Nonlin. Processes Geophys. Discuss.,
https://doi.org/10.5194/npg-2020-39-AC1, 2020

Nonlinear Processes in Geophysics
Discussions

Interactive comment on "Boosting performance in machine learning of geophysical flows via scale separation" *by* Davide Faranda et al.

Davide Faranda et al.

davide.faranda@lsce.ipsl.fr

Received and published: 18 December 2020

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 will fully address 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 will take 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 will be 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 will use a more compact and consistent notation.

(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 will take 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 will move some of the key concepts presented in the conclusions, in the introduction section.

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 will make these concepts more clear.

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 strategy we decided to adopt, but after considering different strategies, including those suggested by the referee, we decided to still adopt 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. First, we need to clarify that we did not use a bootstrap methodology, which relies on resampling. Instead, we adopt a Monte Carlo simulation approach, so that each of the 10000 samples is generated under the null hypothesis. This is possible because we are considering simulated systems, for which we can obtain as many independent samples as we want. This choice is made necessary by the fact that, due to limits in the length of each time series, we cannot observe the entire invariant distribution of the process, and therefore we cannot use the theoretical distribution of the test statistics under the null hypothesis. This situation would happen in any case, independently on the chosen test statistic, 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. Not using this procedure would make fixing the level of the test not sufficient to control the probability of Type 1 error. The reasons for the choice of the chi-squared test reside mainly in the construction of the test statistics: this is built considering values of the empirical probability density functions (pdf) over the entire domain, i.e. using all the bins of the histograms. 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. 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 to reject the null hypothesis even in presence of substantially trivial differences between the distributions. 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 will be added to the new version of the paper.

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 will change 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 will give precision about this reference in the new version of the manuscript

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).

We will add these references 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 will add 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 will be 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 will clarify 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).

We will clearly state when we talk ESNs or RNN in the new version of the manuscript.

L90: “We estimate Wout 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. In the new version of the manuscript we will show that the value of lambda is chosen via cross-validation for the Lorenz 1963. Indeed, we will repeat the procedure to find the good level for the Lorenz 1996 and the PM examples, as the value has been directly taken from the cross validation for the lorenz 1963. Thanks for the suggestion.

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.

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. We will improve the readability as suggested by changing the notation and making clear these points.

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 will specify 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?

Root mean squared error is by definition a mean over several values (not necessarily ordered in time): the errors on each single time step are usually averaged on the whole available sample to compute a single forecasting performance metric. However, if one is interested in assessing a model performance over a given time horizon on time series data, say τ, the computation of RMSE can be limited to the period from t to $t + \tau$. For example, suppose we want to evaluate the performance of a statistical model in predicting the next $\tau = 3$ days in a temperature time series; then, for every day in the sample we would only compute the RMSE of days t+1, t+2, t+3. Here we do the same, but using several values of τ to find the maximum predictability horizon, over which the method loses efficacy.

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 will follow the suggestion of the reviewer.

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

Thank you, we will add 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 will be 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 here the idea is 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 will be 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. Since the referee demands it, we will simplify this part by replacing wavelet spectra with conventional spectra.

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 will add 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 Σ? 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 will be 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 CoarseâAËŸ Rgraining. ËĞ 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 dataâAËŸ Rdriven weather fore- ËĞ 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 500âAËŸ RhPa Geopotential Height From Historical ËĞ Weather Data. J. Adv. Model. Earth Syst. 11, 2680–2693 (2019)

We thank the reviewer for the references, which will be properly added to the

new version of the manuscript.

Interactive comment on Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2020-39, 2020.