Nonlinear Processes in Geophysics Discussions

Nonlin. Processes Geophys. Discuss., author comment AC2
https://doi.org/10.5194/npg-2021-20-AC2, 2021

Reply on RC2

Raphael Kriegmair et al.

Author comment on "Using neural networks to improve simulations in the gray zone" by Raphael Kriegmair et al., Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2021-20-AC2, 2021

I have read with interest the manuscript: "Using neural networks to improve simulations in the gray zone" by Raphael Kriegmair et al. and found it of potential interest for the public of Nonlinear Processes of Geophysics. However, before the paper could be considered for publication, I would like the authors to answer/consider the following specific comments on their work. I would be very happy to read a revised version of their paper.

We thank the reviewer for his important comments. These and other reviewer's comments have led to significant changes in the manuscript:

- ***The partition of training/validation data has been changed. All plots are updated accordingly. The main conclusions remain unchanged.***
- ***A new Figure was added showing the trajectory of both nature runs.***
- ***A new section (section 3: Verification methods) was introduced to clarify our verification metrics.***
- ***The bar plots for the single time step predictions (In the new manuscript Figures 4,7 and 8) are displayed differently and contain more information.***

Specific Comments

1) Introduction: While reading the introduction I was surprised that the authors talk about "gray zone" always avoiding mentioning the concept of turbulence (which is, by the way, mentioned in the title of one of the references provided). In my own view, and I do hope that the authors agree, the gray zone is an effect of coexisting turbulence cascades (direct and inverse) and the emergence of specific phenomena at certain scales due to the physical and geometrical constraints of the system. For exemple, in atmospheric motions, cumulus clouds and more generally convective atmospheric phenomena are constrained, in scale, by the height of the tropopopause. Similarly cyclones and anticyclones have a radius depending on Earth rotation and so on. The authors could discuss this issue and provide additional references for the gray zone with repsect to the concepts of turbulent cascades. See for exemple:

- Lovejoy, S., and D. Schertzer. "Towards a new synthesis for atmospheric dynamics: Space–time cascades." Atmospheric Research 96.1 (2010): 1-52.
- Marino, Raffaele, et al. "Inverse cascades in rotating stratified turbulence: fast growth of large scales." EPL (Europhysics Letters) 102.4 (2013): 44006.
- Faranda, Davide, et al. "Computation and characterization of local subfilter-scale

energy transfers in atmospheric flows." Journal of the Atmospheric Sciences 75.7(2018): 2175-2186.

We agree that a turbulence-based perspective on the gray zone is a useful part of the motivation of our paper, and thank the reviewer for the suggested references. We have added the following paragraph to the introduction, referring the interested reader to the excellent review of Honnert et al. (2020) for a detailed discussion.

"Viewing the atmosphere as a turbulent flow, with up- and downscale cascades, phenomena like synoptic cyclones and cumulus clouds emerge where geometric or physical constraints impose length scales on the flow (Lovejoy and Schertzer 2010, Marino et al. 2013, Faranda et al. 2018). If a numerical model is truncated near one of these scales, the corresponding phenomenon will be only partially resolved and the simulation will be inaccurate. In particular, the properties of the phenomenon may be determined by the truncation length, rather than by the physical scale. A thorough review of the gray zone problem from a turbulence perspective is provided by Honnert et al. (2020)."

2) Experiment set-up: Here the authors attempt to describe their model largely using other existing references but, even digging into the cited literature, it is complicated to understand what is the exact model used. I strongly advise to: i) write the full equations of the model (if it is too long, you can think of doing an appendix), ii) when you say "We pick one simulation from each extreme and compare results to identify general and flow dependent aspects", please show some trajectory of your model in space & time (at least part of it when the system has settled in a stationary states). Figure 12 indeed shows some space snapshot of the system's stat but it comes too late in the manuscript to be useful for the casual reader.

We agree that a visual aid for the model is helpful at this point. We added the trajectory of the model in space & time for both orographies as the reviewer suggests (Figure 1). The reference Kent et al., 2017 is the original publisher of the model we use and they describe the model in detail (including equations). We use their published model code. We have also made our own code available. We therefore believe it is not necessary to republish the equations.

3) Parameters used in this study:

-"The coarse graining factor in this study is set to 4" why is that? the authors should provide a jutification of this value. Any reviewer or reader would question the choice of the value 4 as the only one explored in the paper. I strongly reccomend to see what happens for power-2 values, at least to some extent. In the cited paper by Faranda et al. we have seen that the coarse-grain factors can greatly affect the performances of ML methods. This item should deserve particular attention in the revision of the paper.

It is very likely that our results will be sensitive to the choice of coarse-graining factor, as the reviewer notes, but the information gained from testing a wider range of factors is unlikely to provide useful information for the problem we are considering. Smaller factors are unlikely to be of much interest in practice, since the different resolutions are very similar. Larger factors would change the nature of the learning task by changing the physical problem, This is now discussed in the revised manuscript at L81ff.:

"The coarse graining factor in this study is set to 4, which is analogous to the range of scales found in the gray zone where deep cumulus convection is partially resolved (e.g. 2.5-10 km). Faranda et al. (2018) show that the choice of coarse graining factor can substantially affect the performance of ML methods. In our case, however, choosing a larger factor would correspond to a coarse model grid length that is larger than the typical cloud size, changing the nature of the problem from learning to improve poorly resolved existing features in the coarse simulation to parameterizing features that might not be seen at all."

-"T=200000 time steps". How can we say that this time series is long enough? what is the Lyapunov time of the system? please justify this value as, again, the length of the available dataset is a crucial parameter in ML studies.
The decorrelation length scale of the model is around 4 hours and T=200000 corresponds to approximately 57 days. We have added the following:
" A time series of $T=200 000$ time steps, which is equivalent to approximately 57 days, is generated for both orographies. The first day of the simulation is discarded as spin up, the subsequent 30 days are used for training and the remaining 26 days are used for validation purposes. The decorrelation length scale of the model is approximately 4 hours. "

-"The ANN structure used in this research is described in the following. 5 hidden layers are applied, each using the ReLU activation function. The input layer uses ReLU as well, while the layer uses a linear activation function. All hidden layers have 32 filters. The input and output layer shapes are defined by input and target data. The kernel size is set uniformly to 3 grid points." Please justify the choices "5 layers"; "32 filters" and " 3 grid points". Ideally, you should include additional tests to show that these parameters are a good choice for your analyses and why you have not attempted other combinations.
Given the amount of hyper-parameters to tune for neural networks, a dedicated hyper-parameter search would be beyond the scope of this paper. We tested slightly different configurations and did not see significant changes to the results, nor strong overfitting, giving us confidence that the NN is reasonable for the task. We added the following sentence:
"The ANN architecture and hyperparameters were selected based on a loose tuning procedure, where no strong sensitivities were detected."

4) Convolutional ANN: as for the model used, The convolutional ANN should be defined with equations, with explicetely defined parameters. Again, if this makes the main text too long, you can move this important information in the appendix.
Since we are using the standard implementation of a convolutional neural network, a thoroughly abundant algorithm, we believe that adding equations would add a lot of text while not being useful for most readers. Rather, we now reference Goodfellow et al, a standard textbook on deep learning, that includes all the equations as used in the paper. Also, we now clarify that we use the python library Keras and list the corresponding reference. Finally, our code is made available.

5) Results:
-Figure 2: 5 epochs do not seem enough to conclude anything on the variability. Why using only 5 epochs? you can use 30 and make boxplots instead of just showing 5 points. Otherwise please justify your choice 5x5
We agree that to investigate the variability stemming from epoch number versus the variability stemming from initial weights we would need more samples of each dimension. However, we are not trying to investigate the individual contribution of either of these factors. We want to sample the variability of the ANNs as a whole, for which we believe 25 samples is reasonable. Note that the computation of 50 twin experiments for each of the 250 ANNs (25 ANN realizations * 2 orographies * 5 weightings) is already a lot. The main goal of Figure 3 (new paper version, Figure 2 int the old paper version) is to justify our choice of how we get our 25 samples of ANNs. If, for example, there would barely be sensitivity to the epoch number (which is not the case), we would have been forced to train 25 ANNs for each setting to obtain our 25 samples. % Also, (though I agree we probably could use more than 5), we are limited by the number of epochs we can use per trained ANN, as the loss value of the validation data set has a minimum which we want to target.
We rephrased:

"As the initial training weights of the ANNs and the exact number of epochs performed is to some extent arbitrary, it is desirable to measure the sensitivity of our results to the realization of these quantities. Figure 4 shows the MSE of the validation data set of the last 5 epochs (y-axis) for 5 ANNs with different realizations of initial training weights (x-axis) for both orographies. Since the MSE appears sensitive to both the initial weights and the epoch number, we use both to sample the total ANN variability, resulting in 5$\times$5=25 samples for each ANN training setup that is presented in the remainder of this paper."

-Figure 3: define RMSE
We have added section 3 "verification methods" where we define all the scores we use with equations.

-Section 3.2: it is very difficult to follow the exact way you actually train your ANN with w_mass because you never provided the original equations. Again, my suggestion is to add the relevant equations to understand the ANN dynamics and the way you add w_mass to improve the performances.
w_mass only comes into play in the loss function, which is equation (1) in the manuscript. We have added references for convolutional neural networks (Goodfellow etal, 2016) and the specific python library we use (Chollet et al, 2015). In addition we made our code available (https://doi.org/10.5281/zenodo.4740252, Kriegmair et al., 2020) .

6) Conclusions:
The authors' conclusion are consistent with the material presented in the paper. I have however suggested (see my previous comments) several way for the authors to largely improve their manuscript. In particular, I would expect to see a better model description, as well as additional analyses on the meta-paraemeters used (coarse grain factor, input layers, kernel size, and grid points numbers.