In this paper, the authors present a polynomial regression model that data drivenly predicts the tendency of ocean temperature using the MITgcm ocean model. The study is well conducted, and technically correct. A lot of well-thought-out experiments have been conducted and presented reasonably well with enough details to facilitate reproducibility. The authors have even considered nuances such as correlated training-testing datasets and did a very comprehensive analysis. Having said all that, it must also be pointed out that the statistical model considered here is overly simple (and shows very limited skill of 0.58) and does not contribute to building better (not interpretable) data-driven forecasting models. The authors do acknowledge that throughout the length of the study and have even put an entire section talking about the limitations. In my opinion, although the study is very comprehensive, the model is too simple, and (as pointed out by the authors themselves) do not actually have the capability to iteratively (autoregressively) perform ocean temperature prediction as has been done (although in the context of atmosphere) in several papers that the authors themselves have cited, nor does it show good skill for even one-time-step prediction. Underneath, I list my major concerns with the current version of this paper and while these concerns do not question any technical point in the paper, the method presented does not quite serve as a "model" for geoscientific predictions and thus may not be suitable for publication in this journal.

1. The authors very thoroughly investigate the effect of different input variables in the one-time-step prediction of the temperature tendency, however, the bulk of the temperature difference between time step t and $t + \Delta t$ is 0. Even for a simple polynomial model, this is probably a simple problem. While the authors claim that there is some skill in predicting high value of ΔT, the plot also shows several such predictions missed by the model. This is not surprising owing to the simplicity of the model. *A correlation coefficient of* 0.58 *in validation* really shows that the model is just not skillful. Doing any analysis on such a model, in my opinion, may not be the best path forward in this field of machine learning/data-driven forecasting of weather/climate. Predicting extremes in the field of weather and climate has been done with data-driven models and authors should probably take a look at those literature, e.g., https://arxiv.org/abs/2103.09743, and https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2019MS001958. The major issue that the authors face in predicting the extremes is essentially addressing the class-imbalance problem which is well studied in the ML literature.

2. The authors quite correctly point out that their model cannot do autoregressive/ iterative prediction thus rendering their model not-useful to some extent. Still, interpretability of these models is a big step forward in this field. However, doing so from looking at predictions at a single time step may not be the right approach. There have been several studies that have shown that error propagation in this model is non-trivial and nonlinear. Thus, data-driven models that iteratively forecast the state of the atmosphere/ocean may show variable skill based on how far it directly forecasts and the error analysis may lead to starkly different results e.g., Figure 2 of this paper (https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2020MS002203) shows the effect of changing the time step of prediction in the data-driven model and so does this paper (https://gmd.copernicus.org/preprints/gmd-2021-71/) . While it is definitely true that a physical variable that actually affects ocean temperature would probably lead to better skill if used as an input, this observation has been reported in the conext of

atmospheric dynamics in a more complex deep learning model (https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2019MS00705) and is quite intuitive to be honest. It is still worth mentioning that the authors do a remarkable job at presenting these analyses which are quite comprehensive in the context of this problem.

I would like to emphasize that the analyses conducted in the study are very comprehensive and would had been much more impactful had it been done for a model that was useful for data-driven prediction. Ofcourse, with an increase in complexity of the model, these analyses would become non-trivial as well. However, at this current stage, the best validation accuracy of the data-driven model is 0.58 which questions the performance of the model especially for a single time step prediction. One of the reasons could be because of under-predicting the extremes which can be dealt in other ways as well.