Supplementary Material: Boosting performance in Machine Learning of Turbulent and Geophysical Flows via scale separation

D. Faranda*

LSCE- IPSL, CEA Saclay l'Orme des Merisiers,

CNRS UMR 8212 CEA-CNRS-UVSQ,

Université Paris-Saclay, 91191 Gif-sur-Yvette, France.

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

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

LSCE-IPSL, CEA Saclay l'Orme des Merisiers,

CNRS UMR 8212 CEA-CNRS-UVSQ,

Université Paris-Saclay, 91191 Gif-sur-Yvette, France.

V. Gautard

DRF/IRFU/DEDIP//LILAS Departement d'Electronique

des Detecteurs et d'Informatique pour la Physique,

CEA Saclay l'Orme des Merisiers,

91191 Gif-sur-Yvette, France.

Content

The supplementary material contains methods description (I) and 8 supplementary figures (II). Section (A) contains the numerical code for the computation of ESN; (B) the description of the moving average discrete and continuous filters; (C) the statistical test for the invariant distributions; (D) description and additional analyses of the systems and data analysed.

I. METHODS

A. Numerical code for ESN

We report here the MATLAB code used for the computation of the Echo State Network. This code is adapted from the original code available here: https://mantas.info/code/simple$_esn/whichcomeswit$
$onefor the training phase and one for the prediction phase$

ESN training

```
function [Win, W, Wout]=RNN_training(data,Nres)

%This function train the Echo State network using the data provided.
%INPUTS:
%data: a matrix of the input data to train, arranged as space X time
%Nres: the number of neurons N to be used in the training
%OUTPUTS:
%Win: the input weight matrix which consists of random weights
%W: the network of neurons
%Wout: the output weights, they are adjusted to match the next iterations

inSize = size(data,1);
trainLen= size(data,2);
Win = (rand(Nres,1+inSize)-0.5) .* 1;
W = rand(Nres,Nres)-0.5;
% normalizing and setting spectral radius
opt.disp = 0;
rhoW = abs(eigs(W,1,'LM',opt));
W = W .* ( 1.25 /rhoW);
% memory allocation
X = zeros(1+inSize+Nres,trainLen-1);
```

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

```
Yt = data(:,2:end)';
x = zeros(Nres,1);
for t = 1:trainLen-1
u = data(:,t);
x = tanh( Win*[1;u] + W*x );
X(:,t) = [1;u;x];
end
reg = 1e-8;  % regularization coefficient
Wout = ((X*X' + reg*eye(1+inSize+Nres)) \ (X*Yt))';
end
```

ESN Prediction

```
function [Y_pred]=RNN_prediction(data,Win, W, Wout)

% This function returns the recurrent Echo State Network prediction
%INPUT:
%data: the full data matrix of the data to predict in the form (space*time)
%Win: input weights
%W: neurons matrix
%Wout: output weights
%OUTPUT:
%Y_pred: the RNN prediction
Y_pred = zeros(size(data,1),size(data,2) );
x = zeros(size(W,1),1);
u=data(:,1);
for t = 1:size(data,2)
x = tanh( Win*[1;u] + W*x );
y = Wout*[1;u;x];
Y_pred(:,t) = y;
u = y;
end
end
```

B. Moving average filter

The simple moving average filter can be seen a nonparametric time series smoother (see e.g. [1], chapter 1.5). It can be applied to smooth out (relatively) high frequencies in a time series, both to de-noise the observations of a process or to estimate trend components, if present. It consists of replacing the observation $x(t)$ by a value $y(t)$, obtained by averaging the previous w observations. If the time is discrete (like in the Pomeau Manneville system) it is defined as:

$$y(t) = \frac{1}{w} \sum_{i=0}^{w-1} x(t-i),$$

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

$$y(t) = \frac{1}{w} \int_{t-w}^{t} x(s) ds$$

In practice the computation always refers to the discrete time case, as even continuous time systems are sampled at finite time steps. Since Echo State Networks are known to be sensitive to noise (see e.g. [2]), we exploit the simple moving average filter to smooth out high-frequency noise and assess the results for different smoothing windows w. We find that the choice of the moving averaging window w must respect two conditions: it should be large enough to smooth the noise but smaller than the characteristic time τ of the large-scale fluctuations of the system. For chaotic systems, τ can be derived knowing the rate of exponential divergence of the trajectories, a quantity linked to the Lyapunov exponents [3], and τ is knonw as Lyapunov time.

C. Statistical tests for Invariant distributions

As a first diagnostic of the performance of ESN, we aim at assessing whether or not 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 perform a χ^2

test [4], conducted as follows. Let X be a random variable - in our case the system observable - with domain R_X and probability density function $f_X(x)$, and let x be a sample from X. Let now $h_i(x)$ be an approximation of $f_X(x)$, namely the histogram of x over $i = 1, \ldots, M$ bins.Note that, if x spans the entire phase space, $h_i(x)$ is the numerical approximation of the Sinai-Ruelle-Bowen measure of the system [5]. Let now $\hat{x}$ be the forecast sample, $g_X(\hat{x})$ its probability density function and $\hat{h}_i(\hat{x})$ be 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:

$$H_0: \quad f_X(x) = g_X(\hat{x}) \quad \text{for all } x \in R_X$$
$$H_1: \quad f_X(x) \neq g_X(\hat{x}) \quad \text{for any } x \in R_X$$

Under H_0, $h_i(x)$ is the expected value of $\hat{h}_i(\hat{x})$, which implies that observed differences $(\hat{h}_i(\hat{x}) - h_i(x))$ are due to random errors, and are then independent and identically distributed Gaussian random variables. Statistical theory shows that, given H_0 true, the test statistics

$$T = \sum_{i=1}^{M} \frac{(\hat{h}_i(\hat{x}) - h_i(x))^2}{h_i(x)} \tag{1}$$

is distributed as a chi-squared random variable with M degrees of freedom, $\chi^2(M)$. Then, to test the null hypothesis at the level α, the observed value of the test statistics T is compared to the critical value corresponding to the $1 - \alpha$ percentile of the chi-square distribution, $T_c = \chi^2_{1-\alpha}(M)$: if $T > T_c$, the null hypothesis must be rejected.

In our setup, we encounter two limitations in using this standard χ^2 test. First, problems may arise when $h_i(x) = 0$, i.e. if the support of the sample distribution is wider than the support of the invariant distribution of the system. We observe this in a relatively small number of cases; since aggregating the bins would introduce unwanted complications, we decide to discard the pathological cases, controlling the effect empirically as described below. Moreover, even producing relatively large samples, we are not able to actually observe the invariant distribution of the considered system, which would require much longer simulations. As a consequence, we observe excessive rejection rates when testing samples generated under H_0.

We decide to control these two effects by using a Monte Carlo approach. To this purpose,

we simulate 10000 samples under the null hypothesis (i.e. using the system equation), and we compute the test statistic for each one according to equation 1. Then, we use the $(1-\alpha)$ percentile of the empirical distribution of T - instead the theoretical $\chi^2(M)$ - to determine the critical threshold T_c. As a last remark, we should notice that we are in the case of repeated tests, as the performance of the ESN is tested 100 times. In such cases, testing each sample separately at the chosen level α induces an increase in the observed rejection rate: in fact, extreme cases become more likely when many samples are drawn, even from H_0, and tested, resulting in an increased probability to erroneously reject the null hypothesis. To avoid this problem, we apply the Bonferroni correction ([6]), testing each one of the m available samples at the level $\alpha' = \frac{\alpha}{m}$. The level α used in the paper is 0.05

D. Systems analyzed

Lorenz 1963 equations

The Lorenz [7] system is a simplified model of Rayleigh-Benard convection, derived by E.N. Lorenz. It is an autonomous continuous dynamical system with three variables x, y and z parametrizing respectively the convective motion, the horizontal temperature gradient and the vertical temperature gradient. It writes:

$$\begin{aligned} \frac{dx}{dt} &= \sigma(y-x) \\ \frac{dy}{dt} &= -xz + \varrho x - y, \\ \frac{dz}{dt} &= xy - bz, \end{aligned} \tag{2}$$

where σ, r and b are three parameters, σ mimicking the Prandtl number and ϱ the reduced Rayleigh number. The Lorenz model is usually defined using equations (2), with $\sigma = 10$, $\varrho = 28$ and $b = 8/3$. The trajectory used in this article is shown in Supplementary Figure 1 and it has been obtained via integrating numerically the Lorenz equations with an Euler scheme ($dt = 0.001$). The dependence on the ESN learning from the training length are studied in Supplementary Figure 2, which suggests that $\sim 10^5$ time steps is a sufficient choice for the training set length. The maximum Lyapunov exponent of the system is $\lambda = 0.9$, so that the Lyapunov time $\tau \approx \mathcal{O}\left(\frac{1}{\lambda}\right) \approx 1.1$. Supplementary Figure 3 shows the benefit

of applying a moving average filter of window size $w = 10dt$ to perform ESN prediction. Panel (a) shows 10 trajectories obtained with (red) and without (green) moving average and compared to the reference trajectory (blue). As suggested by the visual inspection, the RMSE analysis (panel b) shows an evident gain of performance when the moving average procedure is applied. Note that this improvement relies on the choice of an averaging window representing the best trade-off between effective de-noising and loss of information, so that too large smoothing windows result in a deterioration of the performance of the network (other than a less accurate representation of the system dynamics). For example, supplementary Figure 4 shows the decrease in performance of the ESN between our chioce of $w = 10$ and the larger value $w = 50$.

Pomeau Manneville intermittent map

Several dynamical systems, including Earth climate, display intermittency, i.e. the time series of a variable issued by the system can experience sudden chaotic fluctuations, as well as a predictable behavior where the observables have small fluctuations. In atmospheric dynamics, such behavior is observed in the switching between zonal and meridional phases of the mid-latitude dynamics if a time series of the wind speed at one location is observed: when a cyclonic structure passes through the area, the wind has high values and large fluctuations, when an anticyclonic structure is present the wind is low and and fluctuations are smaller [8, 9]. It is then of practical interest to study the performance of ESN in Pomeau Manneville predictions as they are then hopefully applicable to climate data.

In particular, the Pomeau-Manneville [10] map is probably the simplest example of intermittent behavior, produced by a 1D discrete deterministic map given by:

$$x_{t+1} = \mathrm{mod}(x_t + x_t^{1+a}, 1), \tag{3}$$

where $0 < a < 1$ is a parameter. We use $a = 0.91$ in this study and a trajectory consisting of $5{\times}10^5$ iterations (see Supplementary Figure 5-a) for the trajectory and (b) for the invariant density $\rho(x)$). It is well known that Pomeau-Manneville systems exhibit sub-exponential separation of nearby trajectories and then the Lyapunov exponent is $\lambda = 0$. However, one can define a Lyapunov exponent for the non-ergodic phase of the dynamics and extract a characteristic time scale [11]. From this latter reference we can derive a value $\lambda \simeq 0.2$ for $a = 0.91$, implying $w < \tau \simeq 5$. We find that the best results are obtained for $w = 3$.

The NCEP sea-level pressure data

In this study we adopt the 6 hourly sea-level pressure (SLP) field as the meteorological variable proxy for the atmospheric circulation. It traces cyclones (resp. anticyclones) with minima (resp. maxima) of the SLP fields. The major modes of variability affecting mid-latitudes weather are often defined in terms of the empirical orthogonal functions of SLP and a wealth of other atmospheric features [12, 13], ranging from teleconnection patterns to storm track activity to atmospheric blocking can be diagnosed from the SLP field. We base our study on NCEP/NCAR reanalysis version 2 [14] data over the period 1979-2019, with a horizontal resolution of 2.5°.

In analogy with the time moving average filter (see section B) of this supplementary material, we investigate the effect of spatial coarse-graining the SLP fields by a factor c and perform the learning on the reduced fields. We use the nearest neighbor approximation, which consist in taking from the original dataset, the closest value to the coarse grid. Compared with methods based on averaging or dimension reduction techniques such as the Emprical Orthogonal Functions, the nearest neighbors approach has the advantage of not removing the extremes and preserve cyclonic and anticyclonic structures. An illustration of the obtained coarse grained field for the 01/01/1981 is provided in Supplementary Figure 6. For $c = 2$ we obtain an horizontal resolution of 5° and for $c = 4$ a resolution 10°. For $c = 4$ the information on the SLP field close to the poles is lost. However, in the remaining of the geographical domain, the coarse grained field still capture the positions of cyclonic and anticyclonic structures. Indeed, as shown in [15], this coarse grain field still preserves the dynamical properties of the original one. There is therefore a certain amount of redundant information on the original 2.5° horizontal resolution SLP fields.

The 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 is shown in Supplementary Figure 7. Panels a-c) show the distance from the invariant density, using the χ^2 divergence. Here it is clear that by increasing w, we get better forecast with smaller network sizes N, while small differences are found when the SLP fields are coarse grained. A large difference for the predicatibility expressed as saturation time τ_s, $s = 1.5$ hPa (panels d-f) emerges when SLP fields are coarse grained. We gain up to 10h in the predictability horizon with respect to the forecasts performed on the original fields ($c = 0$).

This gain is also reflected by the initial error η (panels g-i). From the combination of all the indicators, after a visual inspection, we can identify the best-set of parameters: $w = 12$ h, $N = 200$ and $c = 4$. Indeed this is the case such that, with the smallest network we get the almost the minimal χ^2 distance, the highest predictability (32 h) and one of the lowest initial errors. We also remark that, for $c = 0$ (panels (c) and (i)), the fit always diverges for small network sizes. We have used the best-set of parameters in the main text.

We compare in details the results obtained for two 10 years prediction with $w = 0$ h and $w = 12$ h at $N = 200$ and $c = 4$ fixed. At the beginning of the forecast time (Supplementary Video 1), the target field (panel a) is close to both that obtained with $w = 0h$ (panel b) and $w = 12$ h (panel c). However, when looking at a very late time (Supplementary Video 2), of course we do not expect to see agreement among the three datasets. Indeed we are well beyond the predictability horizon. However, we remark that the dynamics for the run $w = 0$ h is steady: positions of cyclones and anticyclones barely evolve with time. Instead, the run $w = 12$ h shows a richer dynamical evolution with generation and annihilation of cyclones.

In order to assess the two performances of ESN with and without moving average in a more quantitative way we present the space-time distributions in Supplementary Figure 8-a). The distribution obtained for the moving average $w = 12$ h has more realistic tails and matches better than the run $w = 0$ h that of the target data. Supplementary Figure 8b-d) shows the wavelet spectrograms (or scalograms) [16]. The scalogram is the absolute value of the continuous wavelet transform of a signal, plotted as a function of time and frequency. The target data spectrogram (b) presents a rich structure at different frequencies and some interannual variability. The wavelet spectrogram of non-filtered ESN run $w = 0$ h (c) shows no short time variability and too large interseasonal and interannual variability. The spectrogram of the target data is better matched by the run with $w = 12$ h (d) which shows that on time scales of days to weeks there is a larger variability.

[1] P. J. Brockwell and R. A. Davis, *Introduction to time series and forecasting* (springer, 2016).

[2] Z. Shi and M. Han, Support vector echo-state machine for chaotic time-series prediction, IEEE

Transactions on Neural Networks **18**, 359 (2007).

[3] A. Wolf, J. B. Swift, H. L. Swinney, and J. A. Vastano, Determining lyapunov exponents from a time series, Physica D: Nonlinear Phenomena **16**, 285 (1985).

[4] W. G. Cochran, The $\chi 2$ test of goodness of fit, The Annals of Mathematical Statistics , 315 (1952).

[5] J.-P. Eckmann and D. Ruelle, Ergodic theory of chaos and strange attractors, in *The theory of chaotic attractors* (Springer, 1985) pp. 273–312.

[6] C. Bonferroni, Teoria statistica delle classi e calcolo delle probabilita, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze **8**, 3 (1936).

[7] E. N. Lorenz, Deterministic nonperiodic flow, Journal of the atmospheric sciences **20**, 130 (1963).

[8] E. R. Weeks, Y. Tian, J. Urbach, K. Ide, H. L. Swinney, and M. Ghil, Transitions between blocked and zonal flows in a rotating annulus with topography, Science **278**, 1598 (1997).

[9] D. Faranda, G. Masato, N. Moloney, Y. Sato, F. Daviaud, B. Dubrulle, and P. Yiou, The switching between zonal and blocked mid-latitude atmospheric circulation: a dynamical system perspective, Climate Dynamics **47**, 1587 (2016).

[10] P. Manneville, Intermittency, self-similarity and 1/f spectrum in dissipative dynamical systems, Journal de Physique **41**, 1235 (1980).

[11] N. Korabel and E. Barkai, Pesin-type identity for intermittent dynamics with a zero lyaponov exponent, Physical review letters **102**, 050601 (2009).

[12] J. W. Hurrell, Decadal trends in the north atlantic oscillation: regional temperatures and precipitation, Science **269**, 676 (1995).

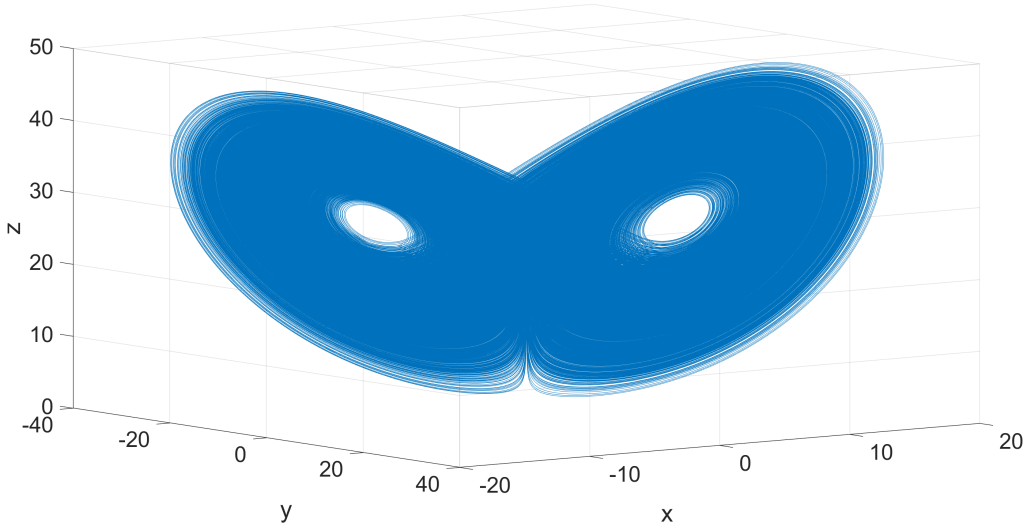
[13] G. Moore, I. A. Renfrew, and R. S. Pickart, Multidecadal mobility of the north atlantic oscillation, Journal of Climate **26**, 2453 (2013).

[14] S. Saha, S. Moorthi, X. Wu, J. Wang, S. Nadiga, P. Tripp, D. Behringer, Y.-T. Hou, H.-y. Chuang, M. Iredell, *et al.*, The ncep climate forecast system version 2, Journal of Climate **27**, 2185 (2014).

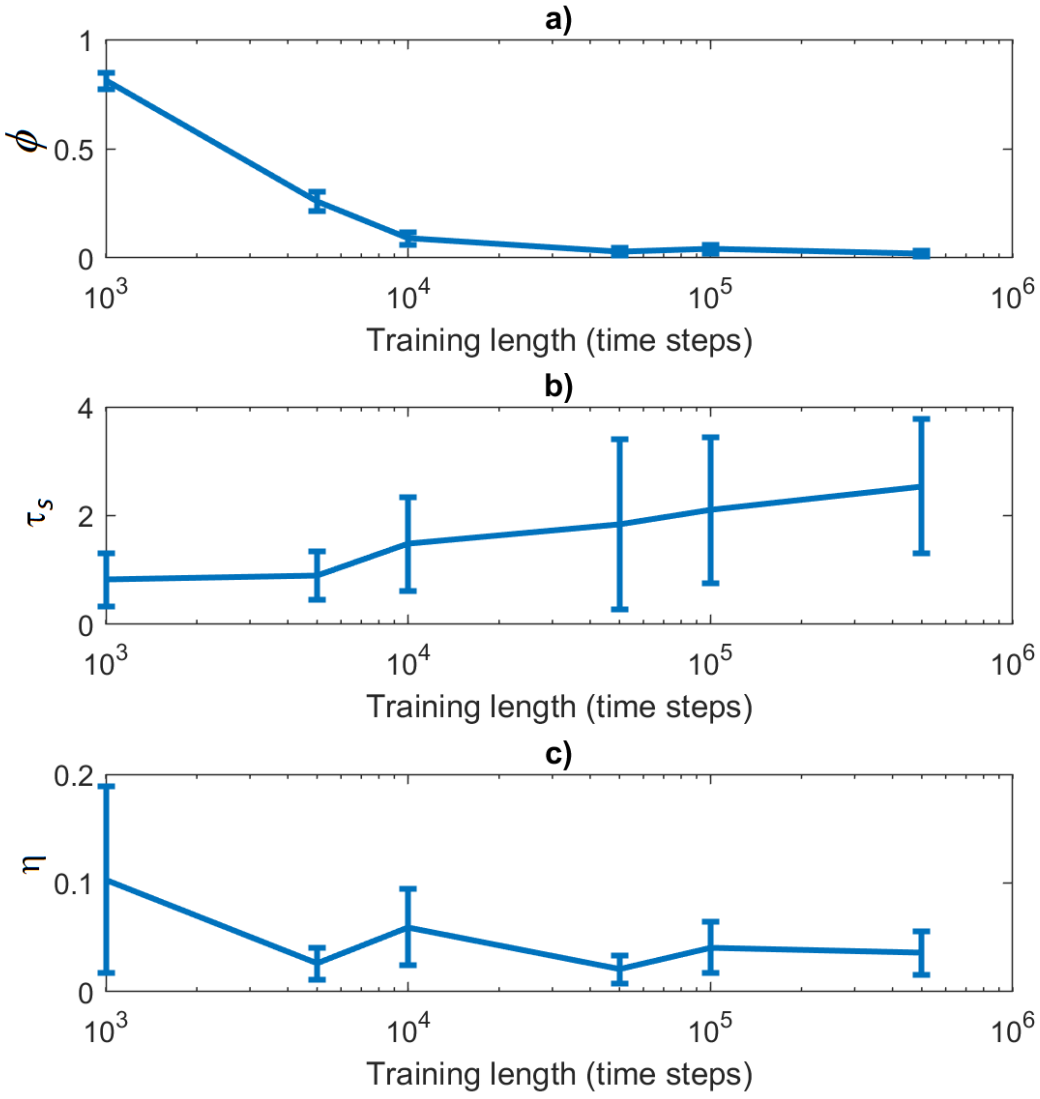
[15] D. Faranda, G. Messori, and P. Yiou, Dynamical proxies of north atlantic predictability and extremes, Scientific reports **7**, 41278 (2017).

[16] L. Hudgins, C. A. Friehe, and M. E. Mayer, Wavelet transforms and atmopsheric turbulence, Physical Review Letters **71**, 3279 (1993).

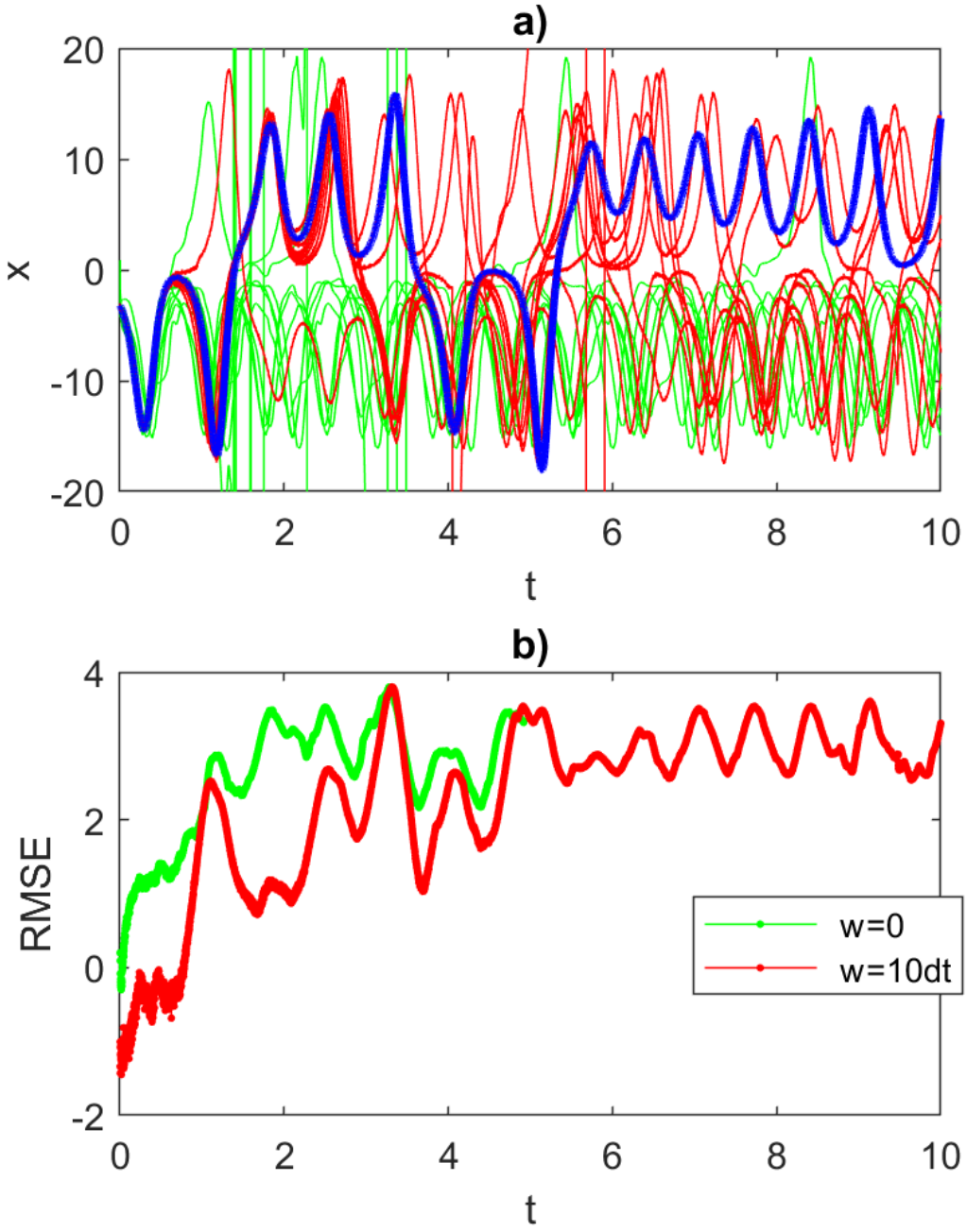
II. SUPPLEMENTARY FIGURES



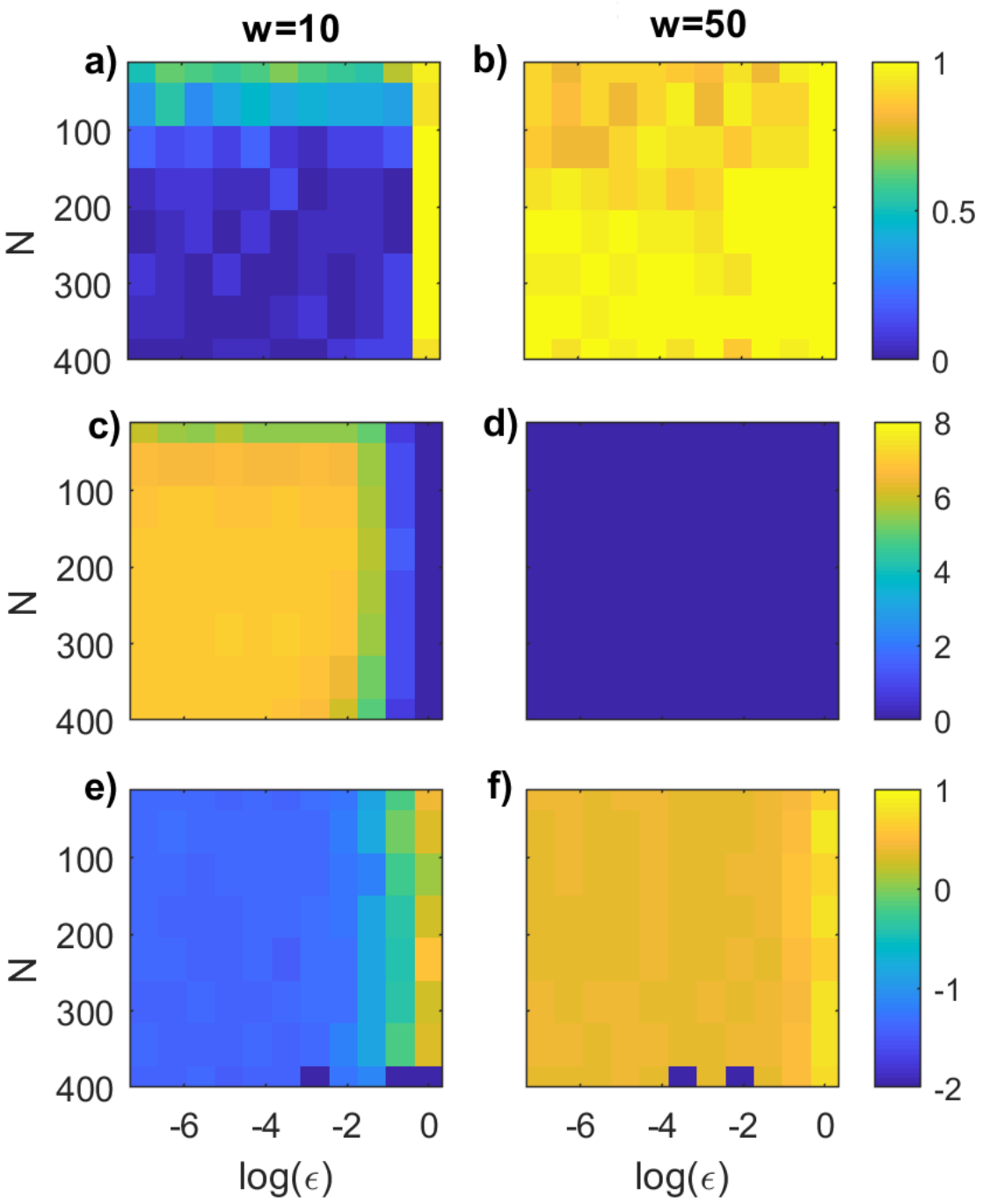
Supplementary Figure 1. Lorenz attractor obtained with an Euler scheme with $dt = 0.001$, $\sigma = 10$, $r = 28$ and $b = 8/3$.



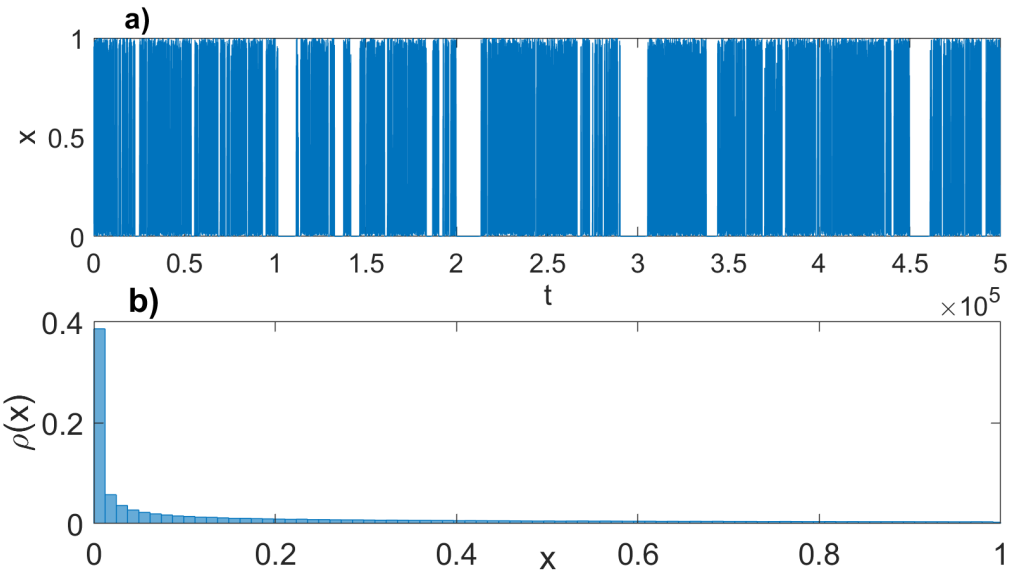
Supplementary Figure 2. Dependence of the ESN performance on the learning window for the Lorenz attractor. a) the rejection rate ϕ of the invariant density test for the x variable; b) the first time t such that the RMSE>1; c) the initial error η. The errorbar represents the average and the standard deviation of the mean over 100 realizations.



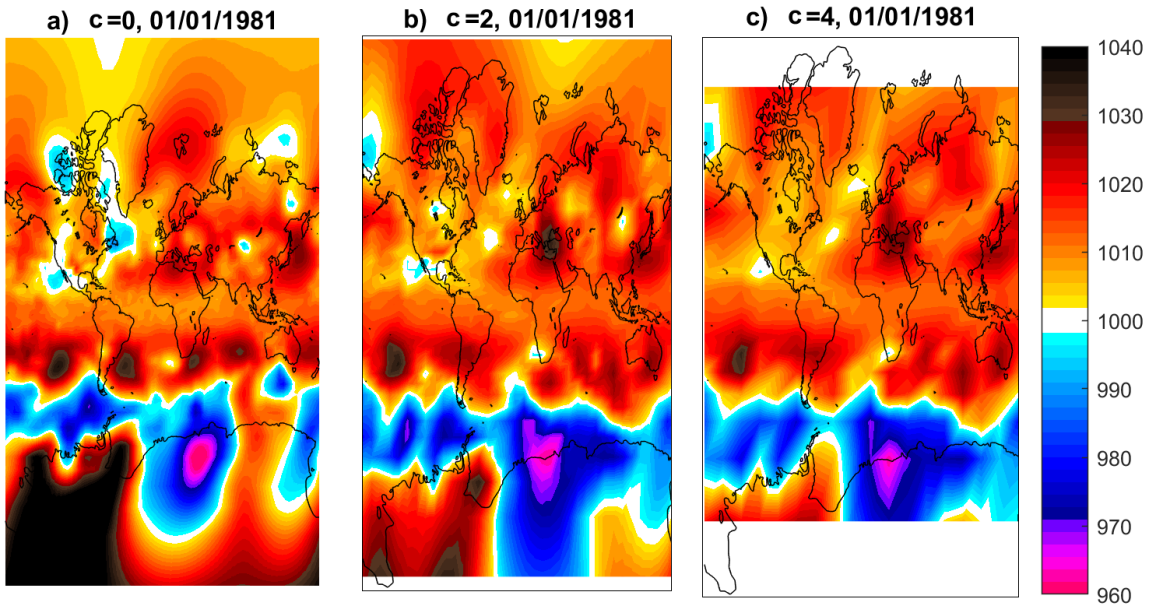
Supplementary Figure 3. a) Trajectories predicted using ESN on the Lorenz 1963 attractor for the variable x. The attractor is perturbed with Gaussian noise with variance $\epsilon = 0.1$. The target trajectory is shown in blue. 10 Trajectories obtained without moving average (green) show an earlier divergence compared to 10 trajectories where the moving average is performed with a window size of $w = 10dt$ (red). Panel (b) shows the evolution of the RMSE, averaged over the trajectories for the cases with $w = 10dt$ (red) and $w = 0$ (green).



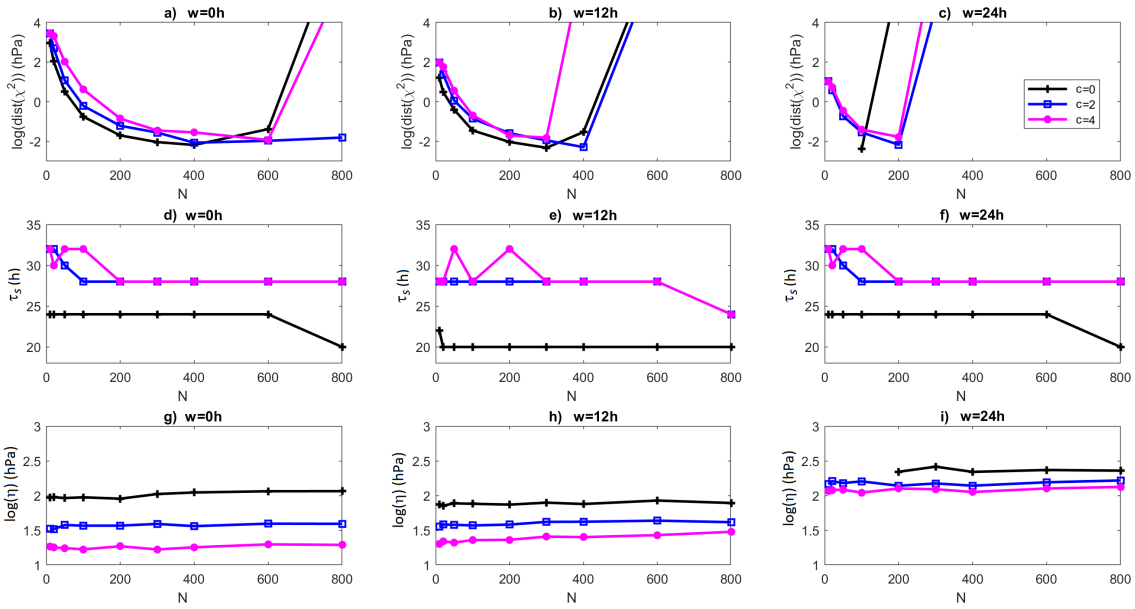
Supplementary Figure 4. Lorenz 1963 analysis for increasing noise intensity ϵ (x-axes), and number of neurons N (y-axes). The color-scale represents: the rejection rate ϕ of the invariant density test for the x variable (a-b); logarithm saturation time $log(\tau_{s=1})$ (c,d); logarithm of initial error $log(\eta)$ (e,f). All the values are average over 100 realizations. Panels (a,c,e) refers to results with averaging window $w = 10dt$, panels (b,d,f) to results with averaging window $w = 50dt$.



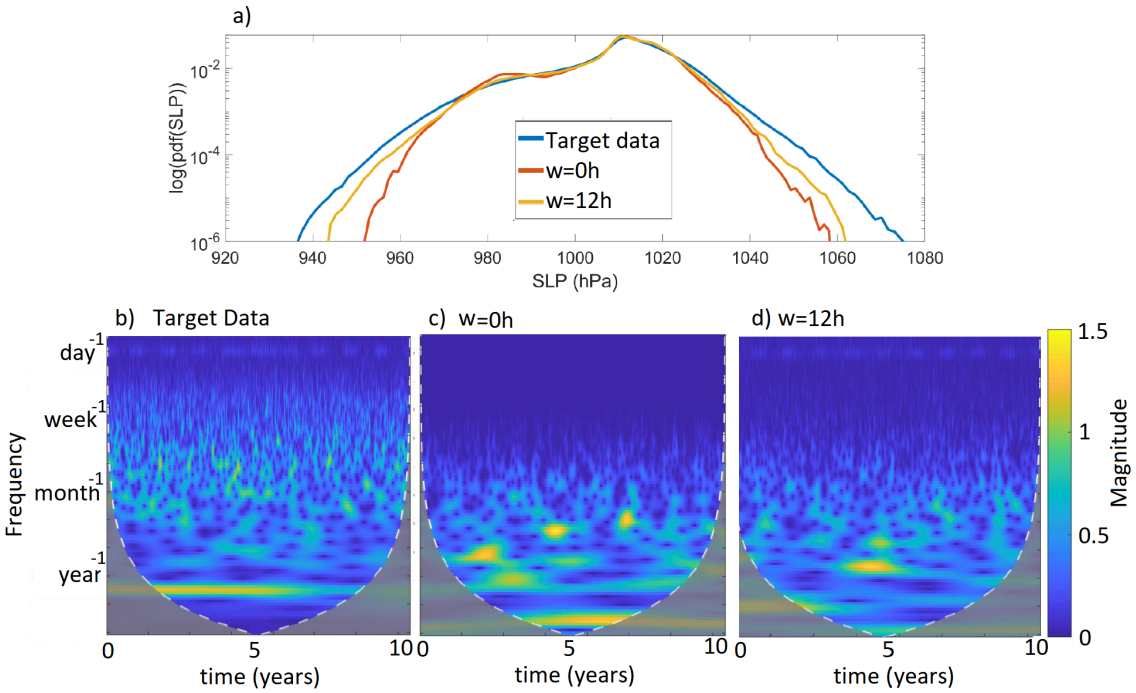
Supplementary Figure 5. a) Trajectory of the Pomeau-Manneville map with $a = 0.91$ and 5×10^5 iterations. b) Invariant density for the same trajectory shown in panel (a).



Supplementary Figure 6. Effects of spatial coarse grain NCEPv2 sea-level pressure (SLP) fields. a) original field for 01/01/1981, b) coarse grained field (factor $c = 2$, c) coarse grained field $c = 4$.



Supplementary Figure 7. 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; d-f) saturation time (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.



Supplementary Figure 8. 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).