Editor Decision: Publish subject to minor revisions (review by editor) (21 Jun 2021) by Amit Apte

Comments to the Author:
Dear authors, As you will see from the reviews, there are several changes suggested by the reviewers that I feel will be quite helpful for improving the manuscript. In particular, it would be great if you can address the following comments from report #2 that I quote: "Improved notation; Simpler plots to accompany the detailed ones; A more careful study of the effects of regularization." Based on the available reviews, I am suggesting publication subject to review by the editor.

Dear Editor,

Thank you very much for your help with this paper and no worries for the delay in the reviewing process. We provide below a complete answer to both reviewers' queries. We do hope that the current version of the manuscript, improved as detailed in the markedup attachment, will be suitable for publication in NPG. We have taken extreme care in fixing the notation as requested by the reviewer. However, we do feel that the dependence of the results on the network size is an interesting outcome of our study, especially in climate sciences where there is a clear alternative to Machine Learning methodologies, namely using the underlying equations. As explained in the answers below, we prefer not to add regularization to explicitly show the cases where overfitting is produced.

Best wishes,
Davide Faranda

REPORT #1

The authors argue that there are serious limitations for applicability of out-of-the-box Echo-state network (ESN) in geophysical flows characterized by intermittency and multiple temporal scales, both for short-term forecasts and long-term attractor prediction. The performance can be improved by training ESN on the time averaged dynamics of large scales, and adding stochastic component accounting for small scales. The results are presented for 3 toy models and sea-level pressure dataset, and are convincing overall. I generally like the paper and it can be published once the following issues are addressed.

We thank the reviewer for the positive feedback on our work. We have taken into account the following comments.

1. Deep learning methods with add-on stochastic component have been explored in references below, those should be cited as part of the review in introduction.

Mukhin, D., Kondrashov, D., Loskutov, E., Gavrilov, A., Feigin, A., & Ghil, M. (2015). Predicting critical transitions in ENSO models. Part II: Spatially dependent models. Journal of Climate, 28(5), 1962–1976.

Seleznev, A., Mukhin, D., Gavrilov, A., Loskutov, E., & Feigin, A. (2019). Bayesian framework for simulation of dynamical systems from multidimensional data us- ing recurrent neural network. Chaos, 29(12), 123115. doi: 10.1063/1.5128372

The references have been added to the manuscript.

2. The raw FFT power spectra in Fig.5&10 are too noisy and overlap, and thus very difficult to compare and interpret. Suggest to use spectral methods with proper smoothing, for example Multitaper Method.

We thank the reviewer for the suggestion. MTM methods would indeed be interesting for quantitative estimates of spectral features. Here, we stay rather qualitative in the spectral description. We therefore prefer to keep the present presentation for those figures.

REPORT #2

Overall, this paper illustrates important points about inadequacies in an existing data-driven approach (reservoir computing) to modeling complex chaotic systems with intermittencies and couplings to unobserved (multi-scale) processes. I agree that this work ought to be published and have the following recommendations to make the work more clear and impactful:

1. Improved notation

2. Simpler plots to accompany the detailed ones

3. A more careful study of the effects of regularization

Dear Matthew, thank you for your review of the paper and your suggestions. We have taken care of answering your comments below.

Detailed comments

Comment on ESNs and memory:
I suggest you distinguish between ESNs and other RNN approaches.
See work by Pantelis Vlachas https://arxiv.org/abs/1910.05266
which suggests that although ESNs have the capability for memory, they often struggle to represent it when compared to fully-trained RNNs.
Note that your references to Shi and Han 2007 and Li et al 2012 all use ESNs with delay-embedding representations.
This essentially defeats the purpose of ESNs, as they are supposed to "learn" this.
A caveat in those papers is the Mackey-Glass DDE, but this is a very simplistic type of memory that can easily be stored by an ESN.
From the work by Vlachas, others, and my own experiments, I have deep suspicions about whether ESNs can learn memory meaningfully at all.
So, I do wish that all the reported experiments could be repeated with a simple ANN, GP regression, Random Feature Map, or other data-driven function approximator that does NOT have the dynamical structure of RNNs/ESNs.
These other approaches I list are generally much easier to train and tune than ESNs and have a more clear interpretation.
In particular, the authors might be interested in Random Feature Map approaches (see Gottwald and Reich 2020 treatment of RF-based regression https://arxiv.org/abs/2007.07383),
as it is essentially an ESN without the recurrence (and is a universal approximator for C1(Rn,Rn)).
I am not requesting an entirely new study with new methods, but a discussion of why ESNs are chosen ought to consider the perspective I outlined above.

Thank you, we have added and discussed this issue in the conclusion section: "Our results, obtained on ESN, should also be distinguished from those obtained using other RNN approaches. Vlachas et al.(2020); Albers et al. (2018) suggest that, although ESNs

have the capability for memory, they often struggle to represent it when compared to fully-trained RNNs. This essentially defeats the purpose of ESNs, as they are supposed to learn memory. In particular, it will be interesting to test whether all experiments reported here could be repeated with a simple artificial neural network, Gaussian Processes regression, Random Feature Map, or other data-driven function approximator that does not have the dynamical structure of RNNs/ESNs (Cheng et al., 2008; Gottwald and Reich, 2021

Equations for ESN:
the "t<T" notation is somewhat unclear...do these indicate a matrix? Perhaps capital letters would be better?

We have revised the notation that was indeed unclear.

Notes on regularization (line 96)
In what settings were the values of lambda investigated?
Why limit to 10^-8? Perhaps 10^-9 is better?
More importantly---I am quite surprised that there was no effect of lambda on performance, given that many results show a "sweet spot" for the network size...larger networks performing worse is a classic sign of overfitting that can be addressed with increased regularization.

We have rephrased this sentence as: " Note that we have investigated different values of λ spanning $10^{-8} <\lambda < 10^{-2}$ on the Lorenz 1963 example and found little improvement only when the network size was large, with λ partially preventing overfitting. Values of $\lambda<10^{-8}$ have not been investigated because too close or below the numerical precision"

Statistical distributional test:
I found this setup rather difficult to read, and the ultimate choice to use Monte Carlo approximations unclear and potentially statistically unsound.

See answer below about Monte Carlo.

-To begin, it would help to define zeta as a function mapping between two spaces. Where does u live? Where does x live? How does this relate to y?

We have added to the text that zeta is a function mapping between two spaces. However, we would like to stress that the choices of x (the dynamical system) and u (the observable) are rather arbitrary and they could live in any metric space allowed in dynamical systems theory. The relation between x and y is obtained by substituting r(t) in Eq. 2 with the expression of Eq. 1.

What is x(t)? This is the first time it appears in the paper (line 108).

In the previous version of the paper, x(t) was already defined in line 85 as "Let $\vec{x}(t)$ be the K-dimensional observable consisting of $t=1,2\dots,T$ time iterations, originating from a dynamical system, and $\vec{r}(t)$ be the N-dimensional reservoir state"

Is R_U in R^1? R^N? a banach space?

R_U and R_V are Banach spaces, whose dimension depends on the choice of the observables. This has been added to the text.

What is meant by the "marginal distribution of the forecast sample"? Marginal over what?

We use the term “marginal distribution” in the time series analysis/stochastic process sense, i.e. the marginal distribution is the probability density function of the observed sample taken as i.i.d., as opposed to the joint distribution, which includes the covariance structure of the stochastic process. For example, an autoregressive process with innovations N(0,sigma) will have a marginal N(0, sigma^2/(1-phi^2)) distribution, but a joint multivariate Normal distribution with geometrically decaying autocorrelation.

-The Monte Carlo approach also confused me---please clarify this a bit more.
Where is the randomness in samples coming from? in time? across initial conditions?
Also what is the goal of the MC process? Is it to estimate a baseline \Sigma for f vs \hat{f}?

The rationale behind adopting a Monte Carlo method follows the need for statistical testing the distributional equivalence while dealing with two main problems, that are stated at lines 125-135: we cannot always use the full histogram because of empty bins, and we cannot observe the invariant distribution of the simulated systems, even with very long simulations. These two shortcomings are expected to produce deviation from the assumptions that make the test statistic actually follow a chi-squared distribution.

It is very common in modern statistical inference to use bootstrap resampling to better approximate the asymptotic distribution of an estimator or of a test statistic in presence of small or noisy datasets (see, e.g. the book by Davison & Hinkley, Bootstrap Methods and their Application 1997, Cambridge University Press). In our case, since we are dealing with simulated systems, we do not need to rely on resampling, and we can instead produce large ensembles (here 10^5 trajectories) and tabulate the observed distribution of the test statistic under the null hypothesis to obtain critical values.

Concerning the randomness, there are two sources, as explained at lines 228-230: the system is perturbed with an additional noise, and each trajectory starts from randomly sampled initial conditions. This guarantees a source of randomness even in cases where the amplitude of the noise is 0. This is true for Lorenz 1963 and the Pommeau-Manneville map, while the Lorenz 1996 always has a perturbation term in the first mode (lines 305-310).

-Throughout the paper, neither Sigma nor phi take on a physical meaning for me---it seems that they are simply used to compare methods based on their ability to reproduce statistical quantities. The statistical validity of chi-squared does not seem to be important in the paper.

Indeed Sigma has no physical meaning and it is defined only for statistical purposes. However, it does have a statistical validity, as it is used to perform an inferential test. Even though we use tabulated critical values under H0 instead of the theoretical chi-squared quantiles, this is only done to overcome limitations in the assumptions. Also phi does not have a physical meaning, as it is rather a purely statistical performance measure for the ESN.

Due to this, I would highly recommend using Kullback Leibler-divergence instead of the chi-squared metric---it measures the information loss between probability measures and seems more suited to the job in the paper.
However, I understand this may be significant extra work---perhaps just make a comment that KL could be another option?
Still, I think this section would be much clearer if it were built around KL-div.

The initial choice of the chi-squared was based on the need to perform a statistical test; while the KL divergence is indeed the most theoretically founded divergence measure

between probability distributions, its sample version is not a statistic with a defined probability distribution, so that it doesn't allow to obtain confidence intervals.

Both the KL and the chi-squared divergences are non symmetric, an ideal property when measuring a "proximity" to a reference probability distribution (see e.g. Basu et al., 2019 about the use of non-symmetric divergences in statistics).

The KL divergence is rarely used to directly define estimators or test statistics, even though it is implicitly used, since in the maximum likelihood (ML) approach, the parameters estimated with ML produce the statistical model (in the selected class) with the minimum KL divergence from the one generating the sample. Other than in estimation, KL is also linked to hypothesis testing, as it can quantify the loss of power of likelihood ratio tests in case of model misspecification (Eguchi et al., 2006).

It has been known for a long time that statistical estimation based on minimum chi-squared is equivalent to other popular objective functions, including the log-likelihood and, thus, the minimum KL (see Berkson, 1980).

Our choice was, indeed, based on the statistical validity of the chi-squared as a meaningful metric that allows performing frequentist statistical testing. In particular, one advantage of the chi-squared is that it is independent of the pdf of the distributions to be compared.

We added this explanation to the text, even though we find it could add confusion rather than clarity, since the chi-squared is a very standard goodness-of-fit test statistic, while the KL is very rarely used in such a setting.

Basu, A., Shioya, H., & Park, C. (2019). *Statistical inference: the minimum distance approach*. Chapman and Hall/CRC.

Berkson, J. (1980). Minimum chi-square, not maximum likelihood!. *The Annals of Statistics*, *8*(3), 457-487.

Eguchi, S., & Copas, J. (2006). Interpreting kullback–leibler divergence with the neyman–pearson lemma. *Journal of Multivariate Analysis*, *97*(9), 2034-2040.

-Also, in this section you should give the examples of zeta that are used---1st coordinate and sum of all coordinates. This will help the reader anticipate what is to come.

we have added this to the text

APE:
-The formulation of s is interesting, as it can equivalently be written as \Delta t times the time average of the derivative \dot{u}.
Would it make sense to then have more stringent demands on divergence when the timestep shrinks?
In the dt -> 0 limit, s->0 which seems odd.
Please justify this choice with respect to the chosen sample rate.

Thank you. We have added to the text the alternative definition of APE. Here we intend the divergence only for fixed timestep, so it is to be intended as a statistical metric rather than an asymptotic quantity.

-Also tau_s is not defined (line 148) (or I could not find it).

The definition was given just below in a sentence, but we have now written a specific equation (7) for clarity

Moving average filter:
-perhaps note that the the average can be left/right or centered, but I suppose you choose right side of the window since you want to keep the system Markovian.

This is right, in the manuscript this reads: "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."

2.3 Testing ESN
When reading the algorithm 1-8, I became confused with y(t) vs x(t). Please make explicit what is data / truth and what is coming out of the ESN---is the ESN trained on and predicting the full state? Or an observable?

thank you for your remark. We have added: "Note that the relation between y(t)and x(t) is given in Eqs 1-2."

Step 5---is the ESN forecast always produced directly after a training trajectory?
This seems like a rather limiting case---it would be better to re-initialize the trained ESN on a new trajectory...how do we know each ESN hasn't overfit to a small region of state space?

Yes we have always produced ESN directly after a training trajectory. We have tested starting the trajectory elsewhere for the Lorenz 1963 attractor and we found that the ESN forecasts depend in a non-trivial way from the chosen point. We would like to explore this interesting dependence in a further study.

Step 7: Why does u depend on the future t>Train? This t>T notation is a bit confusing to me, and needs to be defined explicitly.

We have now corrected the notation. We have added "Note that $\vec{x}(T_{train}<t<T)$ is the ensemble of ~\texit{true values} of the original dynamical system."

Step 11: Please define v(f) explicitly. Also, why does this equation hold? I see how (10) is true because xf is defined in terms of x. But in (11) vf is defined by yf (I assume) which is an output from ESN. Is (11) only approximate? Please clarify.

We have added v(f) definition in Step "11". It is true that equation 10 is always valid but Eq 11 is only approximate because of the filtering procedure. This is now specified in the text.

L63
210: are they iid? **Added to the text**
220: green -> black **Corrected**
Fig 2: Where do the many trajectories come from? Different trained ESNs (on the same or different training sets)? **same training set, this has been specified**
Fig 3b: For large noise, why do smaller networks work better? Is this an overfitting problem that can be fixed with regularization? Same question for 3c...if not overfitting, perhaps larger networks need longer time to initialize?

***From this and subsequent comments, we understand that the reviewer does not consider as a main outcome of our work the dependence of the results on *N*, mostly because we could add regularization to fix this. While we understand and respect the reviewer's viewpoint, we would like to stress that the dependence on *N* can be of great interest for climate applications, as we also look for the most parsimonious ML model to perform certain tasks. Indeed, with respect to other scientific fields where ML techniques are the only way to explore the behavior of complex systems, for climate sciences we know the evolution equations of the dynamics, so that ML techniques should also be appealing from a computational point of view to act as a substitute for equations. Furthermore, adding regularization may solve the overfitting issue , it won't clearly improve the results - at least for the examples considered here. We would therefore keep the figures and the results displayed in terms of both N and noise intensity for the exemples considered.**

Fig 3 overall: It is hard to concretely connect phi (and sigma) to the quality of the estimated invariant statistic.
It would help to plot a blue/green/yellow KDE vs the true KDE so that we can see what phi is really discriminating between.

These two indicators are designed to give a statistical measure of the large-ensemble properties of ESN predictions, rather than to quantify the quality of the single trajectory. As already mentioned, \Sigma is a traditional test statistics to evaluate the proximity of histograms, and \phi is simply defined as the large-ensemble average rejection rate of the test, which gives a failure rate of the ESN in reproducing statistically equivalent distributions.

Concerning the distribution plots, direct comparison of the empirical pdfs or cdfs may indeed give an idea of what type of deviations happen in a single prediction. However, this is a simulation study based on 10^5 trajectories of three different variables, so a visualization as described, and as shown in Fig. 10a for the SLP data, does not appear to be feasible.

Finally, I feel these plots could be much simpler by collapsing over (or fixing) N---this is only possible if you can pick a large enough N and that regularization can prevent the "sweet spot" issues wrt N.

We thank the reviewer for this point, however it is our precise choice to show the behaviour of the ESN as a function of the network size, to investigate its stability and flexibility and for the reasons explicitly given in comment (*)

PM map
Great idea to study this system. L63 is very common and often "too easy".
246: This is not a deterministic map, correct? Also, notation \xi_t would be more consistent than \xi(t).

The original PM map is deterministic, our addition makes it stochastic. We keep the notation \xi(t) for consistency with the Lorenz 1963 example where the notation \xi_t would imply a double subscript.

Fig 4: Why is tau_s better for smaller networks?

It appears that in smaller networks, the ESN better tracks the initial conditions, so that the ensemble shows smaller divergences. This has been added to the text.

Again, I hypothesize this is an issue of regularization. Also, again, this figure feels like information overload for me---if possible, this could be easier to understand when fixed/collapsed over N (and this becomes a supplementary figure).

Thank you, however again we would opt for keeping visible the dependence on *N* for the reasons explicitly given in comment (*)

Comment: Is the intermittency you observe driven more by the noise or the deterministic PM map itself? I wonder if the RC performs poorly due to the randomness of the intermittency (driven by \xi_t), rather than the intermittency itself.

It is driven by the deterministic map. We have added "deterministic" in the text. So that the intermittency is not random but driven by the chaotic behavior.

Can ESN handle the PM map with epsilon=0? Why or why not? This seems like a very simple problem that already breaks the data-driven method. Further comment on what is wrong or what needs to be studied would be very valuable here. Also is log the natural or common log here?

When in the text we discuss the deterministic limit, we mean epsilon=0, this means that ESN cannot handle the deterministic PM map. We can therefore speculate that there is an intrinsic problem in reproducing intermittency driven by the deterministic chaos. Log is the natural logarithm, thus it has been added to the text.

L96
In this problem, we finally have unobserved scales (and, hence, memory)---this brings up the challenge of memory I mention earlier and the work of Vlachas et al.
I would be very curious whether you'd see different results with a vanilla ANN; alternatively, I wonder if LSTM (which seems to have more hope of retaining memory) would perform differently. Comments on this would be welcome.

Thank you for the appreciation of these results. We have indicated the ANN analysis as a possible follow-up study of this paper.

304-305: Please clarify this statement.

We have rephrased as: "The rate of failure ϕ is very high (not shown) because even when the dynamics is well captured by the ESN in terms of characteristics time and spatial scales, the predicted variables are not scaled and centered as those of the original systems"

Fig 6: Excellent Figure!!! **Thank you**
Fig 7: My main complaints arise again here:
1) The dependence on N feels secondary to the main point of this plot, which compares performance for different c, h and X vs X,Y.
So, again, perhaps fix an N and show a box plot?
2) 1 can only be done safely if the N-dependence is simpler...which, again, I hypothesize can be addressed via regularization.
If not, please show this. Currently, the results are confusing, as they show many performance metrics worsening for larger networks---this is in-line with an overfitting hypothesis.
3) Showing the actual KDEs will bring much more light to the differences in Sigma.
Also, are APE and Sigma normalized w.r.t. dimensionality? Their definitions do not suggest so, and thus I wonder if X vs X,Y results are fair comparisons. Please double check and clarify in the text.

We thank the reviewer for these points, however it is our precise choice to show the behaviour of the ESN as a function of the network size, to investigate its stability and flexibility, showing overfitting. The reasons explicitly given in comment (*) apply.

Fig 8: This figure is illuminating, as it shows the explicit dynamcs. The evaluation metrics are excellent for high-throughput comparisons, but their meanings can get lost.
How should we interpret Fig 8? Is this better/worse than other methods? Is this a pathologically good/bad example? Or is it average? For contrast, Vlachas et al. seems to have much more realistic looking ESN fits.

We have added:"This figure shows an average example of the performance of ESN in reproducing Lorenz 1996 system when the fit succeeds. For comparison, we refer to the results by Vlachas et al (2020) which shows that better fits of the Lorenz 1996 dynamics can be obtained using back-propagation algorithms."

NCEP
Fig 9: How can Sigma blow up wrt N but tau_s and eta stay stable?
Is there an issue with long-time stability of the ESNs that is not shown? Vlachas et al. reported on this issue.

Yes, indeed the blow-up is related to the long-time stability of the ESN. The blow only affects global indicator Sigma and not tau_s and eta which refers to short term properties. This has been added to the text.

other
Line 73: should be "Finally, we"
Line 284: "the the"

Thanks, corrected

--Matthew LevineR