Review of "Using Probabilistic Machine Learning to Better Model Temporal Patterns in Parameterizations: a case study with the Lorenz 96 model" by Raghul Parthipan, Hannah M. Christensen, J. Scott Hosking, and Damon J. Wischik

Pavel Perezhogin

March 4, 2023

The paper is devoted to the development of the stochastic parameterization of the subgrid tendency in an idealized case of the two-level Lorenz 96 system. The ground truth system integrates both levels simultaneously, while the slow variables are considered as a coarse system. The interaction with the fast variables is not directly resolved in the coarse model and needs to be parameterized. There is an uncertainty in the state of the unresolved variables, and thus it is natural to propose a stochastic parameterization of the missing physics. The new stochastic parameterization is intended to improve the red noise model (AR1) of stochastic residuals with the Recurrent Neural Networks (RNN) which can be seen as a natural generalization of the classical AR1 approach. The RNN model is reduces to the AR1 process in the simplest case when neural networks (b_θ and s_θ) are parameterized with scalar or identity mappings. The major goal of the presented RNN approach is to improve the temporal correlation of the simulated stochastic residuals. The classical RNN model (which is deterministic) is modified with the Gaussian noise in the output layer to make the prediction stochastic. The free parameters of the RNN model are optimized with the maximum likelihood approach. The authors consider the likelihood which is a probability density to observe the trajectory given the probabilistic model with fixed parameters. This likelihood to observe the trajectory is transformed into a function of RNN model variables using the change of variables approach. The likelihood is also considered as an offline metric to evaluate a probabilistic subgrid model. The proposed RNN model is compared to a simple polynomial model and GAN stochastic model. The RNN model outperforms the baselines in the prediction of the spread of the ensemble in the weather forecasting experiment and has a lower KL-divergence metric in online PDFs of the solution and principal components. The generalization to a different forcing shows that in climate online metrics, the RNN and GAN models are similar, and both generalize better than the polynomial parameterization. Generalization in the weather forecasting problem shows superiority of the RNN model with respect to GAN.

The presented approach is a considerable contribution to the research of stochastic parameterizations. My main suggestions are to improve the description of the training procedure and to demonstrate temporal correlation with ACF function. Also, the manuscript can be improved with the discussion of the limitations of the proposed approach. I recommend a **Minor Revision** prior to the publication, and below address the mentioned points in detail.

In case of any questions: pp2681@nyu.edu.

1 Main issues

1.1 RNN model formulation

The form of the model (Eq. 10-12) may be not familiar to general readers. Below I provide possible details that may be included to the text to improve the readability (up to the author's decision):

- The subgrid parameterization is split into the deterministic part (parameterized by mapping g_θ) and stochastic residual r which is parameterized by RNN.

- The RNN model is given in equations (11-12), and it attempts to predict a sequence of residuals at the next time step $r_{k,t+1}$ given a sequence of $r_{k,t}$.
- The major building block of the RNN is the mapping s_θ, which takes as an input the sequence of residuals $r_{k,t}$ and updates a hidden state $l_{k,t}$ recurrently.
- The output layer of the RNN gives a probabilistic prediction of the residual at the next time step $r_{k,t+1}$ with a Gaussian distribution parameterized by b_θ and z.
- Because sequences $r_{k,t}$ and $r_{k,t+1}$ are related to each other, the stochastic prediction is fed back to the input of the RNN.

Also:

- It would be nice to give a reference to the exactly same form of the RNN model (11-12) if it is possible.

1.2 Training algorithm

Line 156: If it is possible, give a reference for the training of RNNs with (similar) likelihood.

The presented description of the training algorithm and loss function (Line 156) are insufficient to reproduce the results. There are missing explanations that:

- The RNN model is trained given the time series of model state $\mathbf{x}_t$ and the true subgrid forcing U_t (or coupling term).
- Be more explicit in "Where r and l are deterministic functions of $\mathbf{x}_0, ..., \mathbf{x}_n$ derived from equations (10)–(12)", particularly:
  - The residual is defined by $r = U_t - g_\theta(x_t)$.
  - l is obtained from equation (12) which is solved given the known history of r and initial condition for l_0
- An explicit expression for the loss function is required after changing $\log Pr(r|l)$ by Gaussian probability density with $\log \mathcal{N}(b_\theta(l), \sigma^2) = const - \frac{1}{2\sigma^2}\left(U_t - g_\theta(x_t) - b_\theta(l)\right)^2$. In my belief, the final loss involving most of the trainable mappings and parameters $(g_\theta, b_\theta, \sigma)$, can be helpful to understand the model. For example, it is clear that both deterministic part of the model (g_θ) and model of stochastic residual $(b_\theta(l))$ are trained to jointly minimize the MSE of subgrid forcing prediction. This joint optimization clearly makes this model superior to the typical approach when first deterministic part is fitted, and then the residual is simulated independently, as in the Polynomial model. This interpretation of the loss may support the statement in Line 285 "For example, ... hardly suffered". I suggest to put the final loss and its interpretation in the Appendix.

1.3 Temporal correlation

The full paper is devoted to the temporal correlation, but there is no any plot of the lagged-autocorrelation function (ACF). Provide please ACF for: true subgrid forcing and simulated subgrid forcing (for Polynomial, GAN and RNN), and ACF for true and simulated residuals (r in case of RNN and h in case of Polynomial).

1.4 Limitations

Here I suggest topics to be clarified in the Discussion section.

The described approach to train RNN model requires computation of the likelihood of the trajectory $(x_1, x_2, ..., x_n)$, but not the residuals (or subgrid forcing), and it is not common. In the appendix, it is explained that for L96 system both likelihoods can be related to each other (A3):

$$\log Pr(\mathbf{x}_t|\mathbf{x}_{t-1}...\mathbf{x}_1) = \log Pr(\mathbf{r}_t|\mathbf{r}_{t-1}...\mathbf{r}_1) - K\log(\Delta t) \tag{1}$$

It remains unclear if this trick is limited to scalar time series generated by L96. So, the discussion section can be improved with the following topic:

- Does the likelihood of the trajectory remain computationally tractable in the general case of GCM model (many correlated time series).

Also:

- In spite of the joint optimization of deterministic part g_θ and stochastic residuals r, equations for stochastic residuals (11-12) are not informed with the solution $\mathbf{x}_t$. So, the presented model is not conditional (unlike the GAN model). Is it possible to make it conditional?

2 Technical issues

- Line 76, ”eliminating the need for update functions to be manually specified”. Make a reference to β or Eq. (3).
- Consider citation of stochastic LSTM model applied to geophysical data [1].
- Line 4: ”physically-informed” and Line 289 ”physical features”, Line 297 ”mainly those which did not leverage physical structure” what is meant in these cases?
- Line 9 and in other places ”probabilistic metric of hold-out likelihood”. Please specify that it is offline metric (computed on the trajectory of the true system).
- Line 41: ”invented hidden variables”, what does ”invented” mean.
- Equation 5: define ϕ and σ (as lag-one correlation and standard deviation of time series)
- Line 101: in L96 equation for X, check j, k indexes in coupling term.
- Equation (7): change $\lambda(X_t/2)$ to $1/2\lambda(X_t)$
- Line 168: Provide the number of training points in the epoch.
- Line 181: I do not think that $t = t_{init} + \tau$ helps the reader, and most likely wrong.
- Line 182: Why $X_{m,n}(t)$ does not have index ”sample”, but X_m^{sample} has. They should be the same, I think.
- Line 209: Give a short definition of ”in a phase quadrature”
- Line 212: Is $||[PC1, PC2]|| = \sqrt{PC1^2 + PC2^2}$.
- Line 230: Check ”The to simulate”
- Line 256: Clarify ”despite the polynomial exploding”
- Lines 259-262: Reformulate the whole paragraph.
- Line 289: ”the the”
- Figure 9: ”Model which allows to learning from high-resolution data” is not clear. The word ”learning” should be specified. As I understand, it is proposed to simulate unresolved variables with RNN.
- Line 293: ”There are more sophisticated ways to model the system”: reformulate.
- Line 310: Probably better to say ”GAN loss function to train RNN”.
- Check Discussion and Conclusion: the main results of the paper are not stated clearly, i.e. there is no summary.
- Equation (A2): Clear distinction between $\mathbf{x}_t$ and $\mathbf{X}_t$ will help; define the notation for parentheses $(x|y)$; define f – is it all deterministic parts of RHS including deterministic part of the subgrid model?
- Line 343: Give a reference or explain ”change of variables”.

References

[1] Agarwal N, Kondrashov D, Dueben P, Ryzhov E, Berloff P. A Comparison of Data-Driven Approaches to Build Low-Dimensional Ocean Models. Journal of Advances in Modeling Earth Systems. 2021 Sep;13(9):e2021MS002537.