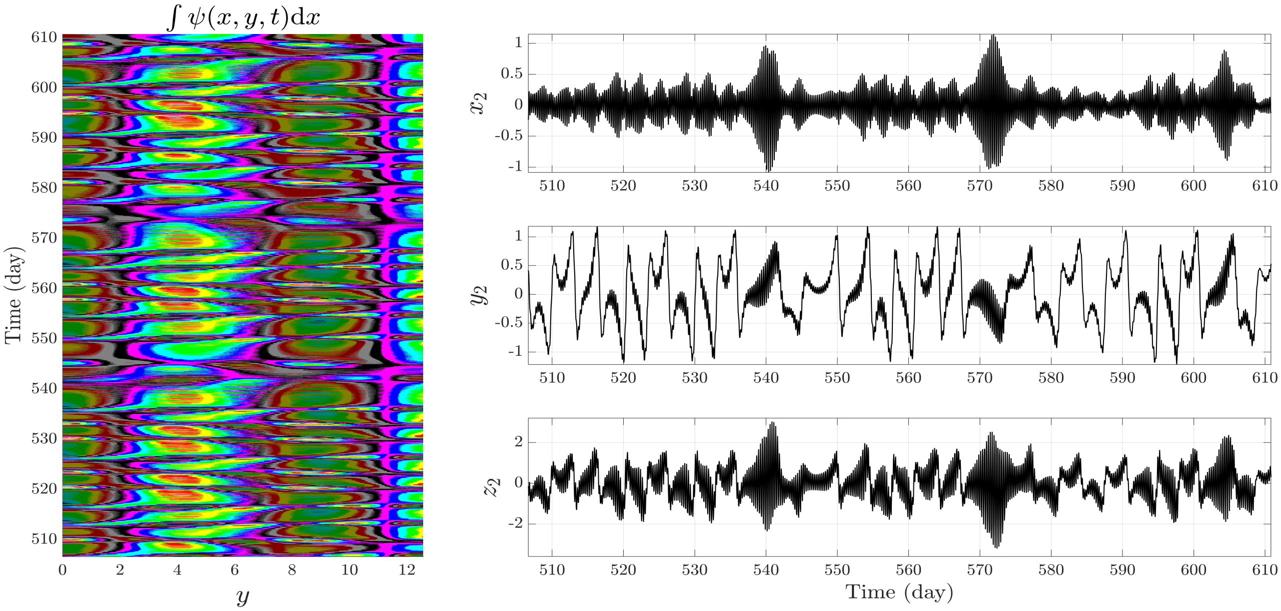

FIG. 10: **HLF solutions**. **Left panel**: Hovmöller plot of the streamfunction (averaged over the $x$-direction). **Right panels**: a few corresponding time series. Note the energetic bursts of fast oscillations corresponding to spontaneous bursts of IGWs. A vanilla NN consists of seeking for a neural network mapping e.g. $y_2$ to $x_2$ and $z_2$. These fast energetic bursts are a serious barrier to learning such a vanilla NN. The streamfunction $\psi$ is constructed from the $\boldsymbol{y}$-components of the L80 model solution according to $\psi(x,y,t) = \sum_{j=1}^{3} y_j(t)\cos(\alpha_j^1 x)\cos(\alpha_j^2 y)$ where the spatial variables $x$ and $y$ (not to be confused with $\boldsymbol{x}$ and $\boldsymbol{y}$ in the L80 model) takes value in a square domain $[0,L]\times[0,L]$ with $L = 4\pi$ and the vectors $\boldsymbol{\alpha}_j = (\alpha_j^1, \alpha_j^2)$, $j = 1,2,3$, are chosen to satisfy the conditions given by [7, Eqs. (16)–(17)]. For the considered parameter regime, we took $\boldsymbol{\alpha}_1 = (\sqrt{2}/2, \sqrt{2}/2)$, $\boldsymbol{\alpha}_2 = ((\sqrt{2}-\sqrt{6})/4, (\sqrt{2}+\sqrt{6})/4)$, and $\boldsymbol{\alpha}_3 = -(\boldsymbol{\alpha}_1 + \boldsymbol{\alpha}_2)$.

(2007).

[26] T. P. Lane, in *Encyclopedia of Atmospheric Sciences (2nd Edition)* (Elsevier, 2015) pp. 171–179.

[27] T. Dror, M. D. Chekroun, O. Altaratz, and I. Koren, Atmos. Chem. Phys. **21**, 12261 (2021).

[28] C. B. Rocha, T. K. Chereskin, S. T. Gille, and D. Menemenlis, Journal of Physical Oceanography **46**, 601 (2016).

[29] C. M. Bender and S. Orszag, *Advanced Mathematical Methods for Scientists and Engineers: Asymptotic Methods and Perturbation Theory*, Vol. 1 (Springer Science & Business Media, 1999).

[30] W. R. Young, Journal of Fluid Mechanics **920**, F1 (2021).

[31] J. McWilliams and P. Gent, J. Atmos. Sci. **37**, 1657 (1980).

[32] P. R. Gent and J. C. McWilliams, J. Atmos. Sci. **39**, 3 (1982).

[33] A. Monin, Akad. Nauk. Izv. Ser. Geofiz. **4**, 76 (1952).

[34] J. Charney, Tellus **7**, 22 (1955).

[35] E. Lorenz, Tellus **12**, 364 (1960).

[36] M. D. Chekroun, H. Liu, and J. C. McWilliams, J. Stat. Phys. **179**, 1073 (2020).

[37] E. N. Lorenz, J. Atmos. Sci. **43**, 1547 (1986).

[38] E. N. Lorenz and V. Krishnamurthy, J. Atmos. Sci. **44**, 2940 (1987).

[39] R. Camassa, Physica D **84**, 357 (1995).

[40] J. Vanneste, J. Atmos. Sci. **65**, 1622 (2008).

[41] R. Temam and D. Wirosoetisno, J. Atmos. Sci. **68**, 675 (2011).

[42] J. Vanneste, Ann. Rev. Fluid Mech. **45**, 147 (2013).

[43] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville, in *International Conference on Machine Learning* (PMLR, 2019) pp. 5301–5310.

[44] F. Ragone, J. Wouters, and F. Bouchet, Proc. Natl. Acad. Sci. USA **115**, 24 (2018).

[45] D. Swain, B. Langenbrunner, J. Neelin, and A. Hall, Nature Climate Change **8**, 427 (2018).

[46] V. M. Galfi and V. Lucarini, Physical Review Letters **127**, 058701 (2021).

[47] V. M. Gálfi, V. Lucarini, F. Ragone, and J. Wouters, La Rivista del Nuovo Cimento **44**, 291 (2021).

[48] V. Lucarini, V. M. Galfi, J. Riboldi, and G. Messori, Environmental Research Letters **18**, 015004 (2023).

[49] G. Dematteis, T. Grafke, M. Onorato, and E. Vanden-Eijnden, Physical Review X **9**, 041057 (2019).

[50] F. Bouchet, J. Rolland, and E. Simonnet, Physical Review Letters **122**, 074502 (2019).

[51] E. Simonnet, J. Rolland, and F. Bouchet, J. Atmos. Sci. **78**, 1889 (2021).

[52] https://raspstephan.github.io/blog/lorenz-96-is-too-easy/.

[53] M. Hagan and M. Menhaj, IEEE transactions on Neural Networks **5**, 989 (1994).

[54] B. Wilamowski and J. Irwin, *Intelligent systems* (CRC press, 2018).