General comments

The article presents work with the aim to use machine learning to nudge a coarse version of the EAMxx-SCREAM model towards the results of a high resolution run with the goal to obtain results corresponding to high-res runs at significantly lower computational costs. For this, the authors follow previous work with another Earth system model, FV3, and use the openly available algorithms that were successfully implemented for FV3. However, in the present article, the results of the nudging were less promising than in the previous work. The authors speculate as to why this is the case. They also argue that, despite the less effective nudging, there are additional benefits stemming from the model developments made, such as Python to C++ bridging added to the model which will also be beneficial for other developments.

In general, this is an interesting and highly timely work that should be published. I also feel very favorable to publishing work that has not proven entirely as successful as initially hoped. However, in the present form, the article feels incomplete and lacks motivation and discussion and thorough investigation as to why the work was not as effective as for the FV3 model, which would make the article useful for readers attempting similar endeavors. I therefore recommend major revisions as detailed in the appendix.

Specific comments

- Since the authors argue that there are benefits from their model developments other than the attempts at nudging to high resolution, these benefits should be stated more clearly in the motivation of the work.
- Also, it is not clear to me whether the changes made to the SCREAM code are openly available (important with respect to the benefits claimed above) – the section on Open data is very superficial.
- What is the throughput of the tested models (high res, coarse without and with nudging)?
- A detailed description of the models before ML is missing. Which parameterizations are active?
- How do the authors deal with the spin-up periods of the models? What are the initial conditions? How do they ensure that the coarse and high res models simulate the same climate state? How long were the runs for training?
- A detailed investigation of the produced nudging tendencies is missing. Here I would expect some suitable figures. Do the nudging tendencies make physical sense, considering the better resolution of physical processes in the high res? How robust are they with the different seeds or with slightly different starting conditions? How large are the nudging tendencies in comparison with the coarse unnudged tendencies? What is the consequence of nudging with 3-hour average fields?
- What motivates the use of the variables used for validation of the ML model (Tab. 2)? I notice that no dynamics variables are present despite the nudging also being applied to the winds?
- There is no detailed description of the ML models trained, please add. Please also add a description of the out-of-sample novelty detector in the methods section.
- How often is the mass clipper applied (in percent)?
- Can the authors make some sense about why they see improvements in some variables and not in others? Please add discussion here, e.g. connected to the extended discussion of the resulting nudging tendencies.
- Please add a table of all the trainings done, how many runs of them failed? (p. 8)

- Figs. 2, 3, 4 – please describe clearly what is seen in the caption, not only discussion, and make sure that the labels in the figures are correct (e.g. Delta pressure for bias in pressure)

Technical corrections

- The second sentence in the introduction (p. 1) is not a complete sentence.
- P. 1 "This project aims to develop a computationally efficient machine learning based emulator for SCREAM" seems misleading, since it is not a complete emulator of the model that is developed, just the ML-based nudging.
- P. 7 first sentence in last paragraph missing "to": While it is possible avoid … "
-
- P. 2: "both approaches have shown successful results, often outperforming state-of-the-art GCMs while only using a fraction of the computational resources" – this statement seems too simplified, please elaborate
- P. 2 First abstract in Section 2: Please add a reference for "maintains good performance across a number of high performance computing systems"