



HAL
open science

Boosting performance in machine learning of geophysical flows via scale separation

Davide Faranda, M. Vrac, Pascal Yiou, F. M. E. Pons, A Hamid, G Carella,
C. G. Ngoungue Langué, S. Thao, V. Gautard

► To cite this version:

Davide Faranda, M. Vrac, Pascal Yiou, F. M. E. Pons, A Hamid, et al.. Boosting performance in machine learning of geophysical flows via scale separation. *Nonlinear Processes in Geophysics*, 2020, 10.5194/npg-2020-39 . hal-03341712v2

HAL Id: hal-03341712

<https://hal.science/hal-03341712v2>

Submitted on 9 Jun 2020 (v2), last revised 12 Sep 2021 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **Boosting performance in machine learning of geophysical flows**
2 **via scale separation**

3 D. Faranda*

4 *Laboratoire des Sciences du Climat et de l'Environnement,*
5 *CE Saclay l'Orme des Merisiers, UMR 8212 CEA-CNRS-UVSQ,*
6 *Université Paris-Saclay & IPSL, 91191 Gif-sur-Yvette, France.*

7 *London Mathematical Laboratory, 8 Margravine Gardens, London, W68RH, UK.*

8 M. Vrac, P. Yiou, F.M.E. Pons, A. Hamid, G. Carella, C.G. Ngoungue Langue, S. Thao

9 *Laboratoire des Sciences du Climat et de l'Environnement,*
10 *CE Saclay l'Orme des Merisiers, UMR 8212 CEA-CNRS-UVSQ,*
11 *Université Paris-Saclay & IPSL, 91191 Gif-sur-Yvette, France.*

12 V. Gautard

13 *DRF/IRFU/DEDIP//LILAS Departement d'Electronique*
14 *des Detecteurs et d'Informatique pour la Physique,*
15 *CE Saclay l'Orme des Merisiers, 91191 Gif-sur-Yvette, France.*

16 (Dated: June 8, 2020)

Abstract

Recent advances in statistical and machine learning have opened the possibility to forecast the behavior of chaotic systems using recurrent neural networks. In this article we investigate the applicability of such a framework to geophysical flows, known to involve multiple scales in length, time and energy and to feature intermittency. We show that both multiscale dynamics and intermittency introduce severe limitations on the applicability of recurrent neural networks, both for short-term forecasts, as well as for the reconstruction of the underlying attractor. We suggest that possible strategies to overcome such limitations should be based on separating the smooth large-scale dynamics from the intermittent/small-scale features. We test these ideas on global sea-level pressure data for the past 40 years, a proxy of the atmospheric circulation dynamics. Better short and long term forecasts of sea-level pressure data can be obtained with an optimal choice of spatial coarse grain and time filtering.

17 I. INTRODUCTION

18 The advent of high-performance computing has paved the way for advanced analyses of high-
19 dimensional datasets [1, 2]. Those successes have naturally raised the question of whether
20 it is possible to learn the behavior of a dynamical system without resolving or even without
21 knowing the underlying evolution equations. Such an interest is motivated on one side by the
22 fact that many complex systems still miss a universally accepted state equation — e.g. brain
23 dynamics [3], macro-economical and financial systems [4] — and, on the other, by the need of
24 reducing the complexity of the dynamical evolution for the systems of which the underlying
25 equations are known — e.g. on geophysical and turbulent flows [5]. Evolution equations are
26 difficult to solve for large systems such as the geophysical flows, so that approximations and
27 parameterizations are needed for meteorological and climatological applications [6]. These
28 difficulties are enhanced by those encountered in the modelling of phase transitions that
29 lead to cloud formation and convection, which are major sources of uncertainty in climate
30 modelling [7]. Machine Learning techniques capable of learning geophysical flows dynamics
31 would help improve those approximations and avoid running costly simulations resolving
32 explicitly all spatial/temporal scales.

* Correspondence to davide.faranda@lsce.ipsl.fr

33 Recently, several efforts have been made to apply machine learning to the prediction of geo-
34 physical data [8], to learn parameterizations of subgrid processes in climate models [9–11],
35 to the forecasting [12–14] and nowcasting (i.e. extremely short-term forecasting) of weather
36 variables [15–17], and to quantify the uncertainty of deterministic weather prediction [18].
37 One of the greatest challenge is to replace equations of climate models with neural network
38 capable to produce reliable long and short term forecast of meteorological variables. A first
39 great step in this direction was the use of Echo State Networks (ESN, [19]) a particular case
40 of Recurrent Neural Networks (RNN) to forecast the behavior of chaotic systems, such as
41 the Lorenz 1963 [20] and the Kuramoto-Sivashinsky [21] dynamics. It was shown that RNN
42 predictions of both systems attain performances comparable to those obtained with the real
43 equations [22, 23]. Good performance of regularized RNN in the short-term prediction of
44 multidimensional chaotic time series was obtained, both from simulated and real data [24].
45 This success motivated several follow-up studies with a focus on meteorological and climate
46 data. These are based on the idea of feeding various statistical learning algorithms with
47 data issued from dynamical systems of different complexity, in order to study short-term
48 predictability and long-term capabilities of RNN in producing a surrogate dynamics of the
49 input data. Recent examples include equation-informed moment-matching for the Lorenz96
50 model [25, 26], multi-layer perceptrons to reanalysis data [27], or convolutional neural net-
51 works to simplified climate simulation models [28, 29]. All these learning algorithms were
52 capable to provide some short-term predictability, but failed in obtaining a long-term be-
53 havior coherent with the input data.

54 In this article we specifically focus on how to improve the performance of ESN in simu-
55 lating long trajectories of large-scale climate fields. With respect to the results presented
56 in [23], we aim at going beyond the predictability horizon and investigate the ability of ma-
57 chine learning algorithms in shadowing the dynamics of observed data. Such applications
58 would avoid the use of general circulation models based on primitive equations to reproduce
59 the evolution of a subset of variables and therefore obtain surrogates dynamics of existing
60 datasets with little computational power. Previous results [27–29] suggest that RNN simu-
61 lations of large-scale climate fields are not as straightforward as those of the chaotic systems
62 considered by [23]. We identify two main mechanisms responsible for these limitations: (i)
63 the non-trivial interactions with small-scale motions carrying energy at large scale and (ii)
64 the intermittent nature of the dynamics. Intermittency triggers large fluctuations of observ-

65 ables of the motion in time and space [30] and can result in non-smooth trajectories within
66 the flow, leading to local unpredictability and increasing the number of degrees of freedom
67 needed to describe the dynamics [31].

68 By applying ESN to multiscale and intermittent systems, we investigate how scale separation
69 improves ESN predictions. Our goal is to reproduce a surrogate of the large-scale dynamics of
70 global sea-level pressure fields, a proxy of the atmospheric circulation. We begin by analysing
71 three different dynamical systems: we simulate the effects of small scales by artificially
72 introducing small-scale dynamics in the Lorenz 1963 equations [20] via additive noise. We
73 investigate the Pomeau-Manneville equations [32] stochastically perturbed with additive
74 noise to have an example of intermittent behavior. We then analyse the performance of
75 ESN in the Lorenz 1996 system [25]. This system dynamics is meant to mimic that of
76 the atmospheric circulation and feature both large-scale and small-scale variables with an
77 intermittent behavior. For all of those systems, as well as for the sea-level pressure data,
78 we show how the performance of ESN in predicting the behavior of the system deteriorates
79 rapidly when small-scale dynamics feedback to large scale is important. The idea of using
80 moving average for scale separation is already established for meteorological variables [33].
81 We choose the ESN framework following the results of [22, 23], and an established literature
82 about its ability to forecast chaotic time series and its stability to noise. For example, [34, 35]
83 analyse and compare the predictive performance of simple and improved ESN on simulated
84 and observed one-dimensional chaotic time series. We aim at understanding this sensitivity
85 in a deeper way, while assessing the possibility to reduce its impact on prediction through
86 simple noise reduction methods. The remaining of this article is organised as follows: first, we
87 give an overview of the ESN method and provide the description of the systems used. Then,
88 we show the results for the perturbed Lorenz 1963 equations, for the Pomeau-Manneville
89 intermittent map, and for the Lorenz 1996 equations. We discuss the improvement in short-
90 term prediction and the long-term attractor reconstruction obtained with the moving average
91 filter. We conclude by testing these ideas on atmospheric circulation data.

92 II. METHODS

93 Reservoir computing is a variant of recurrent neural networks (RNN) in which the input
94 signal is connected to a fixed and random dynamical system called reservoir [36]. The

95 principle of Reservoir computing consists in projecting first the input signal to a space of
 96 high dimension in order to obtain a non-linear representation of the signal; and then perform
 97 a new projection (linear regression or ridge regression) between the high-dimensional space
 98 and the output units. In our study ESN are implemented as follows. The code is given
 99 the in appendix and it shows the parameters used for the computations. Let $u(t)$ be the
 100 K -dimensional observable consisting of $t = 1, 2, \dots, T$ time iterations, originating from a
 101 dynamical system and $r(t)$ be the N -dimensional reservoir state, then:

$$r(t + dt) = \tanh(Wr(t) + W_{in}u(t)) \quad (1)$$

102 where W is the adjacency matrix of the reservoir: its dimensions are $N \times N$, and N is the
 103 number of neurons of the reservoir. In ESN, the neuron layers of classic deep neural networks
 104 are replaced by a single layer consisting of a sparsely connected random network, with
 105 coefficients uniformly distributed in $[-0.5; 0.5]$. W_{in} , with dimensions $N \times K$, is the weight
 106 matrix of the connections between the input layer and the reservoir and the coefficients are
 107 randomly sampled, as for W . The output of the network at time step $t + dt$ is

$$W_{out}r(t + dt) = v(t + dt) \quad (2)$$

108 where $v(t + dt)$ is the ESN prediction, W_{out} with dimensions $K \times N$, is the weight matrix of
 109 the connections between the reservoir neurons and the output layer. We estimate W_{out} via
 110 a ridge regression [37]:

$$W_{out} = v(t + dt)r(t + dt)^T[r(t + dt)r(t + dt)^T - \lambda I]^{-1} \quad (3)$$

111 with $\lambda = 10^{-8}$. In the prediction phase we have a recurrent relationship:

$$r(t + dt) = \tanh(Wr(t) + W_{in}W_{out}r(t)). \quad (4)$$

112 **A. ESN performance indicators**

113 In this paper, we use three different indicators of performance of the ESN:

114

Statistical distributional test

As a first diagnostic of the performance of ESN, we aim at assessing whether the marginal
 distribution of the forecast values for a given dynamical system is significantly different

from the invariant distribution of the system itself. To this purpose, we conduct a χ^2 test [38], designed as follows. Let U be a system observable with support R_U and probability density function $f_U(u)$, and let $u(t)$ be a sample trajectory from U . Let now $\hat{f}_U(u)$ be an approximation of $f_U(u)$, namely the histogram of u over $i = 1, \dots, M$ bins. Note that, if u spans the entire phase space, $\hat{f}_U(u)$ is the numerical approximation of the Sinai-Ruelle-Bowen measure of the dynamical system [39, 40]. Let now V be the variable generated by the ESN forecasting, with support $R_V = R_U$, $v(t)$ the forecast sample, $g_V(v)$ its probability density function and $\hat{g}_V(v)$ the histogram of the forecast sample. We test the null hypothesis that the marginal distribution of the forecast sample is the same as the invariant distribution of the system, against the alternative hypothesis that the two distributions are significantly different:

$$H_0 : f_U(u) = g_V(v) \quad \text{for every } u \in R_U$$

$$H_1 : f_U(u) \neq g_V(v) \quad \text{for any } u \in R_U$$

115 Under H_0 , $\hat{f}_U(u)$ is the expected value for $\hat{g}_V(v)$, which implies that observed differences
 116 $(\hat{g}_V(v) - \hat{f}_U(u))$ are due to random errors, and are then independent and identically dis-
 117 tributed Gaussian random variables. Statistical theory shows that, given H_0 true, the test
 118 statistics

$$\Sigma = \sum_{i=1}^M \frac{(\hat{g}_V(v) - \hat{f}_U(u))^2}{\hat{f}_U(u)} \quad (5)$$

119 is distributed as a chi-squared random variable with M degrees of freedom, $\chi^2(M)$. Then, to
 120 test the null hypothesis at the level α , the observed value of the test statistics Σ is compared
 121 to the critical value corresponding to the $1 - \alpha$ quantile of the chi-square distribution,
 122 $\Sigma_c = \chi_{1-\alpha}^2(M)$: if $\Sigma > \Sigma_c$, the null hypothesis must be rejected in favour of the specified
 123 alternative.

124 In our setup, we encounter two limitations in using the standard χ^2 test. First, problems
 125 may arise when $\hat{f}_U(u)$, i.e. if the sample distribution does not span the entire support of the
 126 invariant distribution of the system. We observe this in a relatively small number of cases;
 127 since aggregating the bins would introduce unwanted complications, we decide to discard
 128 the pathological cases, controlling the effect empirically as described below. Moreover,
 129 even producing relatively large samples, we are not able to actually observe the invariant
 130 distribution of the considered system, which would require much longer simulations. As

131 a consequence, we would observe excessive rejection rates when testing samples generated
 132 under H_0 .

133 We decide to control these two effects by using a Monte Carlo approach. To this purpose, we
 134 use 10000 samples using the system equations under the null hypothesis, and we compute
 135 the test statistic for each one according to Eq. (5). Then, we use the $(1 - \alpha)$ quantile of the
 136 empirical distribution of Σ — instead of the theoretical $\chi^2(M)$ — to determine the critical
 137 threshold Σ_c . As a last remark, we notice that we are making inference in repeated tests
 138 setting, as the performance of the ESN is tested 10000 times. Performing a high number of
 139 independent tests at a chosen level α increases the observed rejection rate: in fact, even if the
 140 samples are drawn under H_0 , extreme events become more likely, resulting in an increased
 141 probability to erroneously reject the null hypothesis. To avoid this problem, we apply the
 142 Bonferroni correction [41], testing each one of the $m = 10000$ available samples at the level
 143 $\alpha' = \frac{\alpha}{m}$, with $\alpha = 0.05$.

144 Averaging the test results over several sample pairs $u(t)$, $v(t)$ we obtain a rejection rate
 145 $0 < \phi < 1$ that we use to measure the adherence of a ESN trajectory $v(t)$ to trajectories
 146 obtained via the equations. If $\phi = 0$, almost all the ESN trajectories can shadow original
 147 trajectories, if $\phi = 1$ none of the ESN trajectories resemble those of the systems of equations.

148 ***Predictability Horizon***

149 As a measure of the predictability horizon of the ESN forecast compared to the equations,
 150 we use the root mean square error (RMSE):

$$RMSE(\tau) = \sqrt{\frac{1}{\tau} \sum_{t=1}^{\tau} (u(t) - v(t))^2} \quad (6)$$

and we define the predictability horizon τ_s as the first time that RMSE exceeds a certain
 threshold s . We link s to the average separation of observations in the observable U and we
 fix

$$s = \frac{1}{T-1} \sum_{t=2}^{T-1} [u(t) - u(t-1)].$$

151 We have tested the sensitivity of results against the exact definition of s .

152 We interpret τ_s as a natural measure of the Lyapunov time ϑ , namely the time it takes for
 153 an ensemble of trajectories of a dynamical system to diverge [42, 43].

154 ***Initial Forecast Error***

155 The initial error is given by $\eta = RMSE(t = 1)$, for the first time step after the initial

156 condition at $t = 0$. We expect η to reduce as the training time increases. In this phase, the
157 the smaller the initial error will be.

158

159 **B. Moving average filter**

160 Equipped with these indicators, we analyze two sets of simulations performed with and
161 without smoothing, which was implemented using a moving average filter. The moving
162 average operation is the integral of $u(t)$ between t and $t - w$, where w is the window size of
163 the moving average. The simple moving average filter can be seen as a nonparametric time
164 series smoother (see e.g. [44], chapter 1.5). It can be applied to smooth out (relatively) high
165 frequencies in a time series, both to de-noise the observations of a process or to estimate
166 trend-cycle components, if present. Moving averaging consists, in practice, of replacing the
167 observation $u(t)$ by a value $u^{(f)}(t)$, obtained by averaging the previous w observations. If
168 the time dimension is discrete (like in the Pomeau-Manneville system) it is defined as:

$$u^{(f)}(t) = \frac{1}{w} \sum_{i=0}^{w-1} u(t - i), \quad (7)$$

169 while for continuous time systems (like the Lorenz 1963 system), the sum is formally replaced
170 by an integral:

$$u^{(f)}(t) = \frac{1}{w} \int_{t-w}^t u(\zeta) d\zeta. \quad (8)$$

171 We can define the residuals as:

$$\delta u(t) = u^{(f)}(t) - u(t) \quad (9)$$

172 In practice, the computation always refers to the discrete time case, as continuous time
173 systems are also sampled at finite time steps. Since Echo State Networks are known to
174 be sensitive to noise (see e.g. [34]), we exploit the simple moving average filter to smooth
175 out high-frequency noise and assess the results for different smoothing windows w . We
176 find that the choice of the moving averaging window w must respect two conditions: it
177 should be large enough to smooth the noise but smaller than the characteristic time τ of
178 the large-scale fluctuations of the system. For chaotic systems, τ can be derived knowing
179 the rate of exponential divergence of the trajectories, a quantity linked to the Lyapunov

180 exponents [45], and τ is known as the Lyapunov time.

181

182 We also remark that we can express explicitly the original variable $u(t)$ as a function of the
183 filtered variable $u^{(f)}(t)$ as:

$$u(t) = w(u^{(f)}(t) - u^{(f)}(t - 1)) + u(t - w) \quad (10)$$

184 we will test this formula for stochastically perturbed systems to evaluate the error introduced
185 by the use of residuals δu .

186 C. Testing ESN on filtered dynamics

187 Here we describe the algorithm used to test ESN performance on filtered dynamics:

- 188 1. Simulate the reference trajectory $u(t)$ using the equations of the dynamical systems,
189 where $u(t)$ has been standardized by subtracting the mean and dividing by its standard
190 deviation.
- 191 2. Perform the moving average filter to obtain $u^{(f)}(t)$.
- 192 3. Extract from $u^{(f)}(t)$ a training set $u_{train}^{(f)}(t)$ with $t \in \{1, 2, \dots, T_{train}\}$.
- 193 4. Train the ESN on $u_{train}^{(f)}(t)$ dataset.
- 194 5. Obtain the ESN forecast $v^{(f)}(t)$ for $t \in \{T_{train} + 1, T_{train} + 2, \dots, T\}$.
- 195 6. Add residuals (Eq. 9) to $v^{(f)}(t)$ sample as $v(t) = v^{(f)}(t) + \delta u$, where δu is randomly
196 sampled from the $\delta u(t)$ with $t \in \{1, 2, \dots, T_{train}\}$.
- 197 7. Compare $v(t)$ and $u(t > T_{train})$ using the metrics ϕ , τ and η .

198 As an alternative to step 6, one can also use Eq. (10) and obtain:

$$v(t) = w(v^{(f)}(t) - v^{(f)}(t - 1)) + v(t - w), \quad (11)$$

199 that does not require the use of residuals $\delta u(t)$.

200 **III. RESULTS**

201 The systems we analyze are the Lorenz 1963 attractor [20] with the classical parameters,
 202 discretized with a Euler scheme and a $dt = 0.001$, the Pomeau-Manneville intermittent
 203 map [32], the Lorenz 1996 equations [25] and the NCEP sea-level pressure data [46].

204

205 ***Lorenz 1963 equations***

206 The Lorenz [20] system is a simplified model of Rayleigh-Benard convection, derived
 207 by E.N. Lorenz. It is an autonomous continuous dynamical system with three variables
 208 $u \in \{x, y, z\}$ parametrizing respectively the convective motion, the horizontal temperature
 209 gradient and the vertical temperature gradient. It writes:

$$\begin{aligned} \frac{dx}{dt} &= \sigma(y - x) + \epsilon\xi_x(t) \\ \frac{dy}{dt} &= -xz + \rho x - y + \epsilon\xi_y(t), \\ \frac{dz}{dt} &= xy - bz + \epsilon\xi_z(t), \end{aligned} \tag{12}$$

210 where σ , ρ and b are three parameters, σ mimicking the Prandtl number and ρ the reduced
 211 Rayleigh number. The Lorenz model is usually defined using Eq. (12), with $\sigma = 10$, $\rho = 28$
 212 and $b = 8/3$. A deterministic trajectory of the system is shown in Figure 1a). It has been ob-
 213 tained via integrating numerically the Lorenz equations with an Euler scheme ($dt = 0.001$).
 214 The systems is perturbed via additive noise: $\xi_x(t)$, $\xi_y(t)$ and $\xi_z(t)$ are random variable all
 215 drawn from a Gaussian distribution. The initial conditions are randomly selected within
 216 a long trajectory of $5 \cdot 10^6$ iterations. First, we study the dependence of the ESN on the
 217 training length in the deterministic system ($\epsilon = 0$, Figure 1b-d). We analyse the behavior
 218 of the rejection rate ϕ (panel b), the predictability horizon τ_s (panel c) and the initial error
 219 η (panel d) as a function of the training sample size. Our analysis suggests that $t \sim 10^2$ is
 220 a minimum sufficient choice for the training window. We compare this time to the typical
 221 time scales of the motion of the sytems, determined via the maximum Lyapunov exponent
 222 λ . For the Lorenz 1963 system, $\lambda = 0.9$, so that the Lyapunov time $\vartheta \approx \mathcal{O}(\frac{1}{\lambda}) \approx 1.1$.
 223 From the previous analysis we should train the network at least for $t > 100\vartheta$. For the other
 224 systems analysed in this article, we take this condition as a lower boundary for the training
 225 times.

226

227 To show the effectiveness of the moving average filter in boosting the machine-learning
 228 performances we produce 10 ESN trajectories obtained without moving average (Figure 2-
 229 green) and with (Figure 2-red) a moving average window $w = 0.01$ and compare them to
 230 the reference trajectory (blue) obtained with $\epsilon = 0.1$. The value of $w = 10dt = 0.01$ respects
 231 the condition $w \ll \vartheta$. Indeed, the RMSE averaged over the two groups of trajectories
 232 (Figure 2-b) shows an evident gain of accuracy (a factor of ~ 10) when the moving average
 233 procedure is applied. We now study in a more systematic way the dependence of the ESN
 234 performance on noise intensity ϵ , network size N and for three different averaging windows
 235 $w = 0$, $w = 0.01$, $w = 0.05$. We produce, for each combination, 100 ESN forecasts. Figure 3
 236 shows ϕ (a), $\log(\tau_{s=1})$ (b) and $\log(\eta)$ (c) computed setting $u \equiv x$ variable of the Lorenz
 237 1963 system (results qualitatively do not depend on the chosen variable). In each panel
 238 from left to right the moving average window is increasing, upper sub-Panels are obtained
 239 using the exact expression in Eq. 11 and lower panels using the residuals in Eq 9. For
 240 increasing noise intensity and for small reservoirs sizes, the performances without moving
 241 average (left subpanels) rapidly get worse. The moving average smoothing with $w = 0.01$
 242 (central sub-panels) improves the performance for $\log(\tau_{s=1})$ (b) and $\log(\eta)$ (c), except when
 243 the noise is too large ($\epsilon = 1$). When the moving average window is too large (right panels),
 244 the performances of ϕ decrease. This failure can be attributed to the fact that residuals δu
 245 (Eq.9) are of the same order of magnitude of the ESN predicted fields for ϵ large. Indeed,
 246 if we use the formula provided in Eq. 11 as an alternative to step 6, we can evaluate the
 247 error introduced in the residuals. The results shown in Figure 3 suggest that residuals can
 248 be used without problems when the noise is small compared with the dynamics. When ϵ is
 249 close to one, the residuals overlay the deterministic dynamics and ESN forecast are poor.
 250 In this case, the exact formulation in Eq. 11 appears much better.

251 *Pomeau-Manneville intermittent map*

252 Several dynamical systems, including Earth climate, display intermittency, i.e., the time
 253 series of a variable issued by the system can experience sudden chaotic fluctuations, as well
 254 as a predictable behavior where the observables have small fluctuations. In atmospheric
 255 dynamics, such a behavior is observed in the switching between zonal and meridional phases
 256 of the mid-latitude dynamics if a time series of the wind speed at one location is observed:
 257 when a cyclonic structure passes through the area, the wind has high values and large
 258 fluctuations, when an anticyclonic structure is present the wind is low and fluctuations are

259 smaller [47, 48]. It is then of practical interest to study the performance of ESN in Pomeau
 260 Manneville predictions as they are a first prototypical example of the intermittent behavior
 261 found in climate data.

262 In particular, the Pomeau-Manneville [32] map is probably the simplest example of inter-
 263 mittent behavior, produced by a 1D (here $u = x$) discrete deterministic map given by:

$$x_{t+1} = \text{mod}(x_t + x_t^{1+a}, 1) + \epsilon\xi(t), \quad (13)$$

264 where $0 < a < 1$ is a parameter. We use $a = 0.91$ in this study and a trajectory consisting of
 265 5×10^5 iterations. The systems is perturbed via additive noise $\xi(t)$ drawn from a Gaussian
 266 distribution. It is well known that Pomeau-Manneville systems exhibit sub-exponential
 267 separation of nearby trajectories and then the Lyapunov exponent is $\lambda = 0$. However, one
 268 can define a Lyapunov exponent for the non-ergodic phase of the dynamics and extract a
 269 characteristic time scale [49]. From this latter reference, we can derive a value $\lambda \simeq 0.2$ for
 270 $a = 0.91$, implying $w < \tau \simeq 5$. We find that the best match between ESN and equations in
 271 terms of the ϕ indicator are obtained for $w = 3$.

272

273 Results for the Pomeau-Manneville map are shown in Figure 4. We first observe that the
 274 ESN forecast of the intermittent dynamics of the Pomeau-Manneville map is much more
 275 challenging than for the Lorenz system as a consequence of the intermittent behavior of this
 276 system. For the simulations performed with $w = 0$, the ESN cannot simulate an intermittent
 277 behavior, for all noise intensities and reservoir sizes. This is reflected in the behavior of the
 278 indicators. In the deterministic limit, the ESN fails to reproduce the invariant density in
 279 80% of the cases ($\phi \simeq 0.8$). For intermediate noise intensities $\phi > 0.9$ (Figure 4-a). The pre-
 280 dictability horizon $\log(\tau_{s=0.5})$ for the short term forecast is small (Figure 4d) and the initial
 281 error large (Figure 4g). The moving average procedure with $w = 3$ partially improves the
 282 performances (Figure 4b,c,e,f,h,i) and it enables ESN to simulate an intermittent behavior
 283 (Figure 5). Performances are again better when using the exact formula (Figure 4b,e,h) than
 284 using the residuals δu (Figure 4c,f,i). Figure 5a) shows the intermittent behavior of the data
 285 generated with the ESN trained on moving averaged data of Pomeau-Manneville system
 286 (red) and compare to the target time series (blue). ESN simulations do not reproduce the
 287 intermittency in the average of the target signal. They only show some second order inter-
 288 mittency in the fluctuations. Figure 5b) displays the power spectra showing in both cases

289 a power law decay, which are typical of turbulent phenomena. Although the intermittent
 290 behavior is captured, this realization of ESN shows that the values are concentrated around
 291 $x = 0.5$ for the ESN prediction, whereas the non-intermittent phase peaks around $x = 0$ for
 292 the target data.

293

294 ***The Lorenz 1996 system***

295 Before running the ESN algorithm on actual climate data, we test our idea in a more
 296 sophisticated, and yet still idealized, model of atmospheric dynamics, namely the Lorenz
 297 1996 equations [25]. This model explicitly separates two scales and therefore will provide a
 298 good test for our ESN algorithm. The Lorenz 1996 system consists of a lattice of large-scale
 299 resolved variables X , coupled to small-scale variables Y , whose dynamics can be intermittent,
 300 so that $u \in \{X, Y\}$. The model is defined via two equations:

$$\begin{aligned} \frac{dX_i}{dt} &= X_{i-1}(X_{i+1} - X_{i-2}) - X_i + F - \frac{hc}{b} \sum_{j=1}^J Y_{j,i}, \\ \frac{dY_{j,i}}{dt} &= cbY_{j+1,i}(Y_{j-1,i} - Y_{j+2,i}) - cY_{j,i} + \frac{hc}{b} X_i \end{aligned} \tag{14}$$

301 where $i = 1, \dots, I$ and $j = 1, 2, \dots, J$ denote respectively the number of large-scale X
 302 and small-scale Y variables. Large-scale variables are meant to represent the meanders
 303 of the jet-stream driving the weather at mid-latitudes. The first term on the right-hand
 304 side represents advection, the second diffusion, while F mimics an external forcing. The
 305 system is controlled via the parameters b and c (the time scale of the the fast variables
 306 compared to the small variables) and via h (the coupling between large and small scales).
 307 From now on, we fix $I = 30$, $J = 5$ and $F = b = 10$ as these parameters are typically used to
 308 explore the behavior of the system [50]. We integrate the equations with an Euler scheme
 309 ($dt = 10^{-3}$) from the initial conditions $Y_{j,i} = X_i = F$, where only one mode is perturbed as
 310 $X_{i=1} = F + \varepsilon$ and $Y_{j,i=1} = F + \varepsilon^2$. Here $\varepsilon = 10^{-3}$. We discard about $2 \cdot 10^3$ iterations to
 311 reach a stationary state on the attractor, and we retain $5 \cdot 10^4$ iterations. When c and h vary,
 312 different interactions between large and small scales can be achieved. A few examples of
 313 simulations of the first mode X_1 and Y_1 are given in Figure 6. Figure 6a,c show simulations
 314 obtained for $h = 1$ by varying c : the larger c the more intermittent the behavior of the fast
 315 scales. Figure 6.b,d) show simulations obtained for different coupling h at fixed $c = 10$:

316 when $h = 0$, there is no small-scale dynamics.

317

318 In the Lorenz 1996 model we can explore what happens to the ESN performances if we turn
319 on and off intermittency and/or the small-to-large-scale coupling, without introducing any
320 additional noise term. Moreover, we can also learn the Lorenz 1996 dynamics on the X vari-
321 ables only, or learn the dynamics on both X and Y variables. The purpose of this analysis is
322 to assess whether the ESN are capable of learning the dynamics of the large-scale variables
323 X alone, and how this capability is influenced by the coupling and the intermittency of the
324 small-scale variables Y . Using the same simulations presented in Figure 6, we train the
325 ESN on the first $2.5 \cdot 10^4$ iterations, and then perform, changing the initial conditions 100
326 different ESN predictions for $2.5 \cdot 10^4$ more iterations. We apply our performance indicators
327 not to the entire I -dimensional X variable (X_1, \dots, X_I) , as the χ^2 test becomes intractable
328 in high dimensions, but rather to the spatial average of the large-scale variables X . The
329 behavior of each variable X_i is similar, so the average is representative of the collective
330 behavior. The rate of failure ϕ is very high (not shown) because even when the dynamics is
331 well captured by the ESN the variables are not scaled and centered as those of the original
332 systems. For the following analysis, we therefore replace ϕ with the χ^2 distance T (Eq. (5)).
333 The use of T allows for better highlighting the differences in the ESN performance with
334 respect to the chosen parameters. The same considerations also apply to the analysis of the
335 sea-level pressure data reported in the next paragraph.

336

337 Results of the ESN simulations for the Lorenz 1996 system are reported in Figure 7. In Fig-
338 ure 7a,c,e) ESN predictions are obtained by varying c at fixed $h = 1$, while in Figure 7b,d,f)
339 by varying h at fixed $c = 10$. The continuous lines refer to results obtained feeding the
340 ESN with only the X variables, dotted lines with both X and Y . For the χ^2 distance T
341 (Figure 7a,b), performances show a large dependence on both intermittency c and coupling
342 h . First of all, we remark that learning both X and Y variables lead to higher distances T ,
343 except for the non intermittent case, $c = 1$. For $c > 1$, the dynamics learnt on both X and
344 Y never settles on a stationary state resembling that of the Lorenz 1996 model. When $c > 1$
345 and only the dynamics of the X variables is learnt, the dependence on N when h is varied is
346 non monotonic and better performances are achieved for $800 < N < 1200$. For this range,
347 the dynamics settles on stationary states whose spatio-temporal evolution resembles that of

348 the Lorenz 1996 model, although the variability of time and spatial scales is different from
349 the target. An example is provided in Figure 8, for $N = 800$.

350

351 Let us now analyse the two indicators of short-term forecasts. Figure 7c,d) display the
352 predictability horizon τ_s with $s = 1$. The best performances are achieved for the non-
353 intermittent case $c = 1$ and learning both X and Y . When only X is learnt, we again
354 get better performances in terms of τ_s for rather small network sizes. The performances
355 for $c > 1$ are better when only X variables are learnt. The good performance of ESN in
356 learning only the large-scale variables X are even more surprising when looking at initial
357 error η (Figure 7), which is one order of magnitude smaller when X, Y are learnt. Despite
358 this advantage in the initial conditions, the ESN performances on (X, Y) are better only
359 when the dynamics of Y is non-intermittent. We find clear indications that large intermit-
360 tency ($c = 25$) and strong small-to-large scale variables coupling ($h = 1$) worsen the ESN
361 performances, supporting the claims made for the Lorenz 1963 and the Pomeau-Manneville
362 systems.

363

364 *The NCEP sea-level pressure data*

365 We now test the effectiveness of the moving average procedure in learning the behavior of
366 multiscale and intermittent systems on climate data issued by reanalysis projects. We use
367 data from the National Centers for Environmental Prediction (NCEP) version 2 [46] with a
368 horizontal resolution of 2.5° . We adopt the global 6 hourly sea-level pressure (SLP) field from
369 1979 to 31/08/2019 as the meteorological variable proxy for the atmospheric circulation.
370 It traces cyclones (resp. anticyclones) with minima (resp. maxima) of the SLP fields.
371 The major modes of variability affecting mid-latitudes weather are often defined in terms
372 of the Empirical Orthogonal Functions (EOF) of SLP and a wealth of other atmospheric
373 features [51, 52], ranging from teleconnection patterns to storm track activity to atmospheric
374 blocking can be diagnosed from the SLP field.

375 In addition to the time moving average filter, we also investigate the effect of spatial coarse-
376 graining the SLP fields by a factor c and perform the learning on the reduced fields. We use
377 the nearest neighbor approximation, which consist in taking from the original dataset the
378 closest value to the coarse grid. Compared with methods based on averaging or dimension
379 reduction techniques such as EOFs, the nearest neighbors approach has the advantage of

380 not removing the extremes (except if the extreme is not in one of the closest gridpoint) and
 381 preserve cyclonic and anticyclonic structures. For $c = 2$ we obtain a horizontal resolution
 382 of 5° and for $c = 4$ a resolution 10° . For $c = 4$ the information on the SLP field close to
 383 the poles is lost. However, in the remaining of the geographical domain, the coarse grained
 384 fields still capture the positions of cyclonic and anticyclonic structures. Indeed, as shown
 385 in [53], this coarse grain field still preserves the dynamical properties of the original one.
 386 There is therefore a certain amount of redundant information on the original 2.5° horizontal
 387 resolution SLP fields.

388 The dependence of the quality of the prediction for the sea-level pressure NCEPv2 data on
 389 the coarse graining factor c and on the moving average window size w is shown in Figure 9.
 390 We show the results obtained using the residuals (Eq. 9). Figure 9a-c) show the distance
 391 from the invariant density, using the χ^2 distance T . Here it is clear that by increasing w ,
 392 we get better forecast with smaller network sizes N . A large difference for the predictability
 393 expressed as predictability horizon τ_s , $s = 1.5$ hPa (Figure 9d-f) emerges when SLP fields
 394 are coarse grained. We gain up to 10h in the predictability horizon with respect to the
 395 forecasts performed on the original fields ($c = 0$). This gain is also reflected by the initial
 396 error η (Figure 9g-i). From the combination of all the indicators, after a visual inspection,
 397 we can identify the best-set of parameters: $w = 12$ h, $N = 200$ and $c = 4$. Indeed this is
 398 the case such that, with the smallest network we get almost the minimal χ^2 distance T , the
 399 highest predictability (32 h) and one of the lowest initial errors. We also remark that, for
 400 $c = 0$ (panels (c) and (i)), the fit always diverges for small network sizes.

401 We compare in details the results obtained for two 10-year predictions with $w = 0$ h and
 402 $w = 12$ h at $N = 200$ and $c = 4$ fixed. At the beginning of the forecast time (Supplementary
 403 Video 1), the target field (panel a) is close to both that obtained with $w = 0$ h (panel b)
 404 and $w = 12$ h (panel c). When looking at a very late time (Supplementary Video 2), of
 405 course we do not expect to see agreement among the three datasets. Indeed we are well
 406 beyond the predictability horizon. However, we remark that the dynamics for the run with
 407 $w = 0$ h is steady: positions of cyclones and anticyclones barely evolve with time. Instead,
 408 the run with $w = 12$ h shows a richer dynamical evolution with generation and annihilation
 409 of cyclones. A similar effect can be observed in the ESN prediction of the Lorenz 96 system
 410 shown in Figure 8b) where the quasi-horizontal patterns indicate less spatial mobility than
 411 the original system (Figure 8a).

412 In order to assess the performances of the two ESNs with and without moving average
 413 in a more quantitative way, we present the space-time distributions in Figure 10a). The
 414 distribution obtained for the moving average $w = 12\text{h}$ has more realistic tails and matches
 415 better than the run $w = 0\text{h}$ that of the target data. Figure 10b-d) shows the wavelet
 416 spectrograms (or scalograms) [54]. The scalogram is the absolute value of the continuous
 417 wavelet transform of a signal, plotted as a function of time and frequency. The target
 418 data spectrogram (b) presents a rich structure at different frequencies and some interannual
 419 variability. The wavelet spectrogram of non-filtered ESN run $w = 0\text{ h}$ (c) shows no short
 420 time variability and too large interseasonal and interannual variability. The spectrogram of
 421 the target data is better matched by the run with $w = 12\text{ h}$ (d) which shows that, on time
 422 scales of days to weeks, there is a larger variability.

423 **IV. DISCUSSION**

424 We have analysed the performance of ESN in reproducing both the short and long-term
 425 dynamics of observables of geophysical flows. The motivation for this study came from the
 426 evidence that a straightforward application of ESN to high dimensional geophysical data
 427 (such as the 6 hourly global gridded sea-level pressure data) does not yield to the same
 428 results quality obtained by [23] for the Lorenz 1963 and the Kuramoto-Sivashinsky models.
 429 Here we have investigated the causes for this behavior and identified two main bottlenecks:
 430 (i) intermittency and (ii) the presence of multiple dynamical scales, which both appear in
 431 geophysical data. In order to illustrate this effect, we have first analysed two low dimensional
 432 systems, namely the Lorenz 1963 [20] and the Pomeau-Manneville [32] equation. To mimic
 433 multiple dynamical scales, we have added noise terms to the dynamics. The performance of
 434 ESN in predicting rapidly drops when the systems are perturbed with noise. Filtering the
 435 noise allows to partially recover predictability. It also enables to simulate some qualitative
 436 intermittent behavior in the Pomeau-Manneville dynamics. This feature could be explored
 437 by changing the degree of intermittency in the Pomeau-Manneville map as well as perform-
 438 ing parameter tuning in ESN. This is left for future work. Here we have used a simple
 439 moving-average filter and shown that a careful choice of the moving-average window can
 440 enhance predictability. As an intermediate step between the low-dimensional models and
 441 the application to the sea-level pressure data, we have analysed the ESN performances on

442 the Lorenz 1996 system [25]. This system was introduced to mimic the behavior of the at-
443 mospheric jet at mid-latitude, and features a lattice of large-scale variables, each connected
444 to small-scale variables. Both the coupling between large and small scales and intermittency
445 can be tuned in the model, giving rise to a plethora of behaviors. For the Lorenz 1996 model,
446 we did not have to apply a moving average filter to the data, as we can train the ESN on the
447 large-scale variables only. Our computations have shown that, whenever the small scales are
448 intermittent, or the coupling is strong, learning the dynamics of the coarse grained variable
449 is more effective, both in terms of computation time and performances. The results also
450 apply to geophysical datasets: here we analysed the atmospheric circulation, represented
451 by sea-level pressure fields. Again we have shown that both a spatial coarse-graining and a
452 time moving-average filter improve the ESN performances.

453

454 Our results may appear rather counter-intuitive, as the weather and climate modelling
455 communities are moving towards extending simulations of physical processes to small scales.
456 As an example, we cite the use of highly-resolved convection-permitting simulations [55]
457 as well as the use of stochastic (and therefore non-smooth) parameterizations in weather
458 models [56]. We have, however, a few heuristic arguments on why the coarse-graining and
459 filtering operations should improve the ESN performances. First of all, the moving-average
460 operation helps both in smoothing the signal and by providing the ESN with a wider tem-
461 poral information. In some sense, this is reminiscent of the embedding procedure [57], where
462 the signal behavior is reconstructed by providing not only information on the previous time
463 step, but on previous times depending on the complexity. The filtering procedure can also
464 be motivated by the fact that the active degrees of freedom for the sea-level pressure data
465 are limited. This has been confirmed by [53] via coarse-graining these data and showing
466 that the active degrees of freedom are independent on the resolution, in the same range
467 explored in this study. In other words, including small scales in the learning of sea-level
468 pressure data, does not provide additional information on the dynamics and push towards
469 over-fitting and saturating the ESN with redundant information. The latter consideration
470 poses also some caveats on the generality of our results: we believe that this procedure is not
471 beneficial whenever a clear separation of scales is not achievable, e.g. in non-confined 3-D
472 turbulence. Moreover, in this study, note that three sources of stochasticity were present:
473 (i) in the random matrices and reservoir, (ii) in the perturbed initial conditions and (iii) in

474 the ESN simulations when using moving average filtered data with sampled *deltau* compo-
475 nents. The first one is inherent to the model definition. The perturbations of the starting
476 conditions allow characterizing the sensitivity of our ESN approach to the initial conditions.
477 The stochasticity induced by the additive noise *deltau* provides a distributional forecast at
478 each time t . Although this latter noise can be useful to simulate multiple trajectories and
479 evaluate their long-term behaviour, in practice, i.e., in the case where a ESN would be used
480 operationally to generate forecasts, one might not want to employ a stochastic formulation
481 with an additive noise, but rather the explicit and deterministic formulation in Eq. 11. This
482 exemplifies the interest of our ESN approach for possible distinction between forecasts and
483 long-term simulations, and therefore makes it flexible to adapt to the case of interest.

484

485 In future work, it will be interesting to use other learning architectures and other methods
486 of separating large- from small-scale components [58–60]. For example, our results give a
487 more formal framework for applications of machine learning techniques on geophysical data.
488 Deep-learning approaches have proven useful in performing learning at different time and
489 spatial scales whenever each layer is specialized in learning some specific features of the
490 dynamics [11, 61]. Indeed, several difficulties encountered in the application of machine
491 learning on climate data could be overcome if the appropriate framework is used, but this
492 requires a critical understanding of the limitations of the learning techniques.

493 **ACKNOWLEDGMENTS**

494 We acknowledge B D’Alena, J Brajard, Balaji, B Dubrulle, R Vautard, N Vercauteren, F
495 Daviaud, Y Sato for useful discussions. This work is supported by the CNRS INSU-LEFE-
496 MANU grant ”DINCLIC”.

-
- 497 [1] M. I. Jordan and T. M. Mitchell, Machine learning: Trends, perspectives, and prospects,
498 Science **349**, 255 (2015).
499 [2] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, nature **521**, 436 (2015).
500 [3] D. S. Bassett and O. Sporns, Network neuroscience, Nature neuroscience **20**, 353 (2017).

- 501 [4] C. Quinlan, B. Babin, J. Carr, and M. Griffin, *Business research methods* (South Western
502 Cengage, 2019).
- 503 [5] J.-X. Wang, J.-L. Wu, and H. Xiao, Physics-informed machine learning approach for recon-
504 structing reynolds stress modeling discrepancies based on dns data, *Physical Review Fluids*
505 **2**, 034603 (2017).
- 506 [6] M. Buchanan, *The limits of machine prediction*, Ph.D. thesis, Nature Publishing Group (2019).
- 507 [7] S. Bony, B. Stevens, D. M. Frierson, C. Jakob, M. Kageyama, R. Pincus, T. G. Shepherd,
508 S. C. Sherwood, A. P. Siebesma, A. H. Sobel, *et al.*, Clouds, circulation and climate sensitivity,
509 *Nature Geoscience* **8**, 261 (2015).
- 510 [8] J.-L. Wu, H. Xiao, and E. Paterson, Physics-informed machine learning approach for aug-
511 menting turbulence models: A comprehensive framework, *Physical Review Fluids* **3**, 074602
512 (2018).
- 513 [9] V. M. Krasnopolsky and M. S. Fox-Rabinovitz, Complex hybrid models combining determinis-
514 tic and machine learning components for numerical climate modeling and weather prediction,
515 *Neural Networks* **19**, 122 (2006).
- 516 [10] S. Rasp, M. S. Pritchard, and P. Gentine, Deep learning to represent subgrid processes in
517 climate models, *Proceedings of the National Academy of Sciences* **115**, 9684 (2018).
- 518 [11] P. Gentine, M. Pritchard, S. Rasp, G. Reinaudi, and G. Yacalis, Could machine learning break
519 the convection parameterization deadlock?, *Geophysical Research Letters* **45**, 5742 (2018).
- 520 [12] J. N. Liu, Y. Hu, Y. He, P. W. Chan, and L. Lai, Deep neural network modeling for big data
521 weather forecasting, in *Information Granularity, Big Data, and Computational Intelligence*
522 (Springer, 2015) pp. 389–408.
- 523 [13] A. Grover, A. Kapoor, and E. Horvitz, A deep hybrid model for weather forecasting, in
524 *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery*
525 *and Data Mining* (ACM, 2015) pp. 379–386.
- 526 [14] S. E. Haupt, J. Cowie, S. Linden, T. McCandless, B. Kosovic, and S. Alessandrini, Machine
527 learning for applied weather prediction, in *2018 IEEE 14th International Conference on e-*
528 *Science (e-Science)* (IEEE, 2018) pp. 276–277.
- 529 [15] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, Convolutional
530 lstm network: A machine learning approach for precipitation nowcasting, in *Advances in*
531 *neural information processing systems* (2015) pp. 802–810.

- 532 [16] X. Shi, Z. Gao, L. Lausen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo, Deep learning
533 for precipitation nowcasting: A benchmark and a new model, in *Advances in neural informa-*
534 *tion processing systems* (2017) pp. 5617–5627.
- 535 [17] M. Sprenger, S. Schemm, R. Oechslein, and J. Jenkner, Nowcasting foehn wind events using
536 the adaboost machine learning algorithm, *Weather and Forecasting* **32**, 1079 (2017).
- 537 [18] S. Scher and G. Messori, Predicting weather forecast uncertainty with machine learning, *Quar-*
538 *terly Journal of the Royal Meteorological Society* **144**, 2830 (2018).
- 539 [19] H. Jaeger, The echo state approach to analysing and training recurrent neural networks-
540 with an erratum note, Bonn, Germany: German National Research Center for Information
541 Technology GMD Technical Report **148**, 13 (2001).
- 542 [20] E. N. Lorenz, Deterministic nonperiodic flow, *Journal of the atmospheric sciences* **20**, 130
543 (1963).
- 544 [21] J. M. Hyman and B. Nicolaenko, The kuramoto-sivashinsky equation: a bridge between pde’s
545 and dynamical systems, *Physica D: Nonlinear Phenomena* **18**, 113 (1986).
- 546 [22] J. Pathak, Z. Lu, B. R. Hunt, M. Girvan, and E. Ott, Using machine learning to replicate
547 chaotic attractors and calculate lyapunov exponents from data, *Chaos: An Interdisciplinary*
548 *Journal of Nonlinear Science* **27**, 121102 (2017).
- 549 [23] J. Pathak, B. Hunt, M. Girvan, Z. Lu, and E. Ott, Model-free prediction of large spatiotem-
550 porally chaotic systems from data: A reservoir computing approach, *Physical review letters*
551 **120**, 024102 (2018).
- 552 [24] M. Xu, M. Han, T. Qiu, and H. Lin, Hybrid regularized echo state network for multivariate
553 chaotic time series prediction, *IEEE transactions on cybernetics* **49**, 2305 (2018).
- 554 [25] E. N. Lorenz, Predictability: A problem partly solved, in *Proc. Seminar on predictability*,
555 Vol. 1 (1996).
- 556 [26] T. Schneider, S. Lan, A. Stuart, and J. Teixeira, Earth system modeling 2.0: A blueprint
557 for models that learn from observations and targeted high-resolution simulations, *Geophysical*
558 *Research Letters* **44**, 12 (2017).
- 559 [27] S. Scher, Toward data-driven weather and climate forecasting: Approximating a simple general
560 circulation model with deep learning, *Geophysical Research Letters* **45**, 12 (2018).
- 561 [28] P. D. Dueben and P. Bauer, Challenges and design choices for global weather and climate
562 models based on machine learning, *Geoscientific Model Development* **11**, 3999 (2018).

- 563 [29] S. Scher and G. Messori, Weather and climate forecasting with neural networks: using general
564 circulation models (gcms) with different complexity as a study ground, *Geoscientific Model*
565 *Development* **12**, 2797 (2019).
- 566 [30] D. Schertzer, S. Lovejoy, F. Schmitt, Y. Chigirinskaya, and D. Marsan, Multifractal cascade
567 dynamics and turbulent intermittency, *Fractals* **5**, 427 (1997).
- 568 [31] G. Paladin and A. Vulpiani, Degrees of freedom of turbulence, *Physical Review A* **35**, 1971
569 (1987).
- 570 [32] P. Manneville, Intermittency, self-similarity and $1/f$ spectrum in dissipative dynamical sys-
571 tems, *Journal de Physique* **41**, 1235 (1980).
- 572 [33] R. E. Eskridge, J. Y. Ku, S. T. Rao, P. S. Porter, and I. G. Zurbenko, Separating different scales
573 of motion in time series of meteorological variables, *Bulletin of the American Meteorological*
574 *Society* **78**, 1473 (1997).
- 575 [34] Z. Shi and M. Han, Support vector echo-state machine for chaotic time-series prediction, *IEEE*
576 *Transactions on Neural Networks* **18**, 359 (2007).
- 577 [35] D. Li, M. Han, and J. Wang, Chaotic time series prediction based on a novel robust echo state
578 network, *IEEE Transactions on Neural Networks and Learning Systems* **23**, 787 (2012).
- 579 [36] X. Hinaut, *Réseau de neurones récurrent pour le traitement de séquences abstraites et de*
580 *structures grammaticales, avec une application aux interactions homme-robot*, Ph.D. thesis,
581 Thèse de doctorat, Université Claude Bernard Lyon 1 (2013).
- 582 [37] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical learning with sparsity: the lasso and*
583 *generalizations* (Chapman and Hall/CRC, 2015).
- 584 [38] W. G. Cochran, The χ^2 test of goodness of fit, *The Annals of Mathematical Statistics* , 315
585 (1952).
- 586 [39] J.-P. Eckmann and D. Ruelle, Ergodic theory of chaos and strange attractors, in *The theory*
587 *of chaotic attractors* (Springer, 1985) pp. 273–312.
- 588 [40] L.-S. Young, What are srb measures, and which dynamical systems have them?, *Journal of*
589 *Statistical Physics* **108**, 733 (2002).
- 590 [41] C. Bonferroni, Teoria statistica delle classi e calcolo delle probabilita, *Pubblicazioni del R*
591 *Istituto Superiore di Scienze Economiche e Commerciali di Firenze* **8**, 3 (1936).
- 592 [42] D. Faranda, V. Lucarini, G. Turchetti, and S. Vaienti, Generalized extreme value distribution
593 parameters as dynamical indicators of stability, *International Journal of Bifurcation and Chaos*

- 594 **22**, 1250276 (2012).
- 595 [43] F. Panichi and G. Turchetti, Lyapunov and reversibility errors for hamiltonian flows, *Chaos,*
596 *Solitons & Fractals* **112**, 83 (2018).
- 597 [44] P. J. Brockwell and R. A. Davis, *Introduction to time series and forecasting* (springer, 2016).
- 598 [45] A. Wolf, J. B. Swift, H. L. Swinney, and J. A. Vastano, Determining lyapunov exponents from
599 a time series, *Physica D: Nonlinear Phenomena* **16**, 285 (1985).
- 600 [46] S. Saha, S. Moorthi, X. Wu, J. Wang, S. Nadiga, P. Tripp, D. Behringer, Y.-T. Hou, H.-y.
601 Chuang, M. Iredell, *et al.*, The ncep climate forecast system version 2, *Journal of Climate* **27**,
602 2185 (2014).
- 603 [47] E. R. Weeks, Y. Tian, J. Urbach, K. Ide, H. L. Swinney, and M. Ghil, Transitions between
604 blocked and zonal flows in a rotating annulus with topography, *Science* **278**, 1598 (1997).
- 605 [48] D. Faranda, G. Masato, N. Moloney, Y. Sato, F. Daviaud, B. Dubrulle, and P. Yiou, The
606 switching between zonal and blocked mid-latitude atmospheric circulation: a dynamical sys-
607 tem perspective, *Climate Dynamics* **47**, 1587 (2016).
- 608 [49] N. Korabel and E. Barkai, Pesin-type identity for intermittent dynamics with a zero lyapunov
609 exponent, *Physical review letters* **102**, 050601 (2009).
- 610 [50] M. R. Frank, L. Mitchell, P. S. Dodds, and C. M. Danforth, Standing swells surveyed showing
611 surprisingly stable solutions for the lorenz'96 model, *International Journal of Bifurcation and*
612 *Chaos* **24**, 1430027 (2014).
- 613 [51] J. W. Hurrell, Decadal trends in the north atlantic oscillation: regional temperatures and
614 precipitation, *Science* **269**, 676 (1995).
- 615 [52] G. Moore, I. A. Renfrew, and R. S. Pickart, Multidecadal mobility of the north atlantic
616 oscillation, *Journal of Climate* **26**, 2453 (2013).
- 617 [53] D. Faranda, G. Messori, and P. Yiou, Dynamical proxies of north atlantic predictability and
618 extremes, *Scientific reports* **7**, 41278 (2017).
- 619 [54] L. Hudgins, C. A. Friehe, and M. E. Mayer, Wavelet transforms and atmopsheric turbulence,
620 *Physical Review Letters* **71**, 3279 (1993).
- 621 [55] G. Fosse, S. Khodayar, and P. Berg, Benefit of convection permitting climate model simula-
622 tions in the representation of convective precipitation, *Climate Dynamics* **44**, 45 (2015).
- 623 [56] A. Weisheimer, S. Corti, T. Palmer, and F. Vitart, Addressing model error through atmo-
624 spheric stochastic physical parametrizations: impact on the coupled ecmwf seasonal forecast-

- 625 ing system, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and*
626 *Engineering Sciences* **372**, 20130290 (2014).
- 627 [57] L. Cao, Practical method for determining the minimum embedding dimension of a scalar time
628 series, *Physica D: Nonlinear Phenomena* **110**, 43 (1997).
- 629 [58] S. Wold, K. Esbensen, and P. Geladi, Principal component analysis, *Chemometrics and intel-*
630 *ligent laboratory systems* **2**, 37 (1987).
- 631 [59] G. Froyland, G. A. Gottwald, and A. Hammerlindl, A computational method to extract macro-
632 scopic variables and their dynamics in multiscale systems, *SIAM Journal on Applied Dynam-*
633 *ical Systems* **13**, 1816 (2014).
- 634 [60] F. Kwasniok, The reduction of complex dynamical systems using principal interaction pat-
635 terns, *Physica D: Nonlinear Phenomena* **92**, 28 (1996).
- 636 [61] T. Bolton and L. Zanna, Applications of deep learning to ocean data inference and subgrid
637 parameterization, *Journal of Advances in Modeling Earth Systems* **11**, 376 (2019).

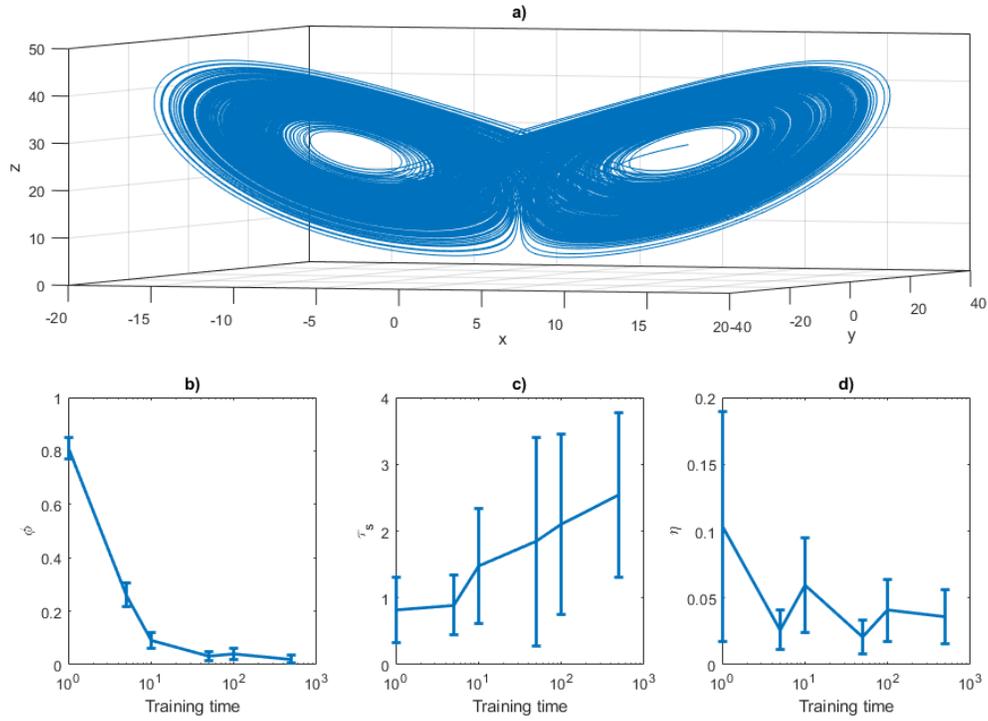


FIG. 1. a) Lorenz 1963 attractor obtained with a Euler scheme with $dt = 0.001$, $\sigma = 10$, $r = 28$ and $b = 8/3$. Panels b-d) show the performances indicator as a function of the training time. b) the rejection rate ϕ of the invariant density test for the x variable; c) the first time t such that the $\text{RMSE} > 1$; d) the initial error η . The error bar represents the average and the standard deviation of the mean over 100 realizations.

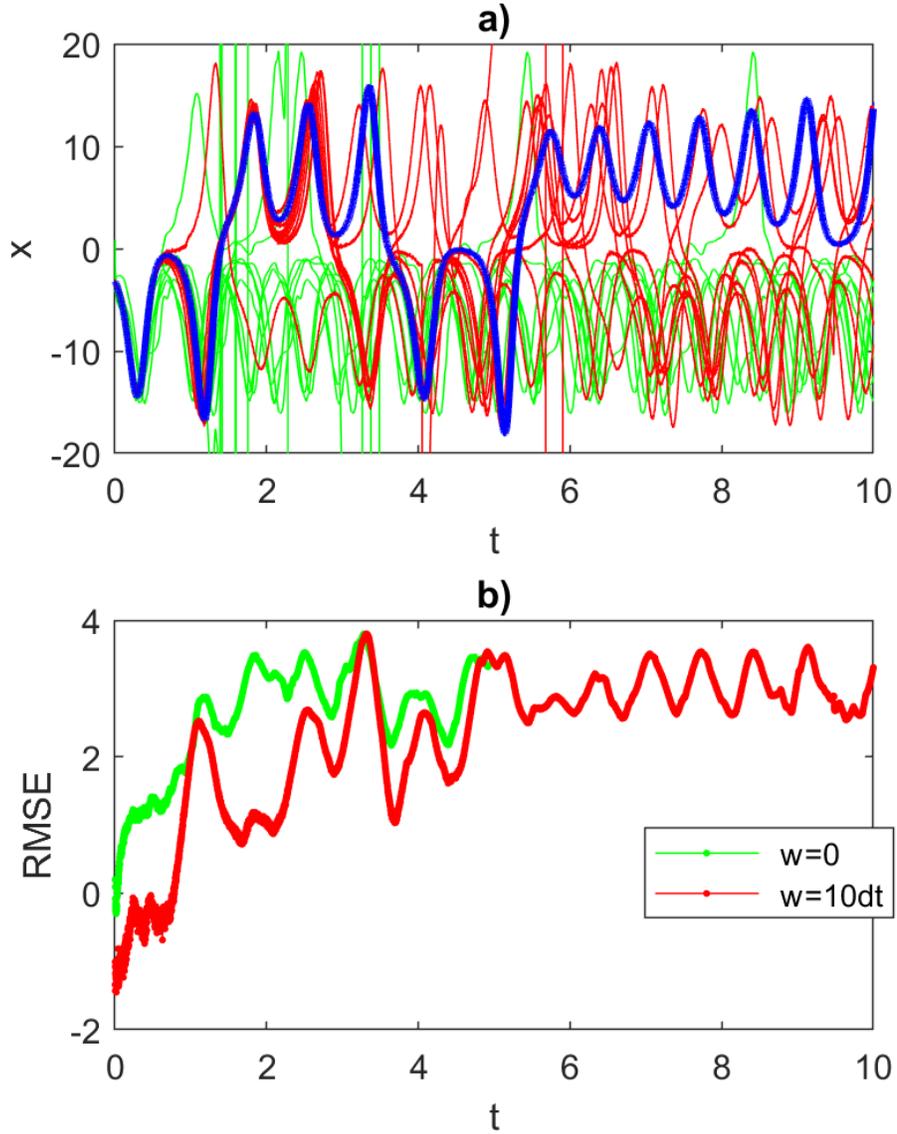


FIG. 2. a) Trajectories predicted using ESN on the Lorenz 1963 attractor for the variable x . The attractor is perturbed with Gaussian noise with variance $\epsilon = 0.1$. The target trajectory is shown in blue. 10 trajectories obtained without moving average (green) show an earlier divergence compared to 10 trajectories where the moving average is performed with a window size of $w = 10dt = 0.01$ (red). Panel (b) shows the evolution of the $\log(\text{RMSE})$, averaged over the trajectories for the cases with $w = 0.01$ (red) and $w = 0$ (green). The trajectories are all obtained after training the ESN for 10^5 time-steps. Each trajectory consists of 10^4 timesteps.

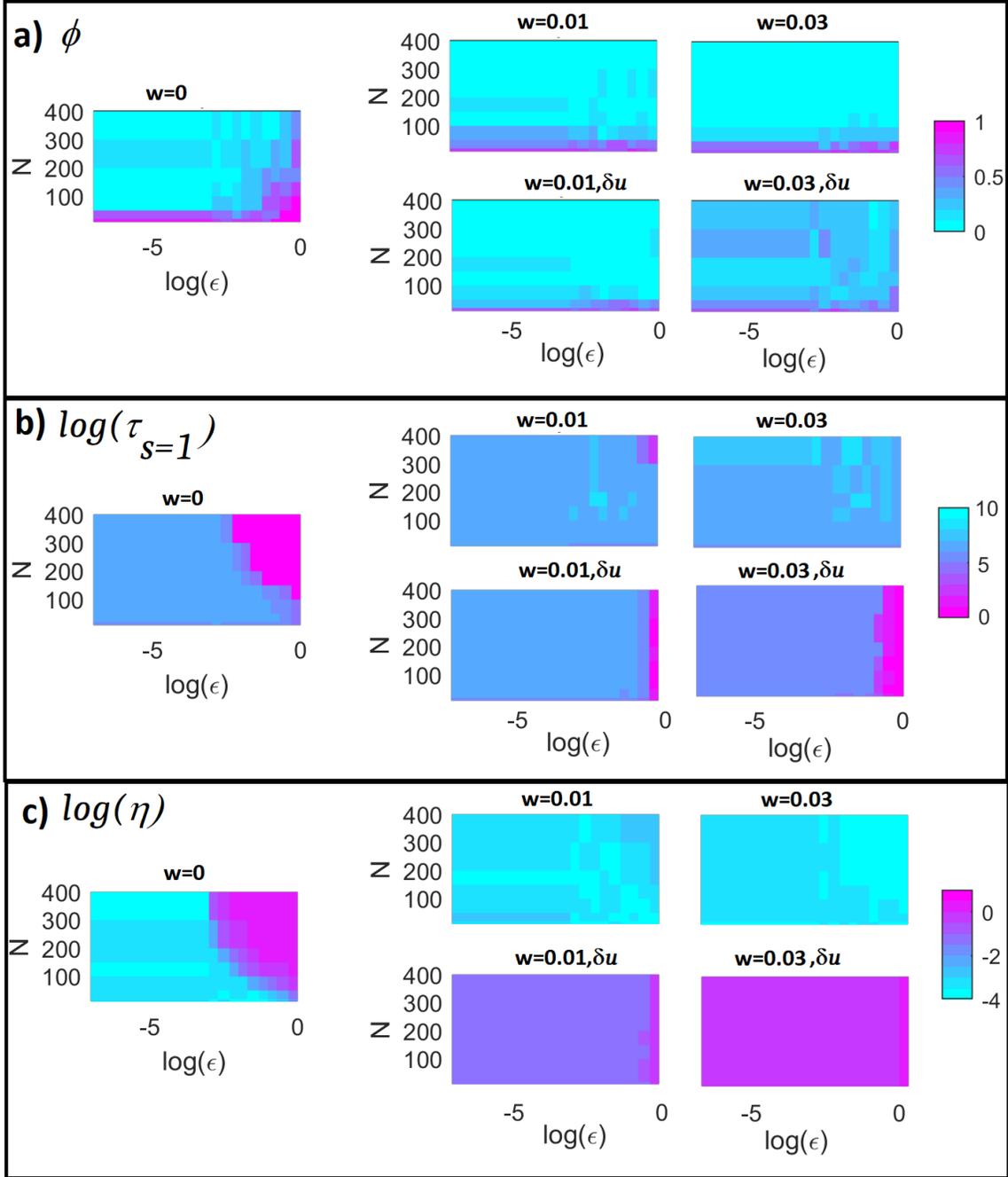


FIG. 3. Lorenz 1963 analysis for increasing noise intensity ϵ (x-axes), and number of neurons N (y-axes). The colorscale represents: ϕ the rate of failure of the χ^2 test (size $\alpha = 0.05$) (a); the logarithm of predictability horizon $\log(\tau_{s=1})$ (b); the logarithm of initial error $\log(\eta)$ (c). All the values are averages over 30 realizations. Left sub-panels refer to results without moving average, central sub-panels with averaging window $w = 0.01$, right hand-side panels with averaging window $w = 0.03$. Upper sub-panels are obtained using the exact expression in Eq. 11 and lower sub-panels using the residuals in Eq. 9. The trajectories are all obtained after training the ESN for 10^5 time-steps. Each trajectory consists of 10^4 time-steps.

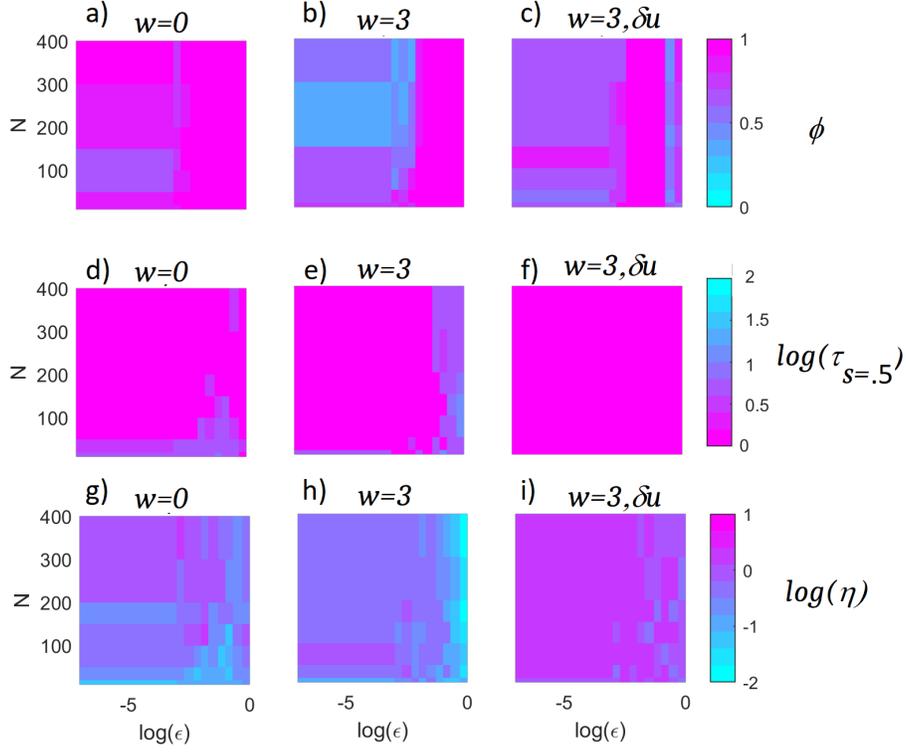


FIG. 4. Analysis of the Pomeau-Manneville system for increasing noise intensity ϵ (x-axes), and number of neurons N (y-axes). The colorscale represents: ϕ the rate of failure of the χ^2 test (size $\alpha = 0.05$) (a-c); the logarithm of predictability horizon $\log(\tau_{s=0.5})$ (d-f); the logarithm of initial error $\log(\eta)$ (g-i). All the values are averages over 30 realizations. Panels a,d,g) refer to results without moving average, b,c,e,f,h,i) with averaging window $w = 3$, c,f,i). Panels b,e,h) are obtained using the exact expression in Eq. 11 and c,f,i) using the residuals in Eq 9. The trajectories are all obtained after training the ESN for 10^5 time-steps. Each trajectory consists of 10^4 timesteps.

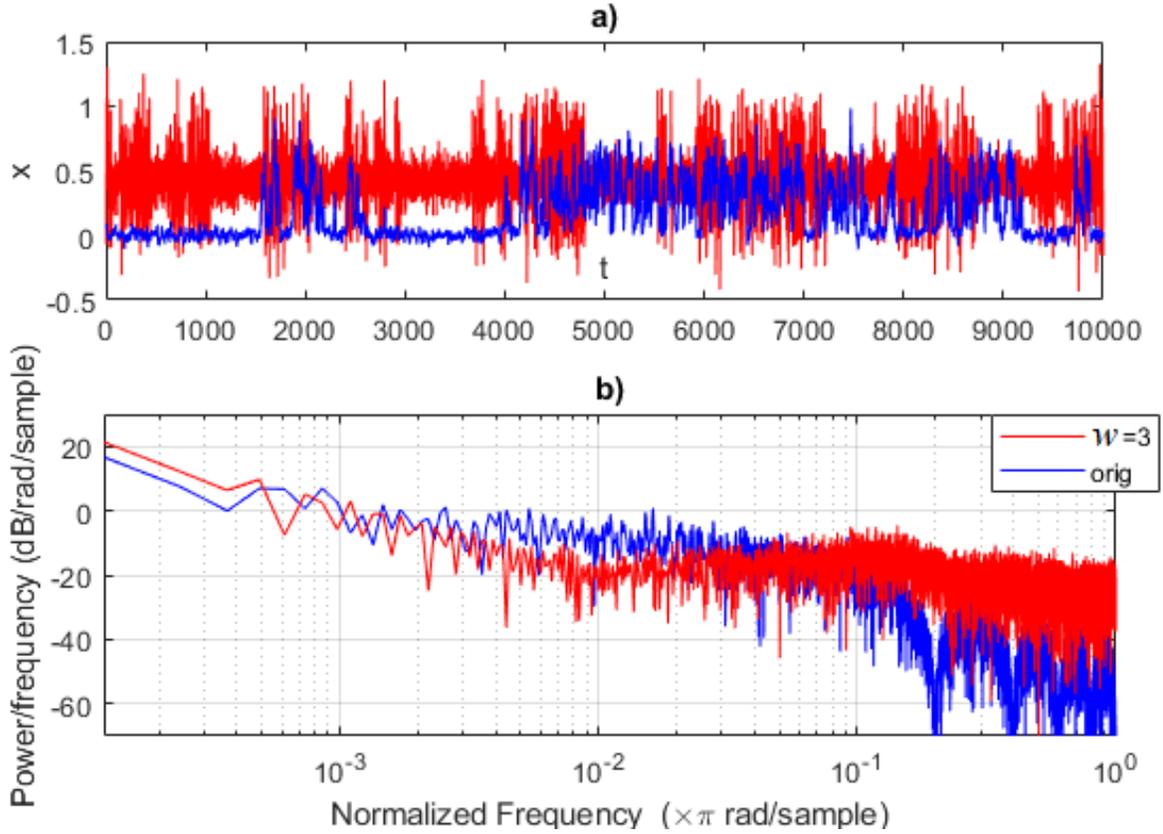


FIG. 5. Pomeau-Manneville ESN simulation (red) showing an intermittent behavior and compared to the target trajectory (blue). The ESN trajectory is obtained after training the ESN for 10^5 time-steps using the moving average time series with $w = 3$. It consists of 10^4 timesteps. Cases $w = 0$ are not shown as trajectories always diverge. Evolution of trajectories in time (a) and Fourier power spectra (b).

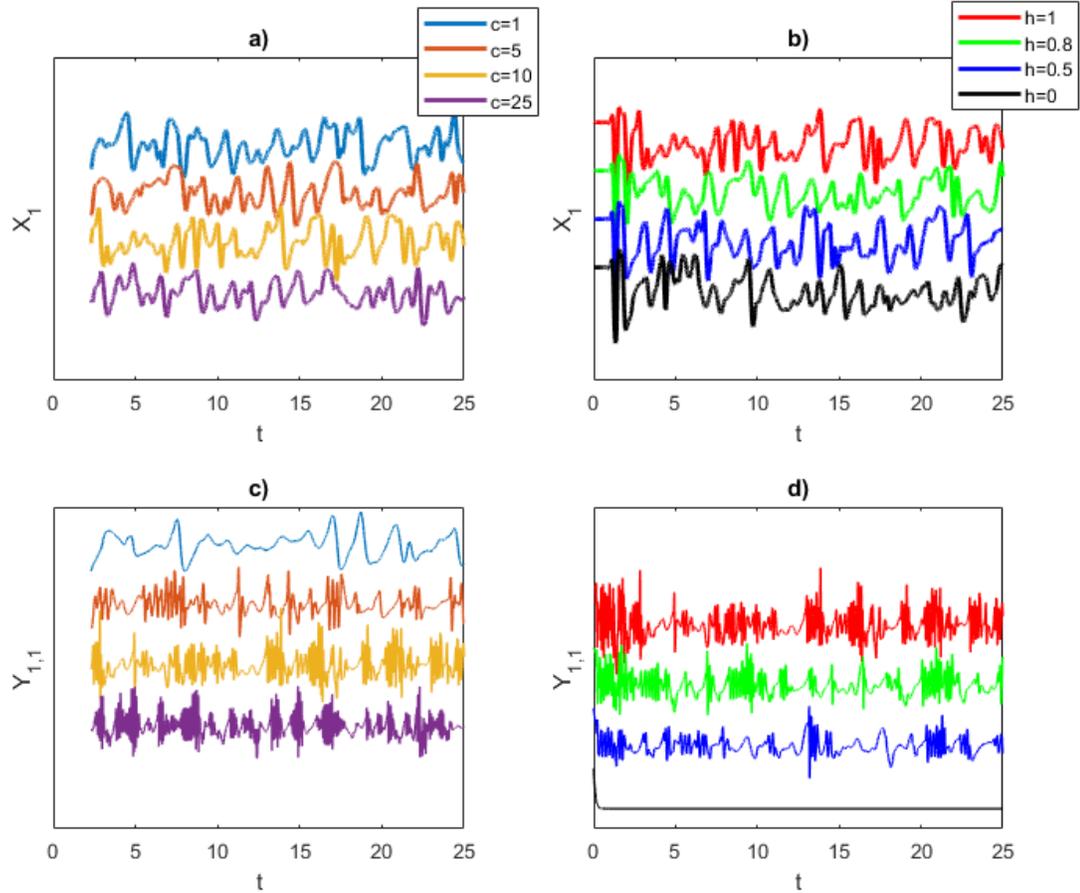


FIG. 6. Lorenz 1996 simulations for the large-scale variable X_1 (a,b) and small-scale variable $Y_{1,1}$ (c,d). Panels (a,c) show simulations varying c at fixed $h = 1$. The larger c , the more intermittent the behavior of the fast scales. Panels (b,d) show simulations varying the coupling h for fixed $c = 10$. When $h = 0$, there is no small-scale dynamics. y -axes are in arbitrary units, time-series are shifted for better visibility.

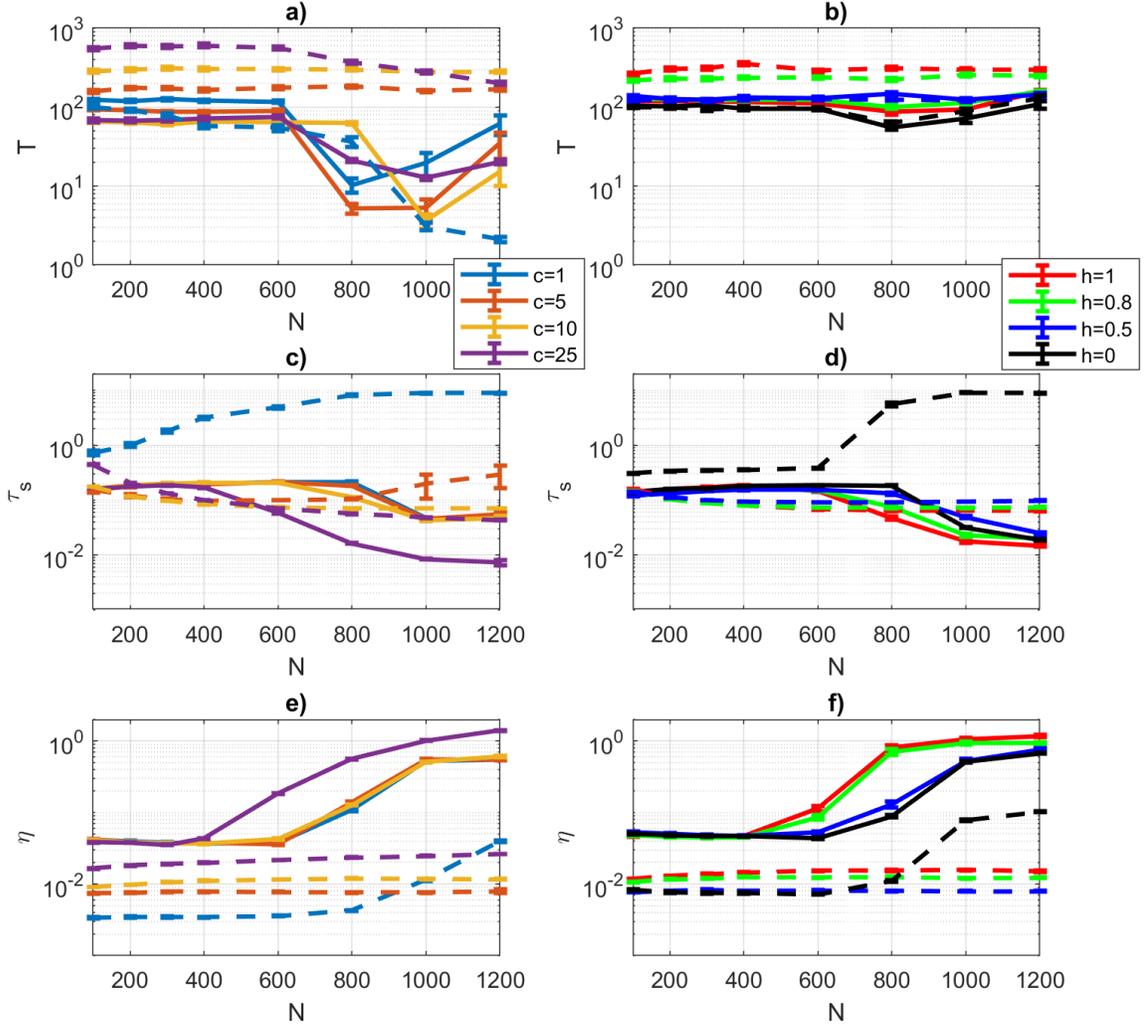


FIG. 7. Lorenz 1996 ESN prediction performance for the large-scale variables X only. a,b) χ^2 distance T ; (c,d) the predictability horizon τ_s with $s = 1$. (e,f) the initial error η in hPa. In (a,c,e) ESN predictions are made varying c at fixed $h = 1$. In (b,d,f) ESN predictions are made varying h at fixed $c = 10$. Continuous lines show ESN prediction performance made considering X variables only, dotted lines considering both X and Y variables.

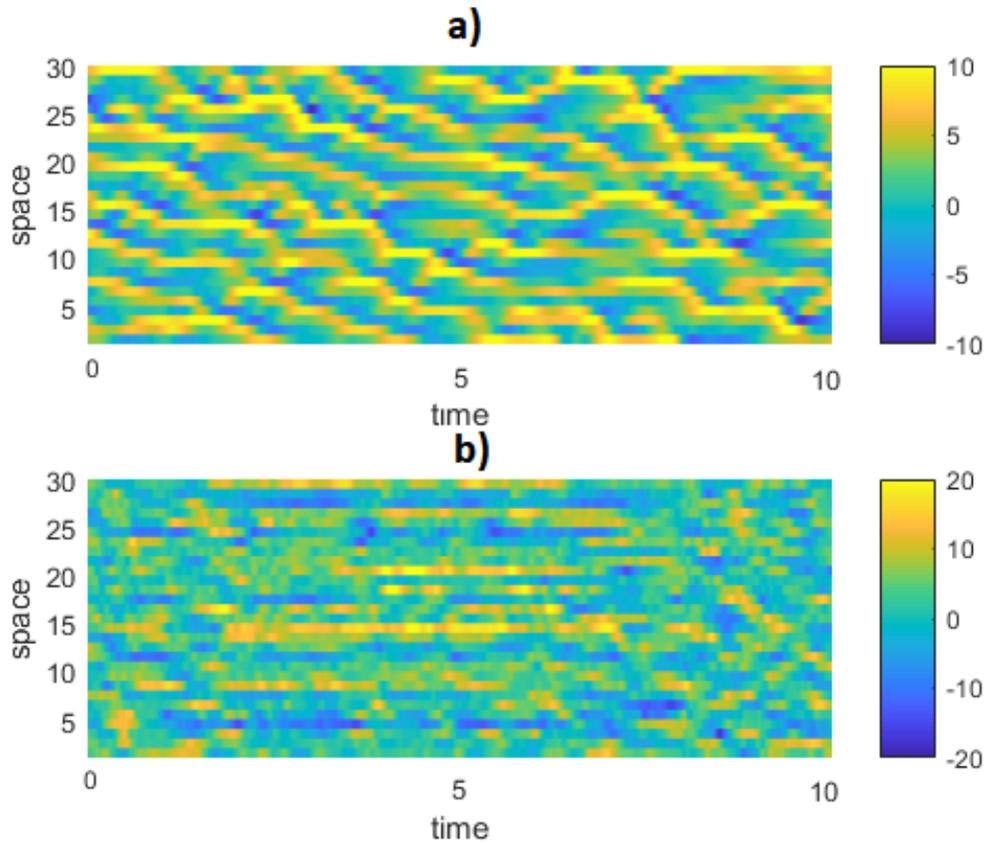


FIG. 8. Example of (a) target Lorenz 1996 spatio-temporal evolution of large-scale variables X for $c = 1, h = 1$ and (b) ESN prediction realized with $N = 800$ neurons. Note that the colors are not on the same scale for the two panels.

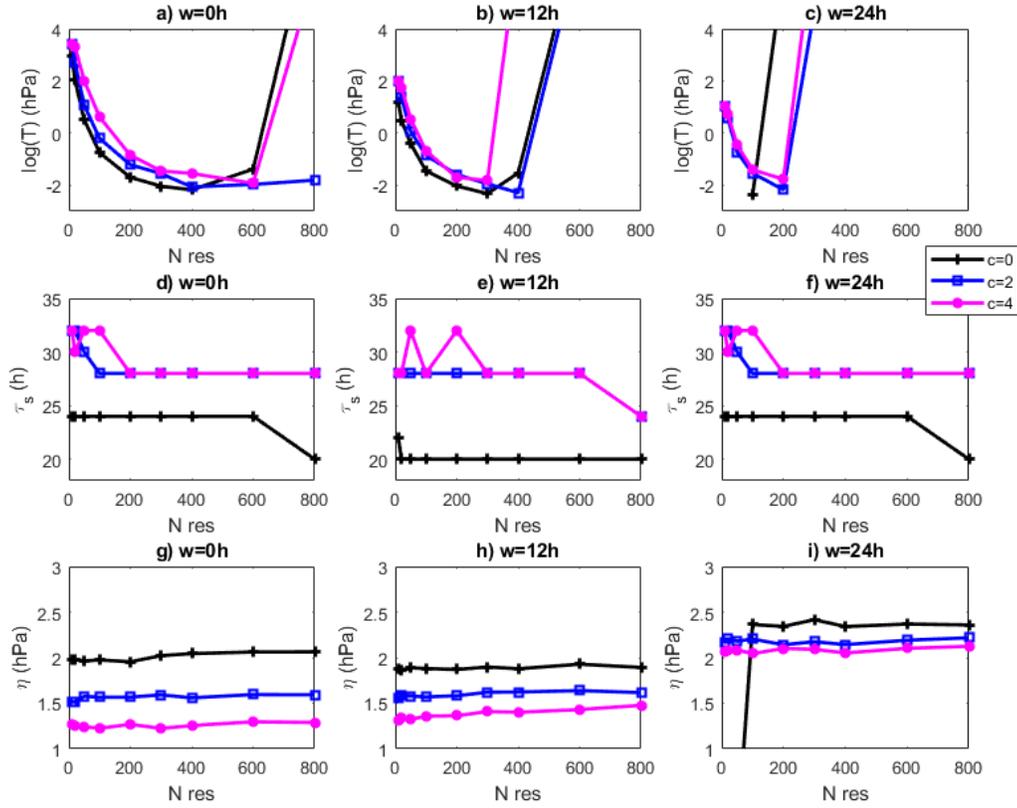


FIG. 9. Dependence of the quality of the results for the prediction of the sea-level pressure NCEPv2 data on the coarse graining factor c and on the moving average window size w . a-c) χ^2 distance T ; d-f) predictability horizon (in hours) τ_s , $s = 1.5$ hPa; g-i) logarithm of initial error η . Different coarse grain factor c are shown with different colors. a,d,g) $w = 0$, b,e,h) $w = 12$ h, c,f,i) $w = 24$ h.

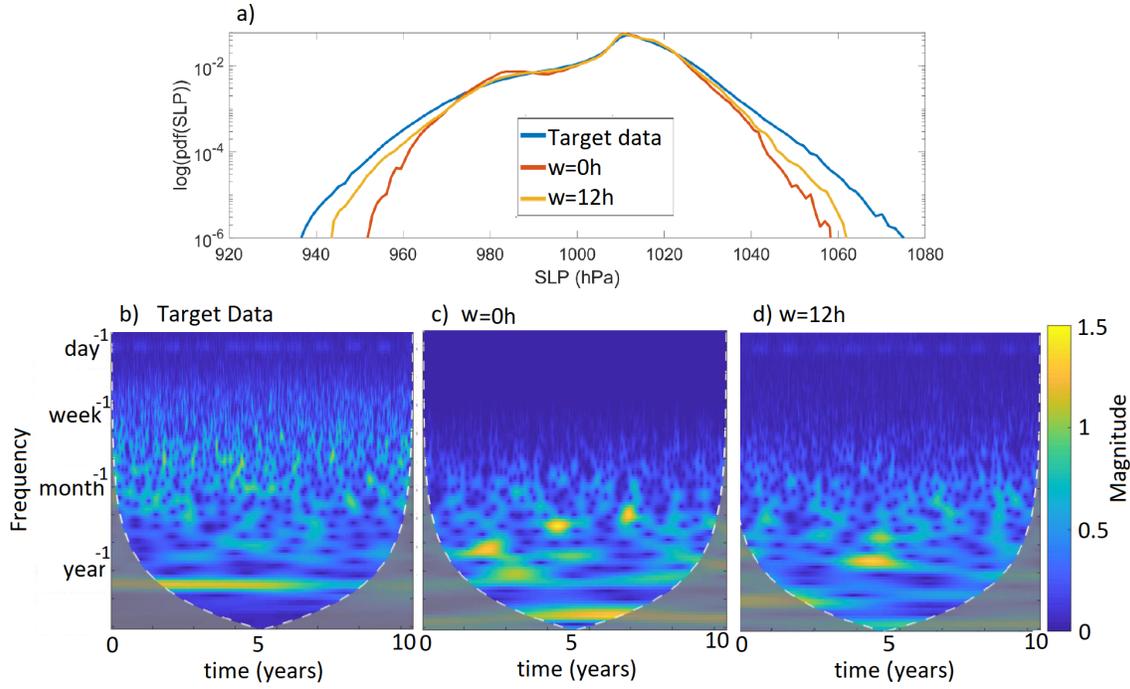


FIG. 10. a) Distributions of 10 years of 6h spatial and temporal data at all grid points obtained for the target NCEPv2 SLP data (blue), an ESN with $c = 4$ and $w = 0$ h (red), and an ESN with $c = 4$ and $w = 12$ h (orange). b-d) wavelet spectrograms for the NCEPv2 SLP target data (b), a run with $c = 4$ $w = 0$ h (c), and with $c = 4$ and $w = 12$ h (d).