**Dear Editor,**

**We have revised the manuscript by following the suggestion of the reviewers. We provide a new version of our study where figures have been improved, notation has been improved and the additional analyses proposed have been integrated. When changing figures and taking into account comments, we have been in the need to redo more realisations for our experiments. Of course, the new realizations included in some of the figures of this paper show statistically the same results, but they can be slightly visually different then the previous version. Furthermore, we have edited some of the answers initially proposed to the referees to reflect the changes in the manuscript. We believe that when reviewing the new version of the manuscript, the referees should find in the present answer letter a more detailed description of changes made in the manuscript than the answers previously published. We regret that reviewer 2 did not provide useful references or suggestions to improve the quality of the manuscript. A marked-up version of the manuscript showing the implemented changes is attached at the end of this answer letter.**
**Finally we believe that our editing efforts are reflected in the general improvement of the quality of the paper, which is now more readable and scientifically sound. In particular, new figure 10 shows a clear evidence of the improvement in forecast of atmospheric fields obtained with the filtering procedure.**

**Best Regards**
**Davide Faranda (on behalf of all the authors)**

***RC1 Anonymous Reviewer***

This manuscript explores the effectiveness of echo-state-networks for a hierarchy of problems. It explores 3 "toy" dynamical systems and then applies the methodology to a data driven weather prediction task. They evaluate both the equilibrium distribution (using a Xi-squared analysis) and initial value forecasts using root-mean-squared error based metrics. By these metrics, they claim that filtering the data before training an ESN generally improves these metrics in cases where the underlying dynamics are "intermittent" or show strong "coupling between timescales". For all the problems except for Lorenz 96 (L96), they pre-filter with moving averages, whereas for L96 they take advantage of the built-in scale separation between the large-scale and small-scale variables. Overall, I thought the results were interesting and relevant to geophysical problems which often feature intermittent and multiscale dynamics, but was not convinced that their claims were valid. See my comments below.

**We thank the reviewer for the appreciation of our work. In the new version of the manuscript we have fully addressed the recommendations. A detailed answer is provided below**

MAJOR COMMENTS

1. The quality of presentation should be improved
(a) In a few cases, the color schemes used were not intelligible to colorblind readers, which significantly hampered my ability to understand their results. There are many multi-panel figures, which are explained only briefly in the text.

**We have taken great care in changing the color scales for colorblind users and we are sorry the reviewer had trouble in understanding some of the results. Multi Panels figures have beeen discussed in more details in the text.**

(b) Notation is used inconsistently and unclearly in some places. Also, this paper introduces redundant notation. Vector and scalar quantities are not differentiated clearly.

**Following the suggestion of the reviewer (see also answers to technical comments), we have used a more compact and consistent notation. The main confusion probably originated by the misuse of u(t) in the formula of ESN. Indeed for the training we use the vectors x(t) of the dynamical systems, while the observable u(t) is a scalar quantity for a given time. Also in section 2.2 and 2.3 we have replaced u(t) with x(t) because the MA filter is operated on the signal and not on the observable. It is now clearly stated what observable is taken for the computation of the metrics used as diagnostic**

(c) The literature review in the introduction was incomplete in a few places. Also, for an article in a geophysical science journal, concepts like CNNs, RNNs, ESNs should all be clearly defined and differentiated from one another. The introduction sometimes incorrectly conflates these concepts.

**We have taken great care to differentiate CNN from RNN and ESN in the new version of the manuscript.**

(d) The conclusion contains many helpful motivations that could have helped guide me through the introduction and the methods sections.

**We have moved some of the key concepts presented in the conclusions, in the introduction section and rephrased a few sentences to make the purpose of the study clearer.**

2. Their ESNs appear to fail to meaningfully reproduce the time series of the Pomeau-Manneville (fig 5) or Lorenz 96 (fig 8) examples. As with any negative result, it is unclear whether some minor methodological improvement could fix it, so I am not sure what insights these examples provides. In particular, some authors have demonstrated substantially nearly optimal performance with data driven techniques for Lorenz 96 (Gagne, et. al. 2020) and with ESNs for similar Kuramoto-Shivashinksy model (Pathak, et. al 2018). Were the authors able to replicate the success of these previous studies?

**Indeed we are able to reproduce the previous results obtained with ESN for the models outlined by the reviewer. However, in this paper we decide to use the simplest possible ESN (i.e. not the tuned one which indeed provides better performances to the single cases) to perform sensitivity studies on noise level and coarse-graining, in an extended, comprehensive and parameters controlled way. As pointed out by the other referee, a deterministic ESN with smooth, continuous activation function cannot be expected to produce trajectories that look spiking/stochastic/rapidly changing. Most previous studies on ESNs were handling relatively smooth signals, and not such rapidly changing signals. Although it does not come as a surprise that utilizing the ESN on the time averaged dynamics and then adding a stochastic residual improves performance, the main insights is the intricate dependence of the ESN performance on the noise structure and the fact that, even for non-smooth signal, ESN with hyperbolic tanh functions can be used to study systems that have a multiscale dynamics. In the new version of the manuscript we have added these considerations to the discussion section of the articles.**

3. The Xi-squared testing procedure seems suspect. It mixes a parametric test (Xisquared) with a boot-strapping based test. Is there any support for this technique in the literature? It would be preferable to use a more well-known statistical test for this problem (e.g. Wilcoxon Rank sum, Kolmogorov-Smirnov).

**We are aware of the limitations of the procedure we decided to adopt, but after considering different strategies, including those suggested by the referee, we decided to still keep this approach. There are two aspects playing a role in this choice: why to use a chi-squared test and why to conduct a Monte Carlo experiment to determine the critical value.**

**The reasons for the choice of the chi-squared test reside mainly in the construction of the test statistics: this is built to compare the probability distribution, considering differences between the empirical probability density functions (pdf) over the entire domain, in practice the histograms. Under the null hypothesis that the two distributions are the same, their differences in the bins are i.i.d normal random variables, and the chi-squared statistics follows indeed a chi-square distribution with M degrees of freedom, with M the number of bins.**
**Other examples of tests that consider the shape of the entire distribution are the Anderson Darling and the Cramer-von-Mises tests. We did see an advantage in using the chi-squared because it is based not on a distance, but on an asymmetric divergence that gives more weight to the reference distribution: since we are not comparing two sample distributions, but a sample distribution and the true distribution of the dynamical system observable, we are willing to place more weight on the latter. These explanations have been added to the new version of the paper.**

**Procedures based on ranks such as the Wilcoxon Rank sum, on the other hand, test more specific null hypotheses. The Wilcoxon rank sum tests the null hypothesis of identical distribution against the specific alternative that one of the two distributions exhibits stochastic dominance (this degenerates to a test on the median in case of Gaussian homoskedastic variables); however, this test could end up not rejecting the null hypothesis in case of pdfs with sensibly different shape, but not clear stochastic dominance or shift in median. In the case of the KS, the test is indeed useful for testing if two distributions are identical in a more general way. However, in our experience the KS test tends, at large sample sizes, to over-reject the null hypothesis in presence of very small differences between the distributions.**

**Concerning the experiment design, we need to clarify that we did not use a bootstrap test, which relies on resampling. Instead, we adopt a Monte Carlo *simulation* approach to approximate the distribution of the test statistics, and then the critical rejection value. Each of the 10000 samples is generated under the null hypothesis, which is possible because we are considering simulated systems, for which we can obtain as many independent samples as we want.**

**We made this choice because, even simulating relatively long trajectories, we cannot observe the entire invariant distribution of the process, and therefore we cannot use the theoretical chi2(M) distribution of the test statistic. This situation would happen in any case, independently on the chosen test, because the problem resides in our inability to observe the invariant distribution. In other words, we run a simulation to obtain tabulated values of the distribution of the test statistics under H0 specific to our study, and we would do this for any adopted statistical test, as using the exact or asymptotic distribution (depending on the test) would make fixing the level of the test not sufficient to control the probability of Type 1 error.**

4. The sea-level pressure example was compelling.

**We thank the referee for the appreciation of the results for the sea-level pressure.**

SPECIFIC COMMENTS

Title: “Boosting performance” This is a quibble, but “boosting” has a rather specific meaning in the machine learning literature https://en.wikipedia.org/wiki/Boosting_(machine_learning). This could be misleading.

**Thank you for this remark, if possible, we have changed the wording “boosting” with “enhancing”**

L8: “with an optimal choice of spatial coarse grain and time filtering” With an optimal choise of spatial coarse-graining

**Corrected**

L20. Buchanan. How does this PhD dissertation relate to the previous assertion. Please be more specific.

**We have corrected this reference which corresponds to a journal article and not to a Phd dissertation as previously stated.**

L26. Gentine There are many other articles on parameterizations which should be mentioned e.g. (Brenowitz and Bretherton 2018, 2019; Yuval and O’Gorman 2020; Kransopalky 2005, 2013; Gettleman et. al 2020).

**After revising them, we agree that these references contain relevant information for the cited step and we have added them to the new version of the manuscript**

L27. This introduction should also mention (Rasp et. al 2020; Weyn et.al 2019) for the pure weather prediction problem

**We have added these references to the new version of the manuscript**

L40. “Recent examples include. . . convolutional neural ntworks, . . . ” C3 With the previous sentence in mind, this wording implies that convolutional neural networks are a type of RNN. I believe the references all used feed-forward architectures.

**Thank you. This have been corrected in the new version of the manuscript**

L65. “Previous results (Scher, 2018; Dueben and Bauer, 2018; Scher and Messori, 2019) suggest that RNN simulations” Again, I don’t think these papers all studied RNNs. At least some used feed-forward architectures.
**We have rephrased this in the new version of the manuscript**

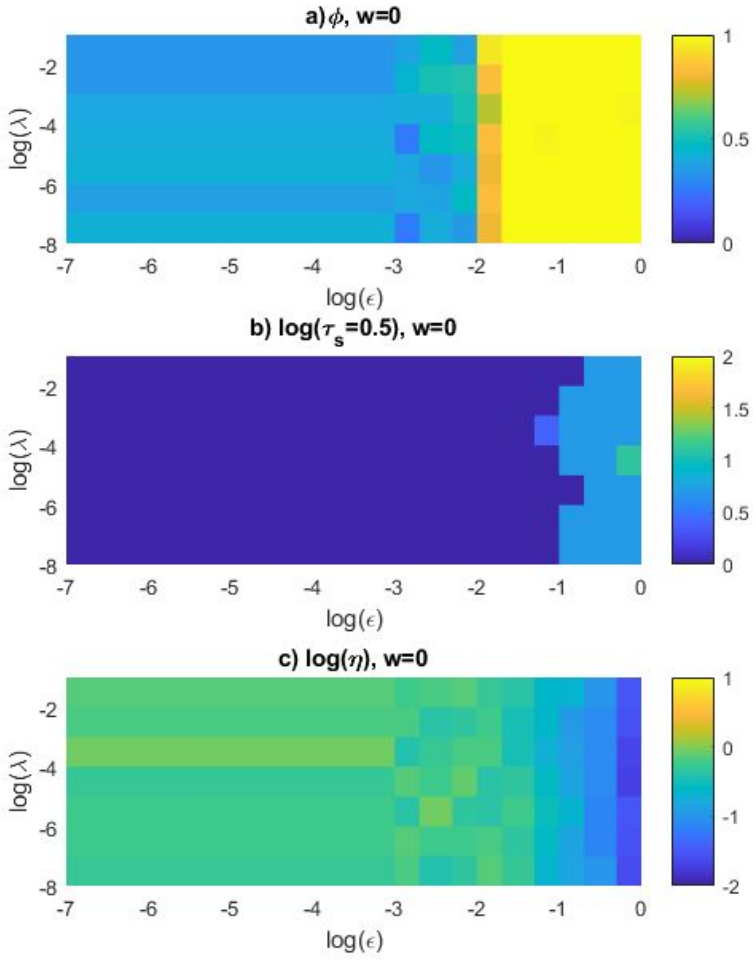
L73-90. Overall, this description does not clarify what ESNs are, and why they work outperform traditional RNNs for some problems (e.g. the vanishing gradients problem).

**Again, we have stated clearly the specificity of the ESN at the beginning of the methods section: “In our study we particularly use ESN, a particular case of RNN where the output and the input have the same dynamical**

**form. In the ESN approach, neuron layers are replaced by a sparsely connected network, with randomly assigned fixed weights. we harvest reservoir states via a nonlinear transform of the driving input and compute the output weights to create reservoir-to-output connections."**

L90: "We estimate w_out via a ridge regression with lambda=" How was this parameter chosen? ESNs are very sensitive to this parameters, and the optimal parameter may vary from problem to problem. This could potentially explain the poor performance on the L96 and Pomeau-Manneville examples below.

**Ridge minimizes the residual sum of squares plus a shrinkage penalty of lambda multiplied by the sum of squares of the coefficients. As lambda increases, the coefficients approach zero. The coefficients are unregularized when lambda is zero. We have tested the dependence on the ridge regression parameter for the Pomeau Manneville map and found that the dependence is small.**




**The figure above shows the dependence of \phi, log(\tau_s=1) and log(\eta) from the value of the ridge parameter. No particular dependence is shown, so we stick to the small value \lambda=10^(-8) for the computations shown in this article.**

L98. “Let, U be . . . ” For readibility, try to re-use previously introduced notation to avoid introducing too many new symbols. For instance “v” is the same as “r” in eq 1-4. Are theses tests univariate? The equations are multivariate.

**We believe that the confusion originated by our wrong use of the notation. Now, for each system we have specified the observable used. The tests are all univariate, for the Lorenz 1963 we consider the variable x only. For the Lorenz 1996 we consider one of the variables, since they are all dynamically and statistically equivalent. For the SLP, we consider the spatial average as observables for the test.**

L120: “we observed excessive rejection rates” How do you quantify this?

**We underline that the sentence is actually: "we would observe excessive rejection rates".**
**Here we are underlining that, due to intrinsic limitations, we can construct a chi-squared test, but not use the standard critical values for the distribution of the test statistic, which would produce excessive rejection rates. Therefore, we construct the test statistic in the usual way, but use Monte Carlo simulation to obtain the distribution of the test statistic under the null hypothesis.**

L121: “we use 10000 samples” What is “a sample”. Is it a single time step of r(t) above (e.g. a K-dimensional vector)? Is it the number of timesteps or is it the number of timesteps times K? This would be clearer if described in terms of the notation used in Eqs 1-4.

**We have specified in the new version of the manuscript that a sample is a series of a univariate [ as specified in the answer for L98] test observables, and we consider 10000 samples extracted from the total, longer time series.**

L135: This formula seems odd. I would normally define predictability by computing RMSE versus the truth for a single timestep. In this case they compute the average MSE accumulated over several timesteps. Also, this formula only makes sense for scalar u and v, but I thought we are in the vector setting?

**We thank the referee for the comment. Unlike our previous response in the discussion, we acknowledge that Indeed the equation was incorrect, after initially using different error measures, we indeed compute the absolute prediction error (APE), and not the RMSE, for each time step.**

Section 2.2: It is unclear why this moving average is described here. It would be clearer if the introduction had introduced a broad outline of the paper.

**We have followed the suggestion of the reviewer and introduced the moving average in the introduction**

L248: “Performances are again better when using the exact formula (Figure 4b,e,h) than using the residuals $\delta u$ (Figure 4c,f,i).” It would be helpful to refer to Eq 11 here.

**Thank you, we have added Eq 11 there**

L250: “ESN simulations do not reproduce the intermittency in the average of the target signal. They only show some second order intermittency in the fluctuations.” Is “the average” supposed to mean “the moving average” rather than “time average”? Is “second order intermittency?”. Is this a formal concept?

**What we mean here is that during the intermittent phases, the PM dynamics oscillate in the range (0.2 1) with an average of about 0.6. In the non-intermittent phases, the PM dynamics is stuck near 0. Therefore, the intermittency is on average (shift from 0.6 to 0) and in variance. This explanation has been added to the new version of the manuscript.**

L270. Forward Euler time steppers are notoriously inaccurate. Why not use a more advanced time stepper (e.g. Runge Kutta) for better accuracy? There are many convenient software packages for integrating ODEs with better schemes (e.g. ode45 in MatLab). What is N? It must be network size, but given all the notational changes it is hard to be sure.

**We remind that the idea is here to have exemples close to the atmospheric or climate data: when considering daily or 6 hourly data, as commonly done in climate sciences and analyses, we hardly are in the case of a smooth RK time stepper. We therefore stick to the Euler method for similarity with the actual climate data. This has been added to the text.**

L331: “We show the results using the residuals (Eq. 9)” Why not show the results with the “exact method” (Eq. 11)? It seems the earlier results implied this technique was more effective.

**Unfortunately the “exact method” cannot be used for the SLP NCEP data. Indeed this dynamics has a spatial component that the “exact” method cannot take into account the spatial component. This is now specified in the paper**

Figure 10 b-d. These panels all look different. I don’t see much reason to prefer panel d to c. Could the authors present a more convincing visualization for the claimed improvement of the moving average filter? Maybe a single power-spectra plot would be more succinct, especially since the author’s don’t comment on the timing of the high-frequency vs low-frequency results.

**The wavelet methodology is a more sophisticated representation of a spectrum. We have replaced wavelet spectra with conventional spectra (now Figure 10b). For consistency, figure 10a is also replaced with the pdf for the observable u , used in figure 9 and in figure 10b. This means that all diagnostic on SLP are now computed using u(t)=<SLP(t)>_{lon,lat}, i.e. the time series of the spatial average of SLP**

L373. “For the Lorenz 1996 mode, we did not apply a moving average filter to the data,. . . ” It would have been nice to see this motivation described in Section 3.

**We have added this motivation in Section 3**

TECHNICAL CORRECTIONS

L73. ‘Reservoir compution"’ There is a missing quote.

L74. “The principle of Reservoir computing” Does “Reservoir” need to be capitalized here? If so, I would expect “computing” to be capitalized as well. “reservoir” is not always capitalized in this manuscript. L76. “In our study ESNs are implemented”
L77. "The code is given in the appendix
L97: “to this purpose” –> “for this purpose”
L239: “we find the best match. . . are obtained for w=3” Correct “are” to “is”. 249: “Figure 5a)” Remove the parenthesis
Line275. “Figure 6.b,d)” This should read “Figure 6 b,d”.
Figures should be referred to with a consistent convention.
L288. “distance T”. C6 Do the authors mean $\Sigma$ ? T is the length of the time series. Figure 8: The text in this graphic is fuzzy. Please save at a higher resolution.
Figure 2a: This plot has too many curves. Red-green is bad for colorblind readers. It is hard to see the author’s point.
Figure 3, 4: These colorscales are not legible for colorblind readers. I could not interpret these figures and relied on the author’s textual description of the results. I suggesting using “viridis” or another sequential colorbar.

**Thank you, technical corrections have been implemented. Colorscales replaced as demanded for colorblind readers.**

REFERENCES

Pathak, J., Hunt, B., Girvan, M., Lu, Z., & Ott, E. (2018). Model-Free Prediction of Large Spatiotemporally Chaotic Systems from Data: A Reservoir Computing Approach. Physical Review Letters, 120(2). https://doi.org/10.1103/physrevlett.120.024102

Gagne, D. J., II, Christensen, H. M., Subramanian, A. C., & Monahan, A. H. (2020). Machine Learning for Stochastic Parameterization: Generative Adversarial Networks in the Lorenz ’96 Model. Journal of Advances in Modeling Earth Systems, 12(3). https://doi.org/10.1029/2019ms001896

Yuval, J. & O’Gorman, P. A. Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. Nat. Commun. 11, 3295 (2020)

Brenowitz, N. D. & Bretherton, C. S. Spatially Extended Tests of a Neural Network Parametrization Trained by CoarseâA˘ Rgraining. ˘ J. Adv. Model. Earth Syst. (2019) doi:10.1029/2019MS001711

Krasnopolsky, V. M., Fox-Rabinovitz, M. S. & Chalikov, D. V. New Approach to Calculation of Atmospheric Model Physics: Accurate and Fast Neural Network Emulation of Longwave Radiation in a Climate Model. Mon. Weather Rev. 133, 1370–1383 (2005)

Krasnopolsky, V. M., Fox-Rabinovitz, M. S. & Belochitski, A. A. Using Ensemble of Neural Networks to Learn Stochastic Convection Parameterizations for Climate and Numerical Weather Prediction Models from Data Simulated by a Cloud Resolving Model. Advances in Artificial Neural Systems 2013, e485913 (2013)

Brenowitz, N. D. & Bretherton, C. S. Prognostic Validation of a Neural Network Unified Physics Parameterization. Geophys. Res. Lett. 17, 2493 (2018)

O'Gorman, P. A. & Dwyer, J. G. Using Machine Learning to Parameterize Moist Convection: Potential for Modeling of Climate, Climate Change, and Extreme Events. J. Adv. Model. Earth Syst. 10, 2548–2563 (2018)

Gettleman et. al. Machine Learning the Warm Rain Process. Rasp, S. et al. WeatherBench: A benchmark dataset for dataâA˘ Rdriven weather fore- ˇ casting. J. Adv. Model. Earth Syst. (2020) doi:10.1029/2020MS002203

Weyn, J. A., Durran, D. R. & Caruana, R. Can Machines Learn to Predict Weather? Using Deep Learning to Predict Gridded 500âA˘ RhPa Geopotential Height From Historical ˇ Weather Data. J. Adv. Model. Earth Syst. 11, 2680–2693 (2019)

**We thank the reviewer for the references, which have been properly added to the new version of the manuscript.**

***RC2 - Anonymous Reviewer***

The authors are utilizing Echo State Networks to predict filtered dynamics in the perturbed Lorenz 1963 equations, the Pomeau-Manneville 89 intermittent map, and the Lorenz 1996 equations. A moving average filter is utilized for scale separation in time. The filtered dynamics are smoother and easier to predict. A residual term is added, either sampled from the training data, or based on an analytic formula derived from the moving average filter. Assuming that the filter width is smaller than the associated large timescales of the processes involved, the large scale processes can be successfully predicted. The authors claim that modeling only the spatially coarse grained and time averaged state can boost performance of ESN. However, the generalization of this argument to more realistic systems is not sufficiently supported by the results, as elaborated in the comment section below. The idea of utilizing a moving average filter for noise reduction and scale separation, or spatial coarse graining is known. I am not sure that the novelty of the paper to apply ESNs to (spatially/time) filtered dynamics, is enough to guarantee publication in the journal. The effect of the unmodeled dynamics (the information lost during filtering) is not taken into account in the model. In most interesting applications, the effect of the unmodeled modes is the problem, and a field of study by itself (closure models in turbulence, small scale models in weather etc.)

**We respect the opinion of the reviewer on our work, but we feel that the motivations provided here and in the following comments for rejection are not supported and sometimes do not contain any element that could help us to improve the manuscript. For example, the reviewer states many times that our results are "known" or "existing in the literature" or "hardly surprising" but not a single reference to previous works which should contain our results is provided. We are therefore unable to assess which part of our work could be not original or to provide an adequate rebuttal to state why our work is instead original. Furthermore, the reviewer says that our results are "not sufficiently supported by the results", but it is not said in which way this is the case. We stress that, for all cases presented, we have used at least three statistical metrics to assess performances, and scanned a large range of coarse graining and noise intensities, as well as performed several realisations of our systems. If this is not enough to warrant publication, then we would like to know why our suggested metrics are not sufficient to support our conclusion. Besides these remarks on the structure of the proposed comments, we will do our best to answer the reviewer's comments and thus to improve the manuscript.**

1 Comments

1. In the three-dimensional Lorenz system, it is logical that the moving average filter produces better results. By construction, noise is added to the system. It does not come as a surprise that the ESN predicting the filtered dynamics (which are smoother) and augmented with the random residual terms, shows superior performance. However, there is no complex multiscale effect taking place, as the whole state information is given to the system (no hidden state, at least nothing is mentioned in the text about it). Moreover, as a reference time-scale, the Lyapunov time of the deterministic system is used, although the system is augmented with noise, which means that the effective Lyapunov time is in essence much shorter, as stochasticity accelerates the divergence of nearby trajectories. In any case, it is important to be critical about the conclusions drawn from this case.

**We agree with the reviewer that "it is logical that the moving average filter produces better results" but what we would like to show is that there is a dependence on the noise intensity on the quality of the results obtained and particularly that the moving average filters is very useful for intermediate noise intensity, namely when the stochastic component starts to affect the deterministic dynamics, but not at the same order of magnitude. We disagree with the statement that " the effective Lyapunov time is in essence much shorter, as stochasticity accelerates the divergence " as this is also dependent on the noise level. The most interesting performances for the filtered ESN are obtained when the noise is yet 2-3 order of magnitude smaller than the typical scales of the deterministic component, and we expect the perturbation on the lyapunov exponents to be of these orders as well. In the conclusion section we have added: Most previous studies on ESNs were handling relatively smooth signals, and not such rapidly changing signals. Although it does not come as a surprise that utilizing the ESN on the time averaged dynamics and then adding a stochastic residual improves performance, the main insights is the intricate dependence of the ESN performance on the noise structure and the fact that, even for non-smooth signal, ESN with hyperbolic tanh functions can be used to study systems that have a multiscale dynamics. In the new version of the manuscript we have added these considerations to the discussion section of the articles.**

2. In the Pomeau-Manneville intermittent map, it is not a surprise that the ESN cannot capture the dynamics, as they are changing very rapidly, even visually they look completely stochastic. A deterministic ESN with tanh (smooth, continuous) activation function cannot be expected to produce trajectories that look spiking/stochastic/rapidly changing. Most previous studies on ESNs were handling relatively smooth signals, and not such rapidly changing signals. At least the nature of the signal has to be taken into account in the selection of the activation function of the reservoir. Thus, it does not come as a surprise that utilizing the ESN on the time averaged dynamics and then adding a stochastic residual improves performance. As expected, the plain ESN diverges, as demonstrated also in previous studies with such non-smooth signals.

**We thank the reviewer for the comment. Indeed this can be a good explanation of our results. The reviewer says again that "it is not a surprise" or "demonstrated also in previous studies". However, no references are provided for us to improve the quality of the manuscript or to give credits to those who have already analysed this problem on another angle. We would be more than happy to include and discuss those references in the manuscript. Furthermore, the reviewer makes a confusion between a visual analysis and what signals truly are. The PM system is piecewise continuous & differentiable, and is hence "relatively" smooth from the mathematical point of view.**

3. In the Lorenz 96 system, as demonstrated in Figure 8, the method fails to capture the long-term climate, as the dynamics predicted by the ESN are clearly different from the groundtruth.

**We only partially agree with the reviewer about the results obtained for Lorenz96. Although the detailed dynamics does look different from that of the original system, there are a few things correctly captured by the ESN, namely the quasiperiodic spatio-temporal oscillations and the fact that ESN produces non-divergent dynamics.**

4. In the sea-level pressure, the moving average filter ESN does not achieve any significant improvement based on the results in Figure 9.

**Here, we would like to gently disagree with the referee comment. The ESN with filter does produce significant improvements, in terms of all the metrics considered, and as noted by the other reviewer.**

5. In the abstract, the authors claim that "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". This is shown for Echo State Networks in the document, but not in general for Recurrent Neural Networks. The argument has to be relaxed to take into account only ESNs, or a relevant reference for other RNN architectures should be given.

**We agree with this comment of the referee, in the new version of the manuscript we have clearly restricted our attention to ESNs.**

6. There is a contradiction in the text, in page 3, the authors state that "We aim at understanding this sensitivity in a deeper way, while assessing the possibility to reduce its impact on prediction through simple noise reduction methods", although one sentence before, they claim that they choose the ESN framework for "...its ability to forecast chaotic time series and its stability to noise". These sentences are contradicting each other. Later in the text, the authors state "Since Echo State Networks are known to be sensitive to noise (see e.g. [34]), ...".

**What we mean here is that ESN are less sensitive to noise than other techniques, but also that our goal is precisely to evaluate such sensitivity and the improvement coming from noise reduction techniques.**

7. The analysis of the performance of the proposed method based on different parameters e.g. intermittency of dynamics/degree of coarse graining, etc. is interesting. However, this is not adequate to warrant publication.

**We are delighted to see that the reviewer admits that our results are interesting. This encourages us to pursue their publication.**

# ~~Boosting performance in~~ Enhancing geophysical flow machine learning ~~of geophysical flows~~ performance via scale separation

Davide Faranda[1,2,3], Mathieu Vrac[1], Pascal Yiou[1], Flavio Maria Emanuele Pons[1], Adnane Hamid[1], Giulia Carella[1], Cedric Ngoungue Langue[1], Soulivanh Thao[1], and Valerie Gautard[4]

[1]Laboratoire des Sciences du Climat et de l'Environnement, CE Saclay l'Orme des Merisiers, UMR 8212 CEA-CNRS-UVSQ, Université Paris-Saclay & IPSL, 91191 Gif-sur-Yvette, France.
[2]London Mathematical Laboratory, 8 Margravine Gardens, London, W68RH, UK.
[3]LMD/IPSL, Ecole Normale Superieure, PSL research University, Paris, France
[4]DRF/IRFU/DEDIP//LILAS Departement d'Electronique des Detecteurs et d'Informatique pour la Physique, CE Saclay l'Orme des Merisiers, 91191 Gif-sur-Yvette, France.

**Correspondence:** Davide Faranda (davide.faranda@lsce.ipsl)

**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~~ short- and long-term forecasts of sea-level pressure data can be obtained with an optimal choice of spatial ~~coarse grain~~ coarse-graining and time filtering.



## 1 Introduction

The advent of high-performance computing has paved the way for advanced analyses of high-dimensional datasets (Jordan and Mitchell, 2015; LeCun et al., 2015). Those successes have naturally raised the question of whether it is possible to learn the behavior of a dynamical system without resolving or even without knowing the underlying evolution equations. Such an interest is motivated on one side by the fact that many complex systems still miss a universally accepted state equation — e.g. brain dynamics (Bassett and Sporns, 2017), macro-economical and financial systems (Quinlan et al., 2019) — and, on the other, by the need of reducing the complexity of the dynamical evolution for the systems of which the underlying equations are known — e.g. on geophysical and turbulent flows (Wang et al., 2017). Evolution equations are difficult to solve for large systems such as the geophysical flows, so that approximations and parameterizations are needed for meteorological and climatological

applications (Buchanan, 2019). These difficulties are enhanced by those encountered in the modelling of phase transitions that lead to cloud formation and convection, which are major sources of uncertainty in climate modelling (Bony et al., 2015). Machine Learning techniques capable of learning geophysical flows dynamics would help improve those approximations and avoid running costly simulations resolving explicitly all spatial/temporal scales.

Recently, several efforts have been made to apply machine learning to the prediction of geophysical data (Wu et al., 2018), to learn parameterizations of subgrid processes in climate models ~~(Krasnopolsky and Fox-Rabinovitz, 2006; Rasp et al., 2018; Gentine et al.,~~ (Krasnopolsky et al., 2005; Krasnopolsky and Fox-Rabinovitz, 2006; Rasp et al., 2018; Gentine et al., 2018; Brenowitz and Bretherton, 20 , to the forecasting ~~(Liu et al., 2015; Grover et al., 2015; Haupt et al., 2018)~~ (Liu et al., 2015; Grover et al., 2015; Haupt et al., 2018; Weyn and nowcasting (i.e. extremely short-term forecasting) of weather variables (Xingjian et al., 2015; Shi et al., 2017; Sprenger et al., 2017), and to quantify the uncertainty of deterministic weather prediction (Scher and Messori, 2018). One of the greatest ~~challenge~~ challenges is to replace equations of climate models with neural ~~network~~ networks capable to produce reliable ~~long and short term forecast~~ long- and short-term forecasts of meteorological variables. A first great step in this direction was the use of Echo State Networks (ESN, (Jaeger, 2001)), a particular case of Recurrent Neural Networks (RNN), to forecast the behavior of chaotic systems, such as the Lorenz (1963) and the Kuramoto-Sivashinsky dynamics (Hyman and Nicolaenko, 1986). It was shown that ~~RNN~~ ESN predictions of both systems attain performances comparable to those obtained with the ~~real~~ exact equations (Pathak et al., 2017, 2018). Good ~~performance of regularized RNN~~ performances were obtained adopting regularized ESN in the short-term prediction of multidimensional chaotic time series~~was obtained~~, both from simulated and real data (Xu et al., 2018). This success motivated several follow-up studies with a focus on meteorological and climate data. These are based on the idea of feeding various statistical learning algorithms with data issued from dynamical systems of different complexity, in order to study short-term predictability and capability of machine learning to reproduce long-term ~~capabilities of RNN in producing a surrogate dynamics of~~ features of the input data dynamics. Recent examples include equation-informed moment-matching for the Lorenz96 model (Lorenz, 1996; Schneider et al., 2017), multi-layer perceptrons to reanalysis data (Scher, 2018), or convolutional neural networks to simplified climate simulation models (Dueben and Bauer, 2018; Scher and Messori, 2019). All these learning algorithms were capable to provide some short-term predictability, but failed ~~in~~ at obtaining a long-term behavior coherent with the input data.

~~In this article we specifically focus on how to improve the performance of ESN in simulating long trajectories of large-scale climate fields. With respect to~~ The motivation for this study came from the evidence that a straightforward application of ESN to high dimensional geophysical data does not yield to the same result quality obtained by Pathak et al. (2018) for the Lorenz 1963 and the ~~results presented in Pathak et al. (2018), we aim at going beyond the predictability horizon and investigate the ability of machine learning algorithms in shadowing the dynamics of observed data. Such applications would avoid the use of general circulation models based on primitive equations to reproduce the evolution of a subset of variables and therefore obtain surrogates dynamics of existing datasets with little computational power. Previous results (Scher, 2018; Dueben and Bauer, 2018; Scher and suggest that RNN~~ Kuramoto-Sivashinsky models. Here we will investigate the causes for this behavior. Indeed, previous results (Scher, 2018; Dueben and Bauer, 2018; Scher and Messori, 2019) suggest that simulations of large-scale climate fields through deep learning algorithms are not as straightforward as those of the chaotic systems considered by ~~(Pathak et al., 2018)~~

Pathak et al. (2018). We identify two main mechanisms responsible for these limitations: (i) the non-trivial interactions with small-scale motions carrying energy at large scale and (ii) the intermittent nature of the dynamics. Intermittency triggers large fluctuations of observables of the motion in time and space (Schertzer et al., 1997) and can result in non-smooth trajectories within the flow, leading to local unpredictability and increasing the number of degrees of freedom needed to describe the dynamics (Paladin and Vulpiani, 1987).

By applying ESN to multiscale and intermittent systems, we investigate how scale separation improves ESN predictions. Our goal is to reproduce a surrogate of the large-scale dynamics of global sea-level pressure fields, a proxy of the atmospheric circulation. We begin by analysing three different dynamical systems: we simulate the effects of small scales by artificially introducing small-scale dynamics in the Lorenz 1963 equations (Lorenz, 1963) via additive noise. We investigate the Pomeau-Manneville equations (Manneville, 1980) stochastically perturbed with additive noise to have an example of intermittent behavior. We then analyse the performance of ESN in the Lorenz 1996 system (Lorenz, 1996). ~~This system dynamics~~ The dynamics of this system is meant to mimic that of the atmospheric circulation~~and feature~~ , featuring both large-scale and small-scale variables with an intermittent behavior. For all of those systems, as well as for the sea-level pressure data, we show how the performance of ESN in predicting the behavior of the system deteriorates rapidly when small-scale dynamics feedback to large scale is important. The idea of using moving average for scale separation is already established for meteorological variables (Eskridge et al., 1997). We choose the ESN framework following the results of Pathak et al. (2017, 2018), and an established literature about its ability to forecast chaotic time series and its stability to noise. For example, Shi and Han (2007); Li et al. (2012) analyse and compare the predictive performance of simple and improved ESN on simulated and observed one-dimensional chaotic time series. We aim at understanding this sensitivity in a deeper way, while assessing the possibility to reduce its impact on prediction through simple noise reduction methods.

The remaining of this article is organised as follows: ~~first~~in section 2, we give an overview of the ESN method ~~and provide the description of the systems used . Then,~~ (2.1), then we introduce the metrics used to evaluate ESN performance (2.2) and introduce the moving average filter used to improve ESN performance (2.3). Results (Section 3) are organised presenting the results by the system analysed. First we show the results for the perturbed Lorenz 1963 equations, then for the Pomeau-Manneville intermittent map, and for the Lorenz 1996 equations. Finally We discuss the improvement in short-term prediction and the long-term attractor reconstruction obtained with the moving average filter. We conclude by testing these ideas on atmospheric circulation data.

## 2 Methods

Reservoir computing ~~"~~ is a variant of recurrent neural networks (RNN) in which the input signal is connected to a fixed~~and random~~ , randomly assigned dynamical system called reservoir (Hinaut, 2013). The principle of Reservoir computing first consists in projecting ~~first~~ the input signal to a ~~space of high dimension~~ high-dimensional space in order to obtain a non-linear representation of the signal; and then ~~perform~~ in performing a new projection ~~(linear regression or ridge regression)~~ between the high-dimensional space and the output units, usually via linear regression or ridge regression. In our study~~ESN are implemented~~

~~as follows.~~ , we use ESN, a particular case of RNN where the output and the input have the same dynamical form. In an ESN, neuron layers are replaced by a sparsely connected network (the reservoir), with randomly assigned fixed weights. We harvest reservoir states via a nonlinear transform of the driving input and compute the output weights to create reservoir-to-output connections. The code is given ~~the in appendix~~in appendix, and it shows the parameters used for the computations. ~~Let $u(t)$~~

We now briefly describe the ESN implementation. Vectors will be denoted in bold and matrices in upper case. Let $\boldsymbol{x}(t)$ be the $K$-dimensional observable consisting of $t = 1, 2 \ldots, T$ time iterations, originating from a dynamical system~~and $r(t)$~~ , and $\boldsymbol{r}(t)$ be the $N$-dimensional reservoir state, then:

$$r\boldsymbol{r}(t + dt) = \tanh(Wr W\boldsymbol{r}(t) + W_{in} u\boldsymbol{x}(t)), \tag{1}$$

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

$$W_{out} r\boldsymbol{r}(t + dt) = v\boldsymbol{y}(t + dt) \tag{2}$$

where ~~$v(t + dt)$~~ $\boldsymbol{y}(t + dt)$ is the ESN prediction, $W_{out}$ with dimensions $K \times N$, is the weight matrix of the connections between the reservoir neurons and the output layer. We estimate $W_{out}$ via a ridge regression (Hastie et al., 2015):

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

with $\lambda = 10^{-8}$. Note that we have investigated different values of $\lambda$ spanning $10^{-8} < \lambda < 10^{-2}$ and found no sensitive differences in the performance of ESN. In the prediction phase we have a recurrent relationship:

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

### 2.1 ESN performance indicators

In this paper, we use three different indicators of performance of the ESN: a statistical distributional test to measure how the distributions of observables derived from ESN match those of the target data, a predictability horizon test and the initial forecast error. They are described below.

#### 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 $\chi^2$ test (Cochran, 1952), designed as follows. Let $U$ be a system observable, linked to the orginal variables of the

systems via a function $\zeta$ such that $u(t) = \zeta(\boldsymbol{x}(t))$ with support $R_U$ and probability density function $f_U(u)$, and let $u(t)$ be a sample trajectory from $U$. Note that $u(t)$ does not correspond to $x(t)$, it is constructed using the observable output of the dynamical system. 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 (Eckmann and Ruelle, 1985; Young, 2002). 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:

$$H_0: \quad f_U(u) = g_V(v) \quad \text{for every } u \in R_U$$
$$H_1: \quad f_U(u) \neq g_V(v) \quad \text{for any } u \in R_U$$

Under $H_0$, $\hat{f}_U(u)$ is the expected value for $\hat{g}_V(v)$, which implies that observed differences $(\hat{g}_V(v) - \hat{f}_U(u))$ 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

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

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 $\alpha$, the observed value of the test statistics $\Sigma$ is compared to the critical value corresponding to the $1-\alpha$ quantile of the chi-square distribution, $\Sigma_c = \chi^2_{1-\alpha}(M)$: if $\Sigma > \Sigma_c$, the null hypothesis must be rejected in favour of the specified alternative.

In our setup, we encounter two limitations in using the standard $\chi^2$ test. First, problems may arise when $\hat{f}_U(u)$, i.e. if the sample distribution does not span the entire 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 would 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 ~~use 10000 samples using the system equations~~ generate $10^5$ samples $u(t) = \zeta(x(t))$ under the null hypothesis, and we compute the test statistic for each one according to Eq. (5). Then, we use the $(1-\alpha)$ quantile of the empirical distribution of $\Sigma$ — instead of the theoretical $\chi^2(M)$ — to determine the critical threshold $\Sigma_c$. As a last remark, we notice that we are making inference in repeated tests setting, as the performance of the ESN is tested ~~10000~~ $10^5$ times. Performing a high number of independent tests at a chosen level $\alpha$ increases the observed rejection rate: in fact, even if the samples are drawn under $H_0$, extreme events become more likely, resulting in an increased probability to erroneously reject the null hypothesis. To avoid this problem, we apply the Bonferroni correction (Bonferroni, 1936), testing each one of the ~~$m = 10000$~~ $m = 10^5$ available samples at the level $\alpha' = \frac{\alpha}{m}$, with $\alpha = 0.05$. Averaging the test results over several sample pairs $u(t)$, $v(t)$ we obtain a rejection rate $0 < \phi < 1$ that we use to measure

the adherence of a ESN trajectory $v(t)$ to trajectories obtained via the equations. If $\phi = 0$, almost all the ESN trajectories can shadow original trajectories, if $\phi = 1$ none of the ESN trajectories resemble those of the systems of equations.

**Predictability Horizon**

As a measure of the predictability horizon of the ESN forecast compared to the equations, we use the ~~root mean square error (RMSE):~~ absolute prediction error (APE):

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

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

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

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

We interpret $\tau_s$ as a natural measure of the Lyapunov time $\vartheta$, namely the time it takes for an ensemble of trajectories of a dynamical system to diverge (Faranda et al., 2012; Panichi and Turchetti, 2018).

**Initial Forecast Error**

The initial error is given by ~~$\eta = RMSE(t=1)$~~ $\eta = APE(t=1)$, for the first time step after the initial condition at $t = 0$. We expect $\eta$ to reduce as the training time increases. ~~In this phase, the the smaller the initial error will be.~~

## 2.2 Moving average filter

Equipped with these indicators, we analyze two sets of simulations performed with and without smoothing, which was implemented using a moving average filter. The moving average operation is the integral of $u(t)$ between $t$ and $t-w$, where $w$ is the window size of the moving average. The simple moving average filter can be seen as a nonparametric time series smoother ~~(see e.g. (Brockwell and Davis, 2016), chapter 1.5)~~ (see e.g. Brockwell and Davis, 2016, 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-cycle components, if present. Moving averaging consists, in practice, ~~of replacing the observation $u(t)$~~ in replacing the trajectory $\boldsymbol{x}(t)$ by a value ~~$u^{(f)}(t)$~~ $\boldsymbol{x}^{(f)}(t)$, obtained by averaging the previous $w$ observations. If the time dimension is discrete (like in the Pomeau-Manneville system) it is defined as:

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

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

$$\underline{u}\boldsymbol{x}^{(f)}(t) = \frac{1}{w}\int_{\underline{t-w}}^{t} \underline{u}_{t}^{t+w}\boldsymbol{x}(\varsigma)d\varsigma. \tag{8}$$

We can define the residuals as:

$$\delta\underline{u}\boldsymbol{x}(t) = \underline{u}\boldsymbol{x}^{(f)}(t) - \underline{u} - \boldsymbol{x}(t). \tag{9}$$

In practice, the computation always refers to the discrete time case, as continuous time systems are also sampled at finite time steps. Since Echo State Networks are known to be sensitive to noise ~~(see e.g. Shi and Han (2007))~~(see e.g. Shi and Han, 2007) , 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 out the noise but smaller than the characteristic time $\tau$ of the large-scale fluctuations of the system. For chaotic systems, $\tau$ can be derived knowing the rate of exponential divergence of the trajectories, a quantity linked to the Lyapunov exponents (Wolf et al., 1985), and $\tau$ is known as the Lyapunov time.

We also remark that we can express explicitly the original ~~variable $u(t)$~~ variables $\boldsymbol{x}(t)$ as a function of the filtered ~~variable $u^{(f)}(t)$~~ variables $\boldsymbol{x}^{(f)}(t)$ as:

$$\underline{u}\boldsymbol{x}(t) = w(\underline{u}\boldsymbol{x}^{(f)}(t) - \underline{u} - \boldsymbol{x}^{(f)}(t-1)) + \underline{u}\boldsymbol{x}(t-w). \tag{10}$$

~~we~~ We will test this formula for stochastically perturbed systems to evaluate the error introduced by the use of residuals ~~$\delta u$~~$\delta\boldsymbol{x}$.

### 2.3 Testing ESN on filtered dynamics

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

1. Simulate the reference trajectory ~~$u(t)$~~ $\boldsymbol{x}(t)$ using the equations of the dynamical systems, ~~where $u(t)$ has been standardized~~ and standardize $\boldsymbol{x}(t)$ by subtracting the mean and dividing by its standard deviation.
2. Perform the moving average filter to obtain ~~$u^{(f)}(t)$~~$\boldsymbol{x}^{(f)}(t)$.
3. Extract from ~~$u^{(f)}(t)$~~ $\boldsymbol{x}^{(f)}(t)$ a training set ~~$u^{(f)}_{train}(t)$~~ $\boldsymbol{x}^{(f)}_{train}(t)$ with $t \in \{1, 2, \ldots, T_{train}\}$.
4. Train the ESN on ~~$u^{(f)}_{train}(t)$~~ $\boldsymbol{x}^{(f)}_{train}(t)$ dataset.
5. Obtain the ESN forecast ~~$v^{(f)}(t)$~~ $\boldsymbol{y}^{(f)}(t)$ for $t \in \{T_{train}+1, T_{train}+2, \ldots, T\}$.
6. Add residuals (Eq. 9) to ~~$v^{(f)}(t)$ sample as $v(t) = v^{(f)}(t) + \delta u$, where $\delta u$~~ $\boldsymbol{y}^{(f)}(t)$ sample as $\boldsymbol{y}(t) = \boldsymbol{y}^{(f)}(t) + \delta\boldsymbol{x}$, where $\delta\boldsymbol{x}$ is randomly sampled from the ~~$\delta u(t)$ with $t \in \{1, 2, \ldots, T_{train}\}$~~$\delta\boldsymbol{x}(t)$ with $t \in \{1, 2, \ldots, T_{train}\}$.

7. ~~Compare $v(t)$ and $u(t > T_{train})$ using~~ Compute the observables $v(t) = \zeta(\boldsymbol{y}(t))$ and $u(t) = \zeta(\boldsymbol{x}(t > T_{train}))$.

8. Using $u(t)$ and $v(t)$, compute the metrics $\phi$, $\tau$ and $\eta$ and evaluate the forecasts.

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

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

that does not require the use of residuals ~~$\delta u(t)$~~$\delta \boldsymbol{x}(t)$.

## 3 Results

The systems we analyze are the Lorenz 1963 attractor (Lorenz, 1963) with the classical parameters, discretized with a Euler scheme and a $dt = 0.001$, the Pomeau-Manneville intermittent map (Manneville, 1980), the Lorenz 1996 equations (Lorenz, 1996) and the NCEP sea-level pressure data (Saha et al., 2014).

### Lorenz 1963 equations

The Lorenz ~~(Lorenz, 1963)~~ system is a simplified model of Rayleigh-Benard convection, derived by E.N. Lorenz (Lorenz, 1963) . It is an autonomous continuous dynamical system with three variables ~~$u \in \{x, y, z\}$~~ $\{x, y, z\}$ parametrizing respectively the convective motion, the horizontal temperature gradient and the vertical temperature gradient. It writes:

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

where $\sigma$, $\varrho$ and $b$ are three parameters, $\sigma$ mimicking the Prandtl number and $\varrho$ the reduced Rayleigh number and $b$ the geometry of convection cells. The Lorenz model is usually defined using Eq. (12), with $\sigma = 10$, $\varrho = 28$ and $b = 8/3$. A deterministic trajectory of the system is shown in Figure 1a). It has been obtained via integrating numerically the Lorenz equations with an Euler scheme ($dt = 0.001$). We are aware that advanced time stepper (e.g. Runge Kutta) would provide better accuracy. However, when considering daily or 6-hourly data, as commonly done in climate sciences and analyses, we hardly are in the case of a smooth time stepper. We therefore stick to the Euler method for similarity with the climate data used in the last section of the paper. The systems is perturbed via additive noise: $\xi_x(t), \xi_y(t)$ and $\xi_z(t)$ are random variable all drawn from a Gaussian distribution. The initial conditions are randomly selected within a long trajectory of $5{\cdot}10^6$ iterations. First, we study the dependence of the ESN on the training length in the deterministic system ($\epsilon = 0$, Figure 1b-d). We analyse the behavior of the rejection rate $\phi$ (panel b), the predictability horizon $\tau_s$ (panel c) and the initial error $\eta$ (panel d) as a function of the training sample size. Our analysis suggests that $t \sim 10^2$ is a minimum sufficient choice for the training window. We compare this time to the typical time

scales of the motion of the sytems, determined via the maximum Lyapunov exponent $\lambda$. For the Lorenz 1963 system, $\lambda = 0.9$, so that the Lyapunov time $\vartheta \approx \mathcal{O}\left(\frac{1}{\lambda}\right) \approx 1.1$. From the previous analysis we should train the network at least for $t > 100\vartheta$. For the other systems analysed in this article, we take this condition as a lower boundary for the training times.

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

**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 a 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 fluctuations are smaller (Weeks et al., 1997; Faranda et al., 2016). It is then of practical interest to study the performance of ESN in Pomeau Manneville predictions as they are a first prototypical example of the intermittent behavior found in climate data.

In particular, the Pomeau-Manneville (Manneville, 1980) map is probably the simplest example of intermittent behavior, produced by a 1D ~~(here $u = x$)~~ discrete deterministic map given by:

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

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. The systems is perturbed via additive noise $\xi(t)$ drawn from a Gaussian distribution. 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 (Korabel and Barkai, 2009). From this latter reference, we can derive a value $\lambda \simeq 0.2$ for $a = 0.91$, implying $w < \tau \simeq 5$. For the Pomeau-Manneville map, we set $u(t) \equiv x(t)$. We find that the best match between ESN and equations in terms of the $\phi$ indicator are obtained for $w = 3$.

Results for the Pomeau-Manneville map are shown in Figure 4. We first observe that the ESN forecast of the intermittent dynamics of the Pomeau-Manneville map is much more challenging than for the Lorenz system as a consequence of the intermittent behavior of this system. For the simulations performed with $w = 0$, the ESN cannot simulate an intermittent behavior, for all noise intensities and reservoir sizes. This is reflected in the behavior of the indicators. In the deterministic limit, the ESN fails to reproduce the invariant density in 80% of the cases ($\phi \simeq 0.8$). For intermediate noise intensities $\phi > 0.9$ (Figure 4-a). The predictability horizon $log(\tau_{s=0.5})$ for the short term forecast is small (Figure 4d) and the initial error large (Figure 4g). The moving average procedure with $w = 3$ partially improves the performances (Figure 4b,c,e,f,h,i) and it enables ESN to simulate an intermittent behavior (Figure 5). Performances are again better when using the exact formula in Eq. 11 (Figure 4b,e,h) than using the residuals ~~$\delta u$~~ $\delta \boldsymbol{x}$ (Figure 4c,f,i). Figure 5a) shows the intermittent behavior of the data generated with the ESN trained on moving averaged data of Pomeau-Manneville system (red) and compare to the target time series (blue). ESN simulations do not reproduce the intermittency in the average of the target signal~~. They~~, which shift from $x \sim 0$ in the non intermittent phase to $0.2 < x < 1$ in the intermittent. ESN simulations only show some second order intermittency in the fluctuations while keeping a constant average. Figure 5b) displays the power spectra showing in both cases a power law decay, which are typical of turbulent phenomena. Although the intermittent behavior is captured, this realization of ESN shows that the values are concentrated around $x = 0.5$ for the ESN prediction, whereas the non-intermittent phase peaks around $x = 0$ for the target data.

**The Lorenz 1996 system**

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

via two sets of equations:

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

where $i = 1, \ldots, I$ and $j = 1, 2, \ldots, J$ denote respectively the number of large-scale $X$ and small-scale $Y$ variables. Large-scale variables are meant to represent the meanders of the jet-stream driving the weather at mid-latitudes. The first term on the right-hand side represents advection, the second diffusion, while $F$ mimics an external forcing. The system is controlled via the parameters $b$ and $c$ (the time scale of the the fast variables compared to the small variables) and via $h$ (the coupling between large and small scales). From now on, we fix $I = 30, J = 5$ and $F = b = 10$ as these parameters are typically used to explore the behavior of the system (Frank et al., 2014). We integrate the equations with an Euler scheme ($dt = 10^{-3}$) from the initial conditions $Y_{j,i} = X_i = F$, where only one mode is perturbed as $X_{i=1} = F + \varepsilon$ and $Y_{j,i=1} = F + \varepsilon^2$. Here $\varepsilon = 10^{-3}$. We discard about $2 \cdot 10^3$ iterations to reach a stationary state on the attractor, and we retain $5 \cdot 10^4$ iterations. When $c$ and $h$ vary, different interactions between large and small scales can be achieved. A few examples of simulations of the first mode $X_1$ and $Y_1$ are given in Figure 6. Figure 6a,c show simulations obtained for $h = 1$ by varying $c$: the larger $c$ the more intermittent the behavior of the fast scales. Figure 6.b,d) show simulations obtained for different coupling $h$ at fixed $c = 10$: when $h = 0$, there is no small-scale dynamics.

~~In~~ For the Lorenz 1996 model~~we can~~, we do not need to apply a moving average filter to the data, as we can train the ESN on the large-scale variables only. Indeed, we can explore what happens to the ESN performances if we turn on and off intermittency and/or the small-to-large-scale coupling, without introducing any additional noise term. Moreover, we can also learn the Lorenz 1996 dynamics on the $X$ variables only, or learn the dynamics on both $X$ and $Y$ variables. The purpose of this analysis is to assess whether the ESN are capable of learning the dynamics of the large-scale variables $X$ alone, and how this capability is influenced by the coupling and the intermittency of the small-scale variables $Y$. Using the same simulations presented in Figure 6, we train the ESN on the first $2.5 \cdot 10^4$ iterations, and then perform, changing the initial conditions 100 different ESN predictions for $2.5 \cdot 10^4$ more iterations. We apply our performance indicators not to the entire $I$-dimensional $X$ variable $(X_1, \ldots, X_I)$, as the $\chi^2$ test becomes intractable in high dimensions, but rather to the ~~spatial~~ average of the large-scale variables $X$. Consistently with our notation, it means that $u(t) \equiv \sum_{i=1}^{I} X_i(t)$ The behavior of each variable $X_i$ is similar, so the average is representative of the collective behavior. The rate of failure $\phi$ is very high (not shown) because even when the dynamics is well captured by the ESN the variables are not scaled and centered as those of the original systems. For the following analysis, we therefore replace $\phi$ with the $\chi^2$ distance ~~T~~ $\Sigma$ (Eq. (5)). The use of ~~T~~ $\Sigma$ allows for better highlighting the differences in the ESN performance with respect to the chosen parameters. The same considerations also apply to the analysis of the sea-level pressure data reported in the next paragraph.

Results of the ESN simulations for the Lorenz 1996 system are reported in Figure 7. In Figure 7a,c,e) ESN predictions are obtained by varying $c$ at fixed $h = 1$, while in Figure 7b,d,f) by varying $h$ at fixed $c = 10$. The continuous lines refer to results obtained feeding the ESN with only the $X$ variables, dotted lines with both $X$ and $Y$. For the $\chi^2$ distance ~~$\mathcal{T}$~~ $\Sigma$ (Figure 7a,b), performances show a large dependence on both intermittency $c$ and coupling $h$. First of all, we remark that learning both $X$ and $Y$ variables lead to higher distances ~~$\mathcal{T}$~~$\Sigma$, except for the non intermittent case, $c = 1$. For $c > 1$, the dynamics learnt on both $X$ and $Y$ never settles on a stationary state resembling that of the Lorenz 1996 model. When $c > 1$ and only the dynamics of the $X$ variables is learnt, the dependence on $N$ when $h$ is varied is non monotonic and better performances are achieved for $800 < N < 1200$. For this range, the dynamics settles on stationary states whose spatio-temporal evolution resembles that of the Lorenz 1996 model, although the variability of time and spatial scales is different from the target. An example is provided in Figure 8, for $N = 800$.

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

**The NCEP sea-level pressure data**

We now test the effectiveness of the moving average procedure in learning the behavior of multiscale and intermittent systems on climate data issued by reanalysis projects. We use data from the National Centers for Environmental Prediction (NCEP) version 2 (Saha et al., 2014) with a horizontal resolution of 2.5°. We adopt the global 6 hourly sea-level pressure (SLP) field from 1979 to 31/08/2019 as the meteorological variable proxy for the atmospheric circulation. It traces cyclones (resp. anti-cyclones) 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 (EOF) of SLP and a wealth of other atmospheric features (Hurrell, 1995; Moore et al., 2013), ranging from teleconnection patterns to storm track activity to atmospheric blocking can be diagnosed from the SLP field.

The dataset consists therefore of a gridded time series $SLP(t)$, consisting of $\sim 33000$ time realization of the pressure field over a grid of spatial size 72 longitudes $\times$73 latitudes. Our observable $u(t) \equiv \langle SLP(t) \rangle_{lon,lat}$ where brackets indicate spatial average. In addition to the time moving average filter, we also 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 EOFs, the nearest neighbors approach has the advantage of not removing the extremes (except if the extreme is not in one of the closest gridpoint) and preserve cyclonic and anticyclonic structures. For $c = 2$ we obtain a 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 fields still capture the positions of cyclonic and anticyclonic structures. Indeed, as shown in ~~(Faranda et al., 2017)~~Faranda et al. (2017), 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 prediction for the sea-level pressure NCEPv2 data on the coarse graining factor $c$ and on the moving average window size $w$ is shown in Figure 9. We show the results obtained using the residuals (Eq. 9) as the exact method is not straightforwardly adaptable to systems with both spatial and temporal components. Figure 9a-c ~~)~~show the distance from the invariant density, using the $\chi^2$ distance ~~$T$~~$\Sigma$. Here it is clear that by increasing $w$, we get better forecast with smaller network sizes $N$. A large difference for the predictability expressed as predictability horizon $\tau_s$, $s = 1.5$ hPa (Figure 9d-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 $\eta$ (Figure 9g-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 almost the minimal $\chi^2$ distance $T$, 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 compare in details the results obtained for two 10-year predictions 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 $w = 0$h (panel b) and $w = 12$h (panel c). 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 with $w = 0$h is steady: positions of cyclones and anticyclones barely evolve with time. Instead, the run with $w = 12$h shows a richer dynamical evolution with generation and annihilation of cyclones. A similar effect can be observed in the ESN prediction of the Lorenz 96 system shown in Figure 8b) where the quasi-horizontal patterns indicate less spatial mobility than the original system (Figure 8a).

In order to assess the performances of the two ESNs with and without moving average in a more quantitative way, we present the ~~space-time distributions~~ probability density functions for $u(t) \equiv \langle SLP(t) \rangle_{lon,lat}$ in Figure 10a~~)~~. 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. Figure 10b-d ~~) shows the wavelet spectrograms (or scalograms) (Hudgins et al., 1993). 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~~ shows the Fourier power spectra for the target data, with the typical decay of turbulent

climate signal. The non-filetered ESN simulation $W = 0$ show a spectrum with very low energy for high frequency and an absence of the daily cycle (no peak at value $10^0$). The simulation with $w = 12$h ~~(d) which shows that, on time scalesof days to weeks, there is a larger variability~~also shows a lower energy for weekly or monthly time-scales but it is the correct peak for the daily cycle and the right energy at subdaily time scales. Therefore, also the spectral analysis shows a real improvment in using moving average data.

## 4 Discussion

We have analysed the performance of ESN in reproducing both the short and long-term dynamics of observables of geophysical flows. The motivation for this study came from the evidence that a straightforward application of ESN to high dimensional geophysical data (such as the 6 hourly global gridded sea-level pressure data) does not yield to the same results quality obtained by (Pathak et al., 2018) for the Lorenz 1963 and the Kuramoto-Sivashinsky models. Here we have investigated the causes for this behavior and identified two main bottlenecks: (i) intermittency and (ii) the presence of multiple dynamical scales, which both appear in geophysical data. In order to illustrate this effect, we have first analysed two low dimensional systems, namely the Lorenz (1963) and the Manneville (1980) equation. To mimic multiple dynamical scales, we have added noise terms to the dynamics. The performance of ESN in predicting rapidly drops when the systems are perturbed with noise. Filtering the noise allows to partially recover predictability. It also enables to simulate some qualitative intermittent behavior in the Pomeau-Manneville dynamics. This feature could be explored by changing the degree of intermittency in the Pomeau-Manneville map as well as performing parameter tuning in ESN. This is left for future work. Our study also suggests that deterministic ESN with smooth, continuous activation function cannot be expected to produce trajectories that look spiking/stochastic/rapidly changing. Most previous studies on ESNs (e.g., Pathak et al., 2018) were handling relatively smooth signals, and not such rapidly changing signals. Although it does not come as a surprise that utilizing the ESN on the time averaged dynamics and then adding a stochastic residual improves performance, the main insights is the intricate dependence of the ESN performance on the noise structure and the fact that, even for non-smooth signal, ESN with hyperbolic tanh functions can be used to study systems that have a intermittent or multiscale dynamics. Here we have used a simple moving-average filter and shown that a careful choice of the moving-average window can enhance predictability. As an intermediate step between the low-dimensional models and the application to the sea-level pressure data, we have analysed the ESN performances on the Lorenz (1996) system. This system was introduced to mimic the behavior of the atmospheric jet at mid-latitude, and features a lattice of large-scale variables, each connected to small-scale variables. Both the coupling between large and small scales and intermittency can be tuned in the model, giving rise to a plethora of behaviors. For the Lorenz 1996 model, we did not have to apply a moving average filter to the data, as we can train the ESN on the large-scale variables only. Our computations have shown that, whenever the small scales are intermittent, or the coupling is strong, learning the dynamics of the coarse grained variable is more effective, both in terms of computation time and performances. The results also apply to geophysical datasets: here we analysed the atmospheric circulation, represented by sea-level pressure fields. Again we have shown that both a spatial coarse-graining

and a time moving-average filter improve the ESN perfomances.

Our results may appear rather counter-intuitive, as the weather and climate modelling communities are moving towards extending simulations of physical processes to small scales. As an example, we cite the use of highly-resolved convection-permitting simulations (Fosser et al., 2015) as well as the use of stochastic (and therefore non-smooth) parameterizations in weather models (Weisheimer et al., 2014). We have, however, a few heuristic arguments on why the coarse-gaining and filtering operations should improve the ESN performances. First~~of all~~, the moving-average operation helps both in smoothing the signal and by providing the ESN with a wider temporal information. In some sense, this is reminiscent of the embedding procedure (Cao, 1997), where the signal behavior is reconstructed by providing not only information on the previous time step, but on previous times depending on the complexity. The filtering procedure can also be motivated by the fact that the active degrees of freedom for the sea-level pressure data are limited. This has been confirmed by Faranda et al. (2017) via coarse-graining these data and showing that the active degrees of freedom are independent on the resolution, in the same range explored in this study. ~~In other words~~Therefore, including small scales in the learning of sea-level pressure data, does not provide additional information on the dynamics and push towards over-fitting and saturating the ESN with redundant information. The latter consideration ~~poses also~~ also poses some caveats on the generality of our results: we believe that this procedure is not beneficial whenever a clear separation of scales is not achievable, e.g. in non-confined 3-D turbulence. Moreover, in this study, ~~note that~~ three sources of stochasticity were present: (i) in the random matrices and reservoir, (ii) in the perturbed initial conditions and (iii) in the ESN simulations when using moving average filtered data with sampled ~~$\delta u$~~ $\delta \boldsymbol{x}$ components. The first one is inherent to the model definition. The perturbations of the starting conditions allow characterizing the sensitivity of our ESN approach to the initial conditions. The stochasticity induced by the additive noise ~~$\delta u$~~ $\delta x$ provides a distributional forecast at each time $t$. Although this latter noise can be useful to simulate multiple trajectories and evaluate their long-term behaviour, in practice, i.e., in the case where ~~a~~ an ESN would be used operationally to generate forecasts, one might not want to employ a stochastic formulation with an additive noise, but rather the explicit and deterministic formulation in Eq. 11. This exemplifies the interest of our ESN approach for possible distinction between forecasts and long-term simulations, and therefore makes it flexible to adapt to the case of interest.

In future work, it will be interesting to use other learning architectures and other methods of separating large- from small-scale components (Wold et al., 1987; Froyland et al., 2014; Kwasniok, 1996). For example, our results give a more formal framework for applications of machine learning techniques on geophysical data. Deep-learning approaches have proven useful in performing learning at different time and spatial scales whenever each layer is specialized in learning some specific features of the dynamics (Bolton and Zanna, 2019; Gentine et al., 2018). Indeed, several difficulties encountered in the application of machine learning on climate data could be overcome if the appropriate framework is used, but this requires a critical understanding of the limitations of the learning techniques.

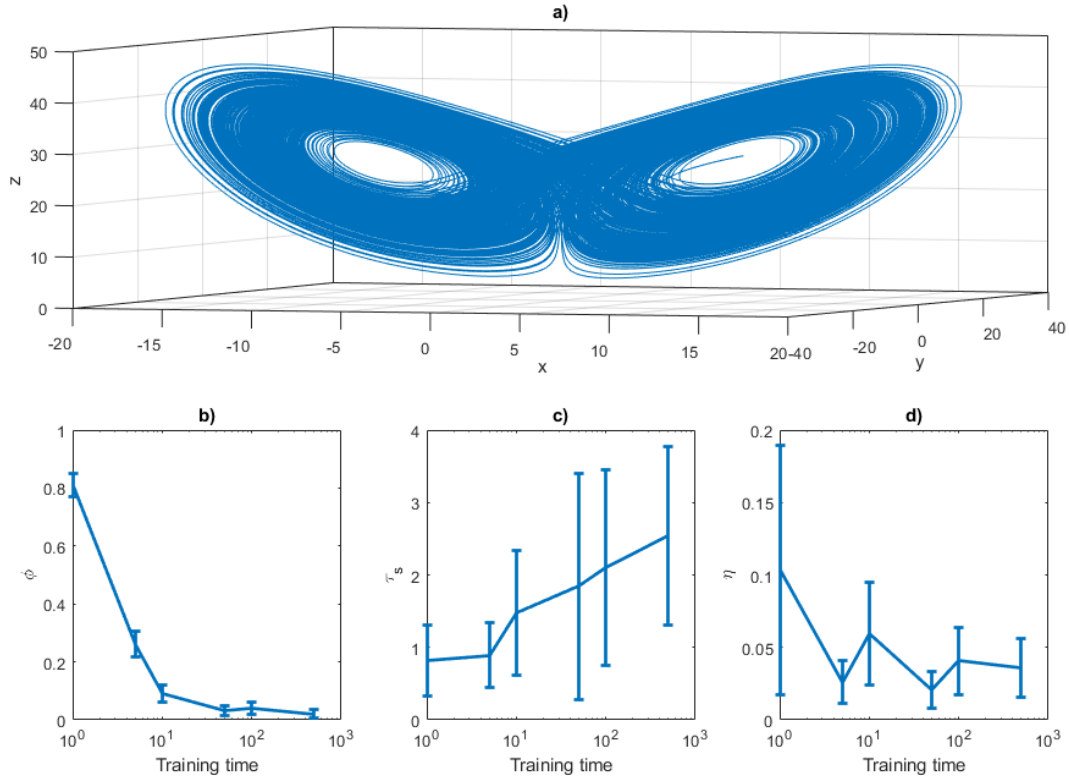



**Figure 1.** a) Lorenz 1963 attractor obtained with a Euler scheme with $dt = 0.001$, $\sigma = 10$, $r = 28$ and $b = 8/3$. Panels b-d) show the performances indicator as a function of the training time. b) the rejection rate $\phi$ of the invariant density test for the $x$ variable; c) the first time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

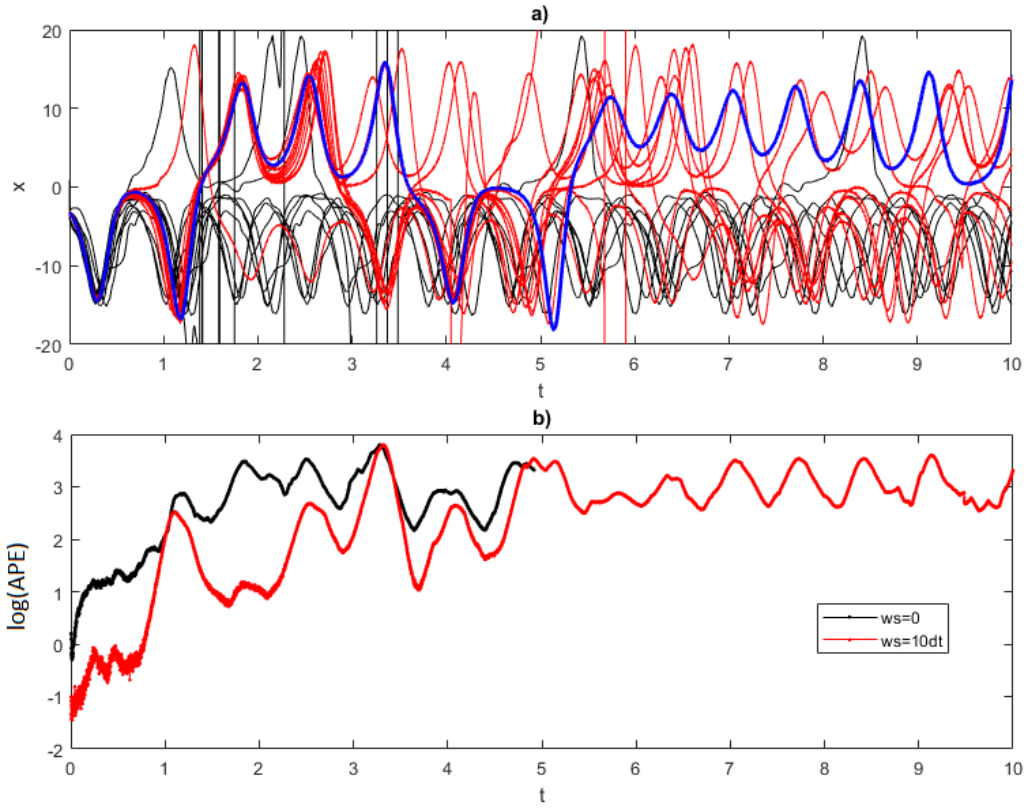



**Figure 2.** a) Trajectories predicted using ESN on the Lorenz 1963 attractor for the variable $x$. The attractor is perturbed with Gaussian noise with variance $\epsilon = 0.1$. The target trajectory is shown in blue. 10 trajectories obtained without moving average (black) show an earlier divergence 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(APE), 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 $10^4$ time steps.

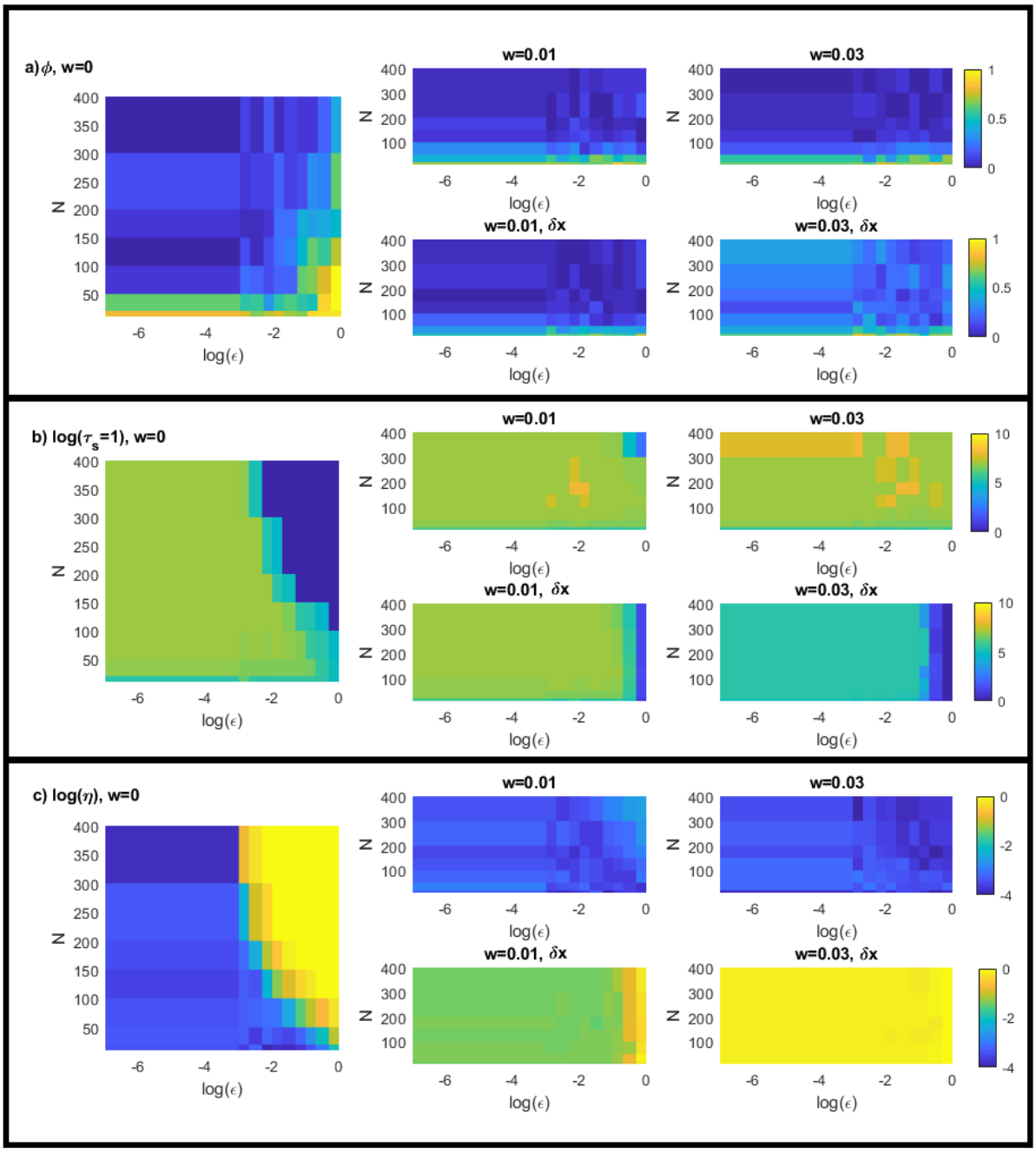



**Figure 3.** Lorenz 1963 analysis for increasing noise intensity $\epsilon$ (x-axes), and number of neurons $N$ (y-axes). The colorscale represents: $\phi$ the rate of failure of the $\chi^2$ test (size $\alpha = 0.05$) (a); the logarithm of predictability horizon $\log(\tau_{s=1})$ (b); the logarithm of initial error $\log(\eta)$ (c). These diagnostics have been computed on the observable $u(t) \equiv x(t)$. 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.

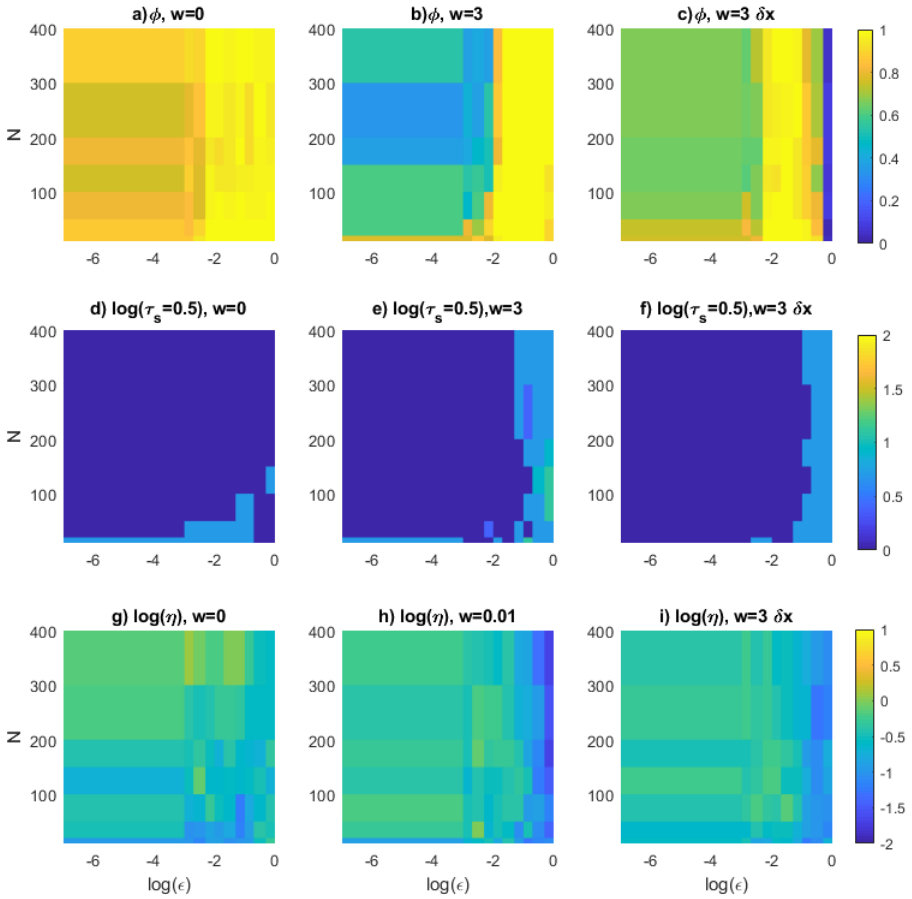



**Figure 4.** Analysis of the Pomeau-Manneville system for increasing noise intensity $\epsilon$ (x-axes), and number of neurons $N$ (y-axes). The colorscale represents: $\phi$ the rate of failure of the $\chi^2$ test (size $\alpha = 0.05$) (a-c); the logarithm of predictability horizon $\log(\tau_{s=0.5})$ (d-f); the logarithm of initial error $\log(\eta)$ (g-i). These diagnostics have been computed on the observable $u(t) \equiv x(t)$ 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 $\delta \boldsymbol{x}$ 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.

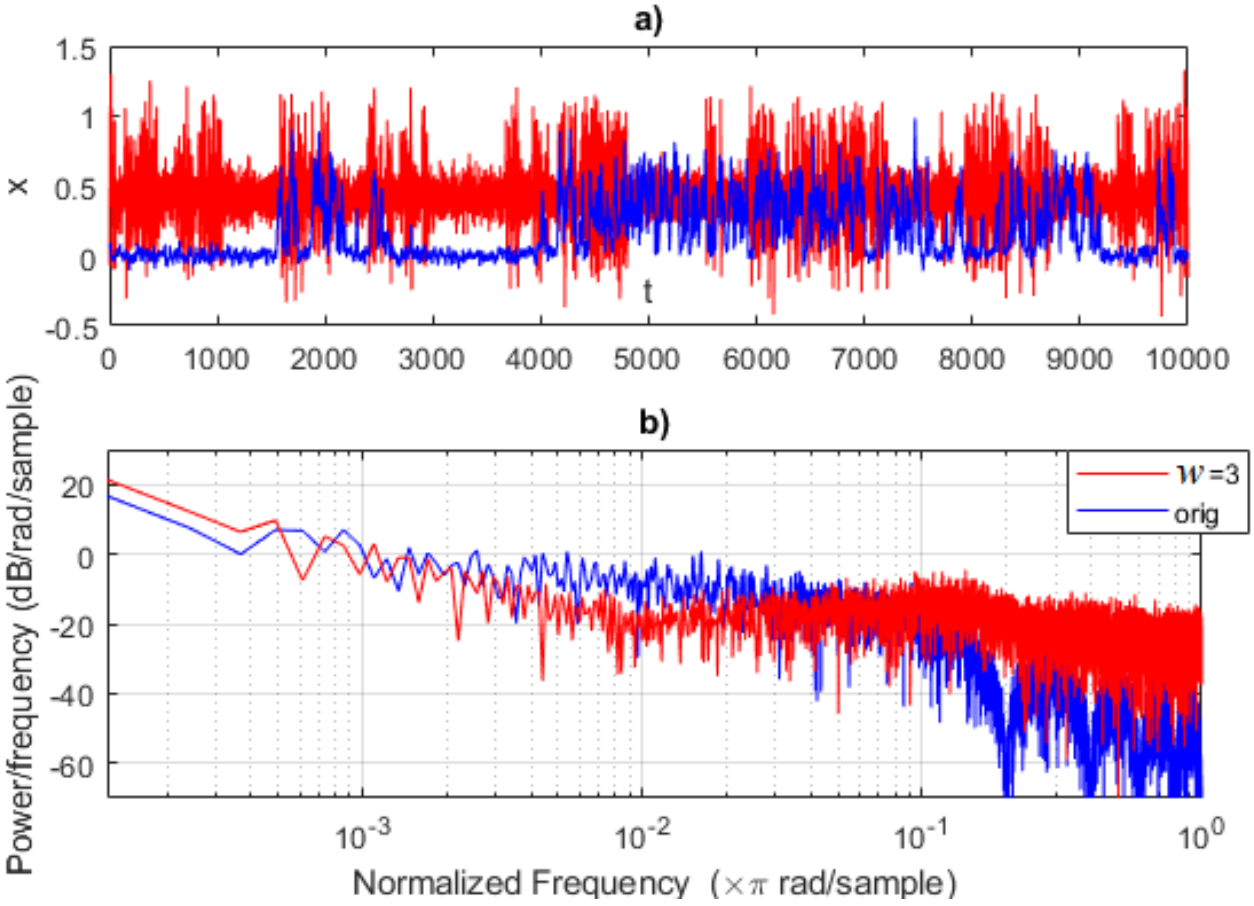



**Figure 5.** Pomeau-Manneville ESN simulation (red) showing an intermittent behavior and compared to the target trajectory (blue). The ESN trajectory is obtained after training the ESN for $10^5$ time-steps using the moving average time series with $w = 3$. It consists of $10^4$ time steps. Cases $w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

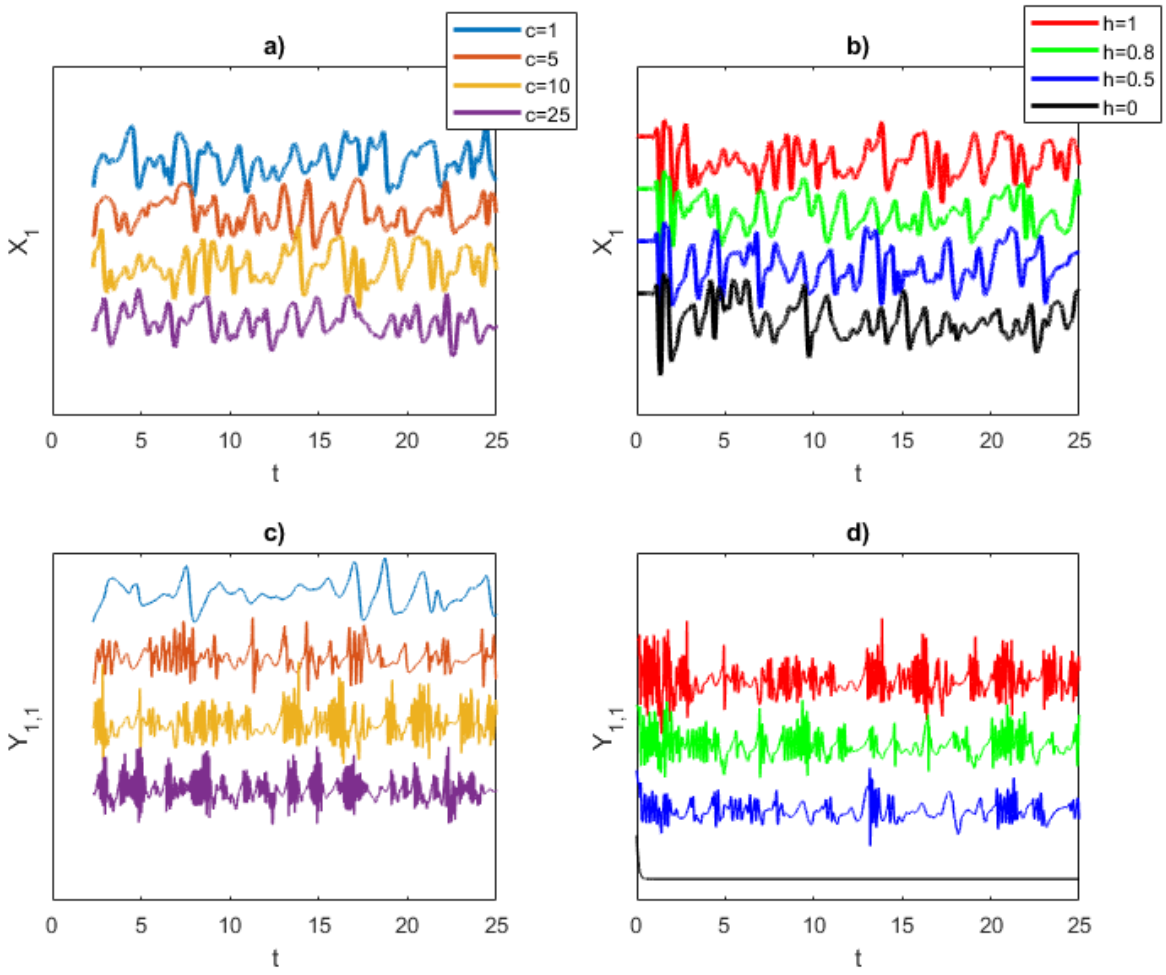



**Figure 6.** Lorenz 1996 simulations for the large-scale variable $X_1$ (a,b) and small-scale variable $Y_{1,1}$ (c,d). Panels (a,c) show simulations varying $c$ at fixed $h = 1$. The larger $c$, the more intermittent the behavior of the fast scales. Panels (b,d) show simulations varying the coupling $h$ 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.

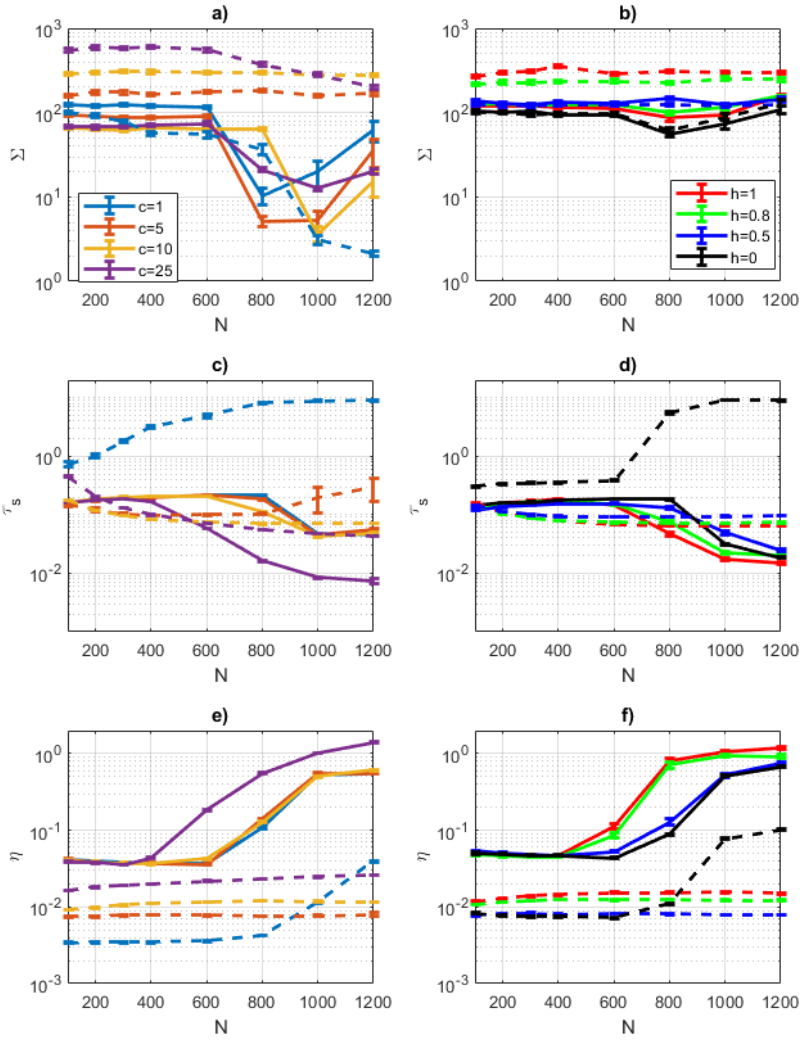



**Figure 7.** Lorenz 1996 ESN prediction performance for $u(t) \equiv \sum_{i=1}^{I} X_i(t)$. a,b) $\chi^2$ distance $\Sigma$; (c,d) the predictability horizon $\tau_s$ with $s = 1$. (e,f) the initial error $\eta$ in hPa. In (a,c,e) ESN predictions are made varying $c$ at fixed $h = 1$. In (b,d,f) ESN 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.

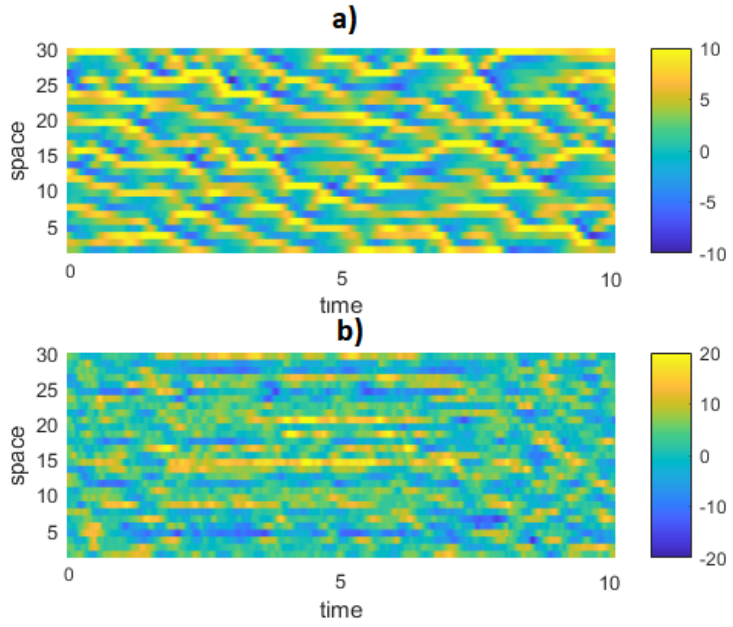



**Figure 8.** Example of (a) target Lorenz 1996 spatio-temporal evolution of large-scale variables $X$ for $c = 1, h = 1$ and (b) ESN prediction realized with $N = 800$ neurons. Note that the colors are not on the same scale for the two panels.

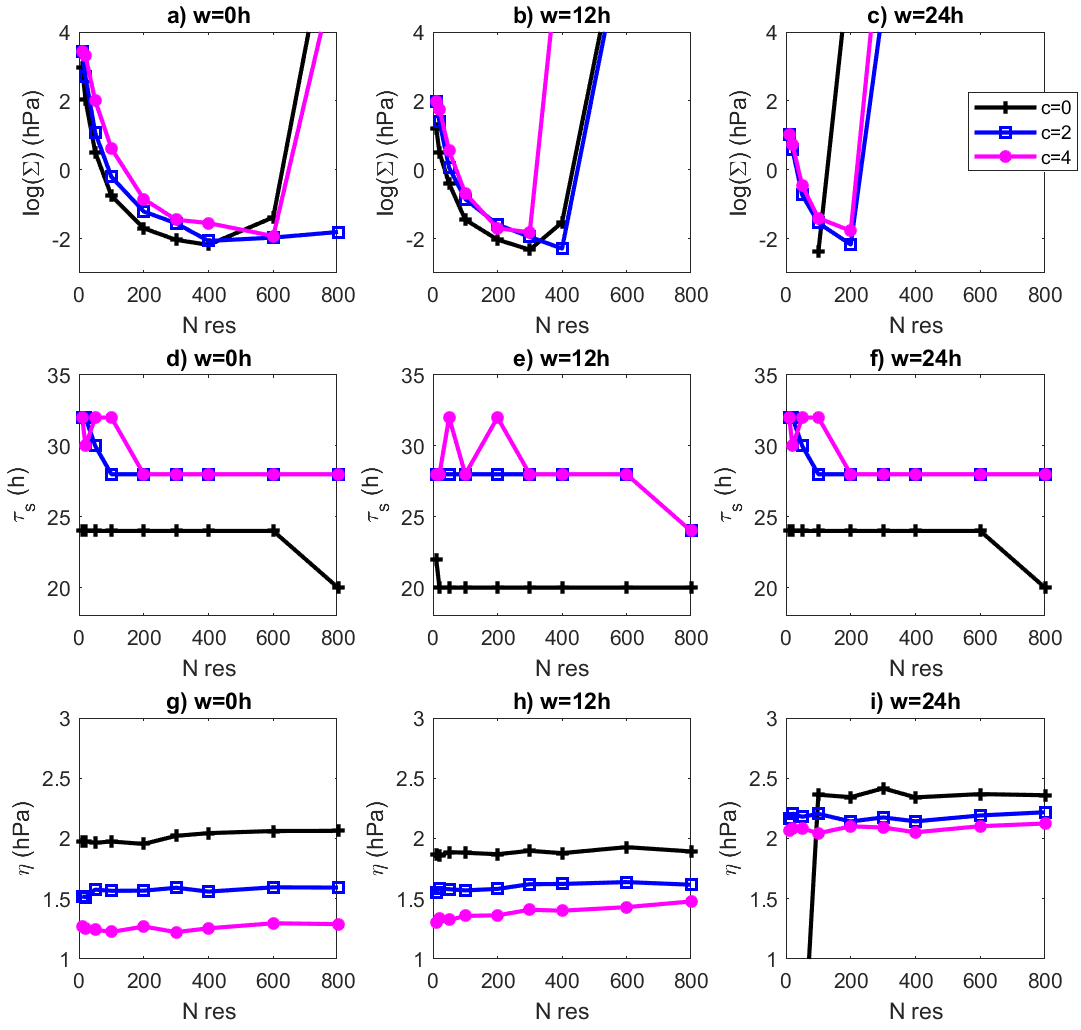



**Figure 9.** Dependence of the quality of the results for the prediction of the sea-level pressure NCEPv2 data on the coarse graining factor $c$ and on the moving average window size $w$. The observable used is $u(t) \equiv \langle SLP(t) \rangle_{lon,lat}$ . a-c) $\chi^2$ distance $log(\Sigma)$; d-f) predictability horizon (in hours) $\tau_s$, $s = 1.5$ hPa; g-i) logarithm of initial error $\eta$. 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.

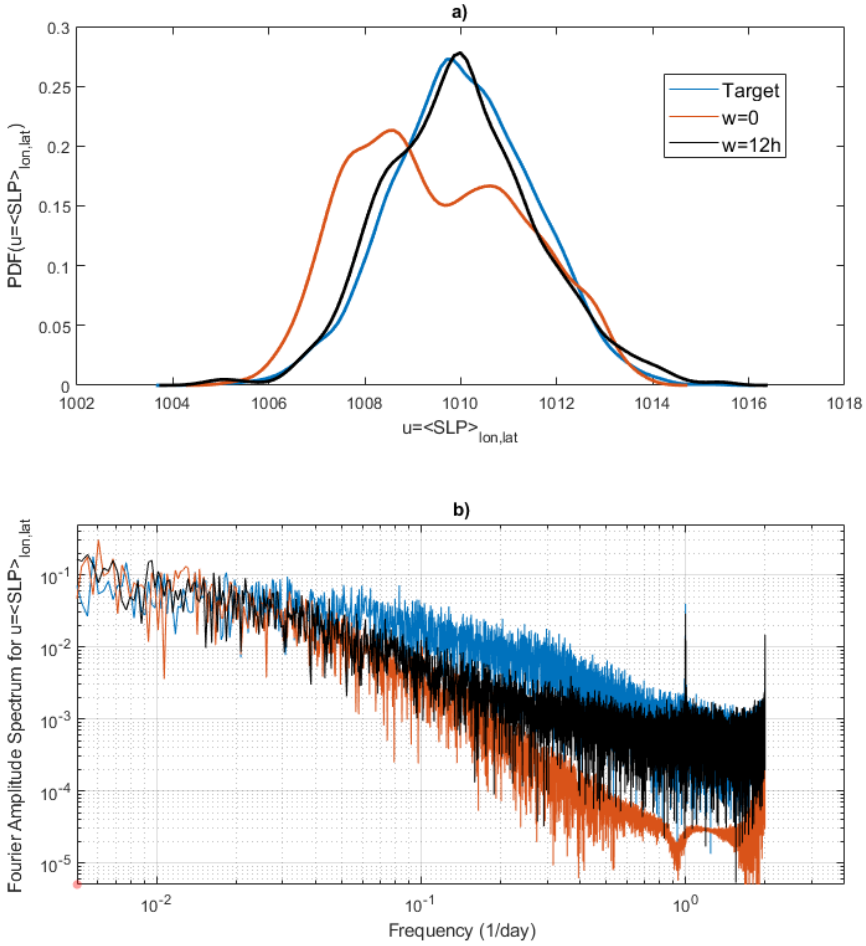



**Figure 10.** a) probability density function and b) Fourier power spectra for $u(t) \equiv \langle SLP(t) \rangle_{lon,lat}$ for the target NCEPv2 SLP data (blue), an ESN with $c = 4$ and $w = 0$ h (red), and an ESN with $c = 4$ and $w = 12$ h (black).

## Appendix A: Numerical code

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/simpleesn

### A1 ESN Training

```
function [Win, W, Wout]=ESN_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);
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
```

### A2 ESN Prediction

```
function [Y_pred]=ESN_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 ESN 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
```

*Code and data availability.* The numerical code used in this article is provided in Appendix A

*Author contributions.* Davide Faranda, Mathieu Vrac, Pascal Yiou and Soulivanh Thao conceived this study. Davide Faranda, Adnane Hamid, Cedric Nguounge Langue and Giulia Carella performed the analysis. Flavio Pons designed and performed the statistical tests. All the authors contributed to writing and discussing the results of the paper.

*Competing interests.* The authors declare that there is no conflict of interest.

*Acknowledgements.* We acknowledge Barbara D'Alena, Julien Brajard, Venkatramani Balaji, Berengere Dubrulle, Robert Vautard, Nikki Vercauteren, Francois Daviaud, Yuzuru Sato for useful discussions. This work is supported by the CNRS INSU-LEFE-MANU grant "DINCLIC".

## References


Bassett, D. S. and Sporns, O.: Network neuroscience, Nature neuroscience, 20, 353, 2017.

Bolton, T. and Zanna, L.: Applications of deep learning to ocean data inference and subgrid parameterization, Journal of Advances in Modeling Earth Systems, 11, 376–399, 2019.

Bonferroni, C.: Teoria statistica delle classi e calcolo delle probabilita, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze, 8, 3–62, 1936.

Bony, S., Stevens, B., Frierson, D. M., Jakob, C., Kageyama, M., Pincus, R., Shepherd, T. G., Sherwood, S. C., Siebesma, A. P., Sobel, A. H., et al.: Clouds, circulation and climate sensitivity, Nature Geoscience, 8, 261, 2015.

Brenowitz, N. D. and Bretherton, C. S.: Prognostic validation of a neural network unified physics parameterization, Geophysical Research Letters, 45, 6289–6298, 2018.

Brenowitz, N. D. and Bretherton, C. S.: Spatially Extended Tests of a Neural Network Parametrization Trained by Coarse-Graining, Journal of Advances in Modeling Earth Systems, 11, 2728–2744, 2019.

Brockwell, P. J. and Davis, R. A.: Introduction to time series and forecasting, springer, 2016.

Buchanan, M.: The limits of machine prediction, Nature Physics, 15, 2019.

Cao, L.: Practical method for determining the minimum embedding dimension of a scalar time series, Physica D: Nonlinear Phenomena, 110, 43–50, 1997.

Cochran, W. G.: The $\chi 2$ test of goodness of fit, The Annals of Mathematical Statistics, pp. 315–345, 1952.

Dueben, P. D. and Bauer, P.: Challenges and design choices for global weather and climate models based on machine learning, Geoscientific Model Development, 11, 3999–4009, 2018.

Eckmann, J.-P. and Ruelle, D.: Ergodic theory of chaos and strange attractors, in: The theory of chaotic attractors, pp. 273–312, Springer, 1985.

Eskridge, R. E., Ku, J. Y., Rao, S. T., Porter, P. S., and Zurbenko, I. G.: Separating different scales of motion in time series of meteorological variables, Bulletin of the American Meteorological Society, 78, 1473–1484, 1997.

Faranda, D., Lucarini, V., Turchetti, G., and Vaienti, S.: Generalized extreme value distribution parameters as dynamical indicators of stability, International Journal of Bifurcation and Chaos, 22, 1250 276, 2012.

Faranda, D., Masato, G., Moloney, N., Sato, Y., Daviaud, F., Dubrulle, B., and Yiou, P.: The switching between zonal and blocked mid-latitude atmospheric circulation: a dynamical system perspective, Climate Dynamics, 47, 1587–1599, 2016.

Faranda, D., Messori, G., and Yiou, P.: Dynamical proxies of North Atlantic predictability and extremes, Scientific reports, 7, 41 278, 2017.

Fosser, G., Khodayar, S., and Berg, P.: Benefit of convection permitting climate model simulations in the representation of convective precipitation, Climate Dynamics, 44, 45–60, 2015.

Frank, M. R., Mitchell, L., Dodds, P. S., and Danforth, C. M.: Standing swells surveyed showing surprisingly stable solutions for the Lorenz'96 model, International Journal of Bifurcation and Chaos, 24, 1430 027, 2014.

Froyland, G., Gottwald, G. A., and Hammerlindl, A.: A computational method to extract macroscopic variables and their dynamics in multiscale systems, SIAM Journal on Applied Dynamical Systems, 13, 1816–1846, 2014.

Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., and Yacalis, G.: Could machine learning break the convection parameterization deadlock?, Geophysical Research Letters, 45, 5742–5751, 2018.

Gettelman, A., Gagne, D. J., Chen, C.-C., Christensen, M., Lebo, Z., Morrison, H., and Gantos, G.: Machine Learning the Warm Rain Process, 2020.

Grover, A., Kapoor, A., and Horvitz, E.: A deep hybrid model for weather forecasting, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 379–386, ACM, 2015.

Hastie, T., Tibshirani, R., and Wainwright, M.: Statistical learning with sparsity: the lasso and generalizations, Chapman and Hall/CRC, 2015.

Haupt, S. E., Cowie, J., Linden, S., McCandless, T., Kosovic, B., and Alessandrini, S.: Machine Learning for Applied Weather Prediction, in: 2018 IEEE 14th International Conference on e-Science (e-Science), pp. 276–277, IEEE, 2018.

Hinaut, X.: Réseau de neurones récurrent pour le traitement de séquences abstraites et de structures grammaticales, avec une application aux interactions homme-robot, Ph.D. thesis, Thèse de doctorat, Université Claude Bernard Lyon 1, 2013.

Hudgins, L., Friehe, C. A., and Mayer, M. E.: Wavelet transforms and atmopsheric turbulence, Physical Review Letters, 71, 3279, 1993.

Hurrell, J. W.: Decadal trends in the North Atlantic Oscillation: regional temperatures and precipitation, Science, 269, 676–679, 1995.

Hyman, J. M. and Nicolaenko, B.: The Kuramoto-Sivashinsky equation: a bridge between PDE's and dynamical systems, Physica D: Nonlinear Phenomena, 18, 113–126, 1986.

Jaeger, H.: The "echo state" approach to analysing and training recurrent neural networks-with an erratum note, Bonn, Germany: German National Research Center for Information Technology GMD Technical Report, 148, 13, 2001.

Jordan, M. I. and Mitchell, T. M.: Machine learning: Trends, perspectives, and prospects, Science, 349, 255–260, 2015.

Korabel, N. and Barkai, E.: Pesin-type identity for intermittent dynamics with a zero Lyaponov exponent, Physical review letters, 102, 050 601, 2009.

Krasnopolsky, V. M. and Fox-Rabinovitz, M. S.: Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction, Neural Networks, 19, 122–134, 2006.

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., and Chalikov, D. V.: New approach to calculation of atmospheric model physics: Accurate and fast neural network emulation of longwave radiation in a climate model, Monthly Weather Review, 133, 1370–1383, 2005.

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., and Belochitski, A. A.: Using ensemble of neural networks to learn stochastic convection parameterizations for climate and numerical weather prediction models from data simulated by a cloud resolving model, Advances in Artificial Neural Systems, 2013, 2013.

Kwasniok, F.: The reduction of complex dynamical systems using principal interaction patterns, Physica D: Nonlinear Phenomena, 92, 28–60, 1996.

LeCun, Y., Bengio, Y., and Hinton, G.: Deep learning, nature, 521, 436, 2015.

Li, D., Han, M., and Wang, J.: Chaotic time series prediction based on a novel robust echo state network, IEEE Transactions on Neural Networks and Learning Systems, 23, 787–799, 2012.

Liu, J. N., Hu, Y., He, Y., Chan, P. W., and Lai, L.: Deep neural network modeling for big data weather forecasting, in: Information Granularity, Big Data, and Computational Intelligence, pp. 389–408, Springer, 2015.

Lorenz, E. N.: Deterministic nonperiodic flow, Journal of the atmospheric sciences, 20, 130–141, 1963.

Lorenz, E. N.: Predictability: A problem partly solved, in: Proc. Seminar on predictability, vol. 1, 1996.

Manneville, P.: Intermittency, self-similarity and 1/f spectrum in dissipative dynamical systems, Journal de Physique, 41, 1235–1243, 1980.

Moore, G., Renfrew, I. A., and Pickart, R. S.: Multidecadal mobility of the North Atlantic oscillation, Journal of Climate, 26, 2453–2466, 2013.

Paladin, G. and Vulpiani, A.: Degrees of freedom of turbulence, Physical Review A, 35, 1971, 1987.

Panichi, F. and Turchetti, G.: Lyapunov and reversibility errors for Hamiltonian flows, Chaos, Solitons & Fractals, 112, 83–91, 2018.

Pathak, J., Lu, Z., Hunt, B. R., Girvan, M., and Ott, E.: Using machine learning to replicate chaotic attractors and calculate Lyapunov exponents from data, Chaos: An Interdisciplinary Journal of Nonlinear Science, 27, 121 102, 2017.

Pathak, J., Hunt, B., Girvan, M., Lu, Z., and Ott, E.: Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach, Physical review letters, 120, 024 102, 2018.

Quinlan, C., Babin, B., Carr, J., and Griffin, M.: Business research methods, South Western Cengage, 2019.

Rasp, S., Pritchard, M. S., and Gentine, P.: Deep learning to represent subgrid processes in climate models, Proceedings of the National Academy of Sciences, 115, 9684–9689, 2018.

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y.-T., Chuang, H.-y., Iredell, M., et al.: The NCEP climate forecast system version 2, Journal of Climate, 27, 2185–2208, 2014.

Scher, S.: Toward Data-Driven Weather and Climate Forecasting: Approximating a Simple General Circulation Model With Deep Learning, Geophysical Research Letters, 45, 12–616, 2018.

Scher, S. and Messori, G.: Predicting weather forecast uncertainty with machine learning, Quarterly Journal of the Royal Meteorological Society, 144, 2830–2841, 2018.

Scher, S. and Messori, G.: Weather and climate forecasting with neural networks: using general circulation models (GCMs) with different complexity as a study ground, Geoscientific Model Development, 12, 2797–2809, 2019.

Schertzer, D., Lovejoy, S., Schmitt, F., Chigirinskaya, Y., and Marsan, D.: Multifractal cascade dynamics and turbulent intermittency, Fractals, 5, 427–471, 1997.

Schneider, T., Lan, S., Stuart, A., and Teixeira, J.: Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations, Geophysical Research Letters, 44, 12–396, 2017.

Shi, X., Gao, Z., Lausen, L., Wang, H., Yeung, D.-Y., Wong, W.-k., and Woo, W.-c.: Deep learning for precipitation nowcasting: A benchmark and a new model, in: Advances in neural information processing systems, pp. 5617–5627, 2017.

Shi, Z. and Han, M.: Support vector echo-state machine for chaotic time-series prediction, IEEE Transactions on Neural Networks, 18, 359–372, 2007.

Sprenger, M., Schemm, S., Oechslin, R., and Jenkner, J.: Nowcasting foehn wind events using the adaboost machine learning algorithm, Weather and Forecasting, 32, 1079–1099, 2017.

Wang, J.-X., Wu, J.-L., and Xiao, H.: Physics-informed machine learning approach for reconstructing Reynolds stress modeling discrepancies based on DNS data, Physical Review Fluids, 2, 034 603, 2017.

Weeks, E. R., Tian, Y., Urbach, J., Ide, K., Swinney, H. L., and Ghil, M.: Transitions between blocked and zonal flows in a rotating annulus with topography, Science, 278, 1598–1601, 1997.

Weisheimer, A., Corti, S., Palmer, T., and Vitart, F.: Addressing model error through atmospheric stochastic physical parametrizations: impact on the coupled ECMWF seasonal forecasting system, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 372, 20130 290, 2014.

Weyn, J. A., Durran, D. R., and Caruana, R.: Can machines learn to predict weather? Using deep learning to predict gridded 500-hPa geopotential height from historical weather data, Journal of Advances in Modeling Earth Systems, 11, 2680–2693, 2019.

Wold, S., Esbensen, K., and Geladi, P.: Principal component analysis, Chemometrics and intelligent laboratory systems, 2, 37–52, 1987.

Wolf, A., Swift, J. B., Swinney, H. L., and Vastano, J. A.: Determining Lyapunov exponents from a time series, Physica D: Nonlinear Phenomena, 16, 285–317, 1985.

Wu, J.-L., Xiao, H., and Paterson, E.: Physics-informed machine learning approach for augmenting turbulence models: A comprehensive framework, Physical Review Fluids, 3, 074 602, 2018.

Xingjian, S., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W.-c.: Convolutional LSTM network: A machine learning approach for precipitation nowcasting, in: Advances in neural information processing systems, pp. 802–810, 2015.

Xu, M., Han, M., Qiu, T., and Lin, H.: Hybrid regularized echo state network for multivariate chaotic time series prediction, IEEE transactions on cybernetics, 49, 2305–2315, 2018.

Young, L.-S.: What are SRB measures, and which dynamical systems have them?, Journal of Statistical Physics, 108, 733–754, 2002.

Yuval, J. and O'Gorman, P. A.: Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions, Nature communications, 11, 1–10, 2020.

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 $\phi$ of the invariant density test for the $x$ variable; c) the first time $t$ such that the RMSE$>1$; d) the initial error $\eta$. The error bar represents the average and the standard deviation of the mean over 100 realizations.

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(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.

Lorenz 1963 analysis for increasing noise intensity $\epsilon$ (x-axes), and number of neurons $N$ (y-axes). The colorscale represents: $\phi$ the rate of failure of the $\chi^2$ 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 $10^4$ timesteps.

Analysis of the Pomeau-Manneville system for increasing noise intensity $\epsilon$ (x-axes), and number of neurons $N$ (y-axes). The colorscale represents: $\phi$ the rate of failure of the $\chi^2$ 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.

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).

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.

Lorenz 1996 ESN prediction performance for the large-scale variables $X$ only. a,b) $\chi^2$ distance $T$; (c,d) the predictability horizon $\tau_s$ with $s = 1$. (e,f) the initial error $\eta$ 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.

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.

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) $\chi^2$ distance $T$; d-f) predictability horizon (in hours) $\tau_s$, $s = 1.5$ hPa; g-i) logarithm of initial error $\eta$. 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.

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).