Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2019-319-RC2, 2020

Geoscientific Model Development Discussions

Interactive comment on "Coupled online learning as a way to tackle instabilities and biases in neural network parameterizations: general algorithms and Lorenz96 case study (v1.0)" *by* Stephan Rasp

Anonymous Referee #2

Received and published: 27 January 2020

This paper address an interesting approach to perform ML-based subgrid parameterization. The method itself is relevant for a publication, but I still think that some points could be discussed further.

General comments:

- I feel some inadequate between the general objective of the method the setup chosen to address this objective. Even the author acknowledges clearly that L96 model is not adequate, it is difficult for me to have a clear idea of what was or not shown by the L96 example. Maybe it can be summarize in the form of an introduction (or conclusion) to section 6:

1) questions addressed by online learning and that L96 confirms (mainly sanity check as far as I understand, but it shows also that you could improve an existing ML configuration)

2) question addressed by online learning and for which L96 is not sufficient and would require more test (e.g. improvement of a forecast of the ML-LR model without HR)

3) question not addressed by online learning as introduced in the paper but that could be implemented (e.g. physical constraints)

4) question not addressed yet by online learning (e.g. real observations, stochastic parametrizations)

I think such a summarized section would help the reader to clarify the contributions of the paper.

- I have also some concern about the justification of the nudging. The nudging aims at avoiding HR and LR to diverge as expected in case of chaotic models. So, my understanding is that it aims at correcting from errors due to initial conditions errors (which are due to past errors themselves). But then, could you explain more in what way restoring the HR model to the LR state could be problematic? You said that it was due to the fact that you can be outside the attractor, but isn't the purpose of ML-LR to be precisely in the HR attractor? Or you mean that you should nudge carefully for the first training step before ML-LR has converged toward the HR attractor? If there is a persistent bias between HR and ML-LR, it means either that the Neural Net is not able to learn it or that you have corrected it by the nudging, so you don't tell the Neural Net to learn it. These are quick questions (maybe not all relevant), but I think a more extensive discussion about this point could be done.

- Last point: as far as I understand, there are two methodological novelties. First, the fact that you learn online (meaning that the LR forecast is already computed using the ML parametrization). Second, the way you construct your training set is a bit different

from all the methods you detailed in the introduction. Computing, for each time step, an "assumed" HR tendency could be done offline using the Non-ML physics and dynamics as a LR model. It would be different either from RPG18 in which there is no HR model per se, and it would be also different from BB18 in which they don't run a LR model but using a coarse-grained HR model. The second innovation is not presented as a specific point (maybe in a sense, it is equivalent SP setup), but it would worse a discussion. In other words: in Fig.2, if I remove the ML physics, and I make the training offline (afterwards), is it equivalent to a previous approach (and if not, would it be valuable)? In still other words: if the training frequency M is very large, what does it give?

Other details comments:

L47: "In our subsequent online tests", could you be more explicit? What are the tests you are referring to?

Eq. (1): even if is implicit could you quickly define all the terms of the Equation: \overbar{v}, \overbar{\phi}, \overbar{\nabla}

L76: "Random forests are also competitive with neural networks for many types of ML problems". I find this assertion a bit too general to be really informative. Is there a particular problem, with a similar degree of complexity as subgrid parameterization, in which RF has similar performances as Neural Networks?

L107-109: HR increment calculation. Even if I support the effort to make the explanation not too technical, I think the description could benefit for a bit more of formalism. If you define what you call tendency and increment using equations, you could introduce very clearly the different terms. Also, you could add the time step to the formulas (\Delta\phi_{\rm nudging}) depend on a time step. Is the nudging term assumed constant between t and t+\Delta_{LR}?

L110: I don't really get here what is called "assumed HR-internal increment". What do you mean exactly by "the nudging and the HR-internal evolution are independent"?

L110: I think that using time reference would clarify a lot the expression. (I understand HR' is calculated at time t+\Delta t and HR is calculated at time t). Also; in theory, you can compute the increment of the HR model with or without nudging. So why do you need to make the "linear superposition assumption?"

L115: this is true if you neglect the error due to initial conditions, but what is the effect of such error on the loss?

L140: I am a bit surprised by the architecture. Several studies (Pathak et al. 2018, Dueben and Bauer 2018, Bocquet et al. 2019) suggest that it is relevant to take into account locality of the parametrization (see also Fig.3 of Wilks 2005 that suggests a quick spatial decorrelation). Could you comment on that?

L145: Here again, could you add a quick justification of the setup. In the introduction, the case you address is that the attractor of the pure LR-model and the ML-LR model is different. Is it not the case here without changing model configuration?

L158: Would it also be possible to add regularization term on the weights to avoid them to have a too strong update? Figure 4: Could you define the true network? If the true network is the one learnt offline with the real tendency, it seems to me that it is exactly the network you could find in an offline approach and for which you say it was subject to divergence. L209-210: I don't understand this point

On the code (found on GitHub): Learning linear model using Adam is not standard and probably not as efficient than using dedicated fit for linear problems.

very minor details:

L64: "an multi-time-step" -> a multi-time-step

P3 Footnote 2: could you expand SPCAM notation

L81: "ML-LR": Could you define the acronym the first time you are using it?

L121-122:" Here, I use the model as described in Schneider et al. (2017a)." the refer-

ence to Schneider is not the most relevant want to introduce the Lorenz model. This configuration is pretty standard and you find the same in numerous articles. Potentially, it is a good place to cite the already mentioned Lorenz 95 paper.

L135: It would be interesting to introduce the polynomial parametrization as in Wilks 2005, as it is rather standard for this model.

Algoritmh 3: SP-CAM -> SPCAM?

L200: SP-CAM (SPCAM) why is it specific to SPCAM?

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2019-319, 2020.