This manuscript presents a CNN-based approach for predicting the distance of the AMOC from its tipping point. The model uses spatial fields as input, including sea surface temperature (SST), sea surface salinity (SSS), their combination, and the full-depth salinity profile, to predict a normalized distance-to-tipping metric ranging from 0 to 1. The authors futher employ SHAP analysis to identify the features that contribute most strongly the model predictions. In addition, the effort to transfer information learned from CLIMBER-X simulations to CESM simulations is potentially valuable. The topic is important and interesting. However, several important issues need to be addressed before the manuscript can be considered for publication.

1. The CNN output is defined as a normalized index, which is designed to avoid explicitly providing information about the freshwater forcing rate during training and testing. However, in CLIMBER-X, different types of tipping may occur, including bifurcation-induced, noise-induced, and rate-induced tipping. The current definition of the distance to the tipping point appears to be mainly applicable to bifurcation-induced tipping under deterministic conditions. Therefore, the authors should provide a clearer explanation of the applicability and limitations of this distance to tipping definition. In particular, it would be helpful to clarify whether this definition is intended to characterize only the distance to a bifurcation threshold, or whether it can also meaningfully describe proximity to noise-induced or rate-induced tipping events. This distinction is important because the timing of noise-induced and rate-induced tipping events can be highly stochastic and may not have a simple linear relationship with the freshwater forcing value.
2. In the CLIMBER-X experiments, the LR model performs comparably to, and in some cases even better than, the CNN. This suggests that the relationship between the input variable fields and the target index may be relatively simple, rather than requiring complex nonlinear spatial features learned by the deep neural networks. To better justify the use of CNNs, the authors should include additional baseline models, such as shallow CNNs or other lightweight machine-learning methods. In addition, the training dataset appears to be relatively small for a deep learning approach. The authors should provide a clearer description of the dataset, including the size and construction of

training samples. This is important because temporally adjacent fields are often highly correlated, meaning that the nominal sample size may overestimate the amount of independent information available for training and evaluation. The manuscript also reports a relatively large number of hyperparameter settings across different experiments, which may partly reflect the limited sample size and raises questions about the consistency and robustness of the model configuration.

3. The manuscript emphasizes that the CNN is trained on CLIMBER-X and then generalized to CESM. However, since part of the CESM simulation is used for validation and hyperparameter selection, the model selection procedure has already incorporated information from the target model. Therefore, this experiment should not be presented as a fully independent cross-model extrapolation. The authors should moderate the relevant claims in the abstract and conclusions. In particular, statements such as “reliable estimates” and “generalize across models” should be softened unless additional validation using fully independent target model simulations is provided.
4. The explainability analysis is potentially valuable: however, the SHAP maps appear to be derived from the best performing CNN realizations selected using test set performance. This may introduce selection bias, particularly for CESM, where the model shows substantial variability across realizations.
5. Some typos should be corrected throughout the manuscript. For example, in line 64, “Each of these interacting modules are discretized” should be revised to “Each of these interacting modules is discretized”. In line 413, “a modest SST increase is occurs” should be revised to “a modest SST increase occurs”. There also appears to be an incorrectly formatted reference citation around line 600.