Deep Learning for Climate Models of the Atlantic Ocean

Anton Nikolaev,[1] **Ingo Richter,**[2] **Peter Sadowski**[1*]
[1]Information and Computer Sciences, University of Hawai'i at Mānoa
[2]Japan Agency for Marine-Earth Science and Technology

Abstract

A deep neural network is trained to predict sea surface temperature variations at two important regions of the Atlantic ocean, using 800 years of simulated climate dynamics based on the first-principles physics models. This model is then tested against 60 years of historical data. Our statistical model learns to approximate the physical laws governing the simulation, providing significant improvement over simple statistical forecasts and comparable to most state-of-the-art dynamical/conventional forecast models for a fraction of the computational cost.

Introduction

General circulation models (GCMs) describe the time-evolution of the atmosphere or ocean using mathematical models of fluids and thermodynamics. These models are good at predicting climate variations in the Pacific Ocean such as the El Niño–Southern Oscillation (ENSO), but the same models perform poorly in predicting an analogous climate pattern in the Atlantic Ocean. Indeed one of the most successful approaches to predicting short term (1-6 month) climate variability in the Atlantic is just a "damped persistence" model — i.e. the prediction that the seasonal climate anomaly will remain constant with a regression (damping) towards the mean.

Data-driven machine learning methods take a different approach to climate forecasting. Rather than integrating the physics equations forward in time, machine learning attempts to learn emergent patterns from data, sacrificing the interpretability and robustness of first principles in favor of black-box statistical models. When trained on real data, these models could capture deficiencies in the physical models. When trained on simulation data, they can provide a fast approximation to computationally-expensive simulations. Deep learning with artificial neural networks, a machine learning approach that is particularly well-suited for high-dimensional data, has recently shown promise in modeling a variety of fluid flow processes (Wang et al. 2019;

de Bezenac, Pajot, and Gallinari 2019; Ham, Kim, and Luo 2019).
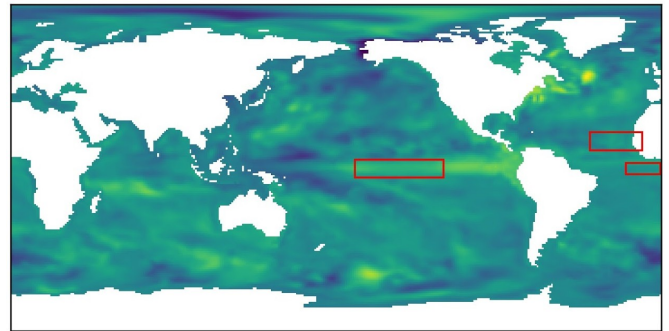

Figure 1: Example heat map of sea surface temperature anomalies for a single month. Physics-based models can forecast variations due to ENSO in the Pacific (center region), but perform poorly in two regions of interest in the Atlantic:: the Northern Tropical Atlantic (NTA; 40°W-10°W, 10°N-20°N) and the eastern equitorial Atlantic (ATL3; 20°W-0°, 3°S-3°N)

In this work we apply deep learning to the challenging task of predicting sea surface temperature (SST) anomalies in two particular regions of the Atlantic (Figure 1) where GCMs are known to perform relatively poorly: the eastern equatorial Atlantic (ATL3), which is subject to pronounced warm and cold events lasting 3-6 months, and the northern tropical Atlantic (NTA). Deep learning methods require large data sets for training, and we use simulated climate processes from Version 2 of the Canadian Earth System Model (CanESM2). The dynamical core of this climate model is based on the first principles Navier-Stokes equations for fluid dynamics, with some unresolved processes such as convection and turbulence represented through parameterization schemes. The latter introduces a few free parameters that are tuned to observational data. This tuning, however, only concerns the mean statistics of the model output and does not provide any information that would allow the model to forecast particular climate events. Running this model forward in time produces simulated climate cycles

---

*peter.sadowski@hawaii.edu

that demonstrate a range of fluctuations under steady radiative forcing. We use this to test the hypothesis that a deep learning model trained on GCM simulations can provide a fast approximation to GCM-based forecasts, and whether such a model performs better than simple persistence forecast models.

Methods

Data

The training data consists of an 800 year time series from CanESM2 simulations, represented as a sequence of one month time steps. The first 600 years are used for training, years 601-700 are used for early stopping and hyper-parameters tuning, and years 701-800 are used as a clean test set for evaluation. After hyper-parameter optimization a final model is trained on the first 700 years with the final 100 years used for early stopping, and we evaluate performance on historical SST anomaly data from years 1958-2017, pre-processed by subtracting the linear climate change trend line.

The CanESM2 data is represented by a grid of 128 longitudinal steps (ranging from $180°$W-$180°$E) and 22 latitudinal steps (ranging from $30°$S-$30°$N). A mask is applied to cells that do not consist entirely of open ocean. For each unmasked cell we have sea surface temperature anomaly, surface wind stress decomposed into longitudinal and latitudinal components τ_u and τ_v, and the depth of the 20 degree Celsius isotherm $z20$, which essentially measures the upper ocean heat content. The data is normalized by mean-subtraction and scaling by the standard deviation, with the mean and variance of each feature calculated over all grid cells over the entire data set. For masked cells the values of all input features are filled with zeroes; predictions at these cells do not contribute to the loss.

Deep Learning

The deep learning approach can leverage *global* information to predict SST at any particular location. However, limiting the information to a local region is advantageous because it helps prevent overfitting. The size of this "receptive field" is something that is optimized during hyper-parameter selection.

In our experiments, a neural network architecture takes in a $(128+k)\times(22+k)\times T\times 4$ tensor, where k is the kernel size and T is the number of months to consider when making predictions. The $T\times 4$ input features at each grid cell are concatenated and treated as input channels. The model consists of a sequence of 2D convolutional layers, with skip connections concatenating the input SST values to the penultimate layer (similar to the widely-used U-net architectures (Ronneberger, Fischer, and Brox 2015)) and adding them to the linear output layer (as in a ResNet (He et al. 2016)). The objective is the Mean Squared Error (MSE) loss computed over the non-masked grid cells.

Hyper-parameters were optimized using the Bayesian Optimisation algorithm implemented in the SHERPA black-box optimization framework (Hertel et al. 2018). A total of 400 neural networks were trained, optimizing over the

Hyper-parameter	Range	Best
Input timesteps	1 - 12	2
Convolution layers	4 - 12	12
Convolution channels	20 - 160	49
Kernel shape	(3,3), (5,5), (7,7)	(3,3)
Initial learning rate	0.0001 - 0.1	0.0001
Batch size	1, 2, 4	1
Early stopping patience	1 - 10	10

Table 1: Hyper-parameter search space explored and optimal values selected by SHERPA after training 100 neural networks using Bayesian Optimization. Kernel shape and batch size were treated as categorical variables with the choices listed.

search space shown in Table 1. The best model consisted of the maximum number of hidden layers (twelve) in our hyper-parameter search space. Many of the models overfit to the data set, and regularization was important — the best model used a small batch size, a small kernel size, a small number of timesteps to consider in the input, and a small channel size. We tried four other modifications that did not improve the performance on the GCM validation set, so were not used in the final model: (1) using locally-connected layers instead of convolutional layers; (2) passing the landmass mask as an input instead of zero-filling; (3) including the month as an extra input channel; (4) dropping the $z20$ input channel.

Results

The model is trained to make predictions for the entire CanESM2 grid, but we focus our analysis on the NTA and the ATL3 regions. In order to evaluate the generalization from simulation to observed data, we evaluate performance on both (1) the final 100 years of the CanESM2 simulation, and (2) the de-trended historical data. In both test sets and both regions, the NN predictions beat the persistence model for lead times of 1-6 months (Figure 2). There is a significant increase in RMSE when transferring the model from the simulation data it was trained on to the historical data, confirming that the simulations are only an imperfect approximation to the real system, but the NN maintains its performance advantage.

| | NTA | | ATL3 | |
	GCM	Historical	GCM	Historical
Persistence	0.27	0.50	0.35	0.51
Deep learning	0.23	0.41	0.26	0.43

Table 2: RMSE on the GCM simulation and historical test sets, in degrees Celsius, averaged over a six-month lead time.

In the NTA, the NN predictions also beat the *damped* persistence model on both test sets. Figure 2 breaks down the performance on the 1958-2017 historical data by lead time for predictions made on February 1st of each year, showing that the forecasting ability degrades with longer lead times

ACC of NTA sst_anom, lead 1 thru 6
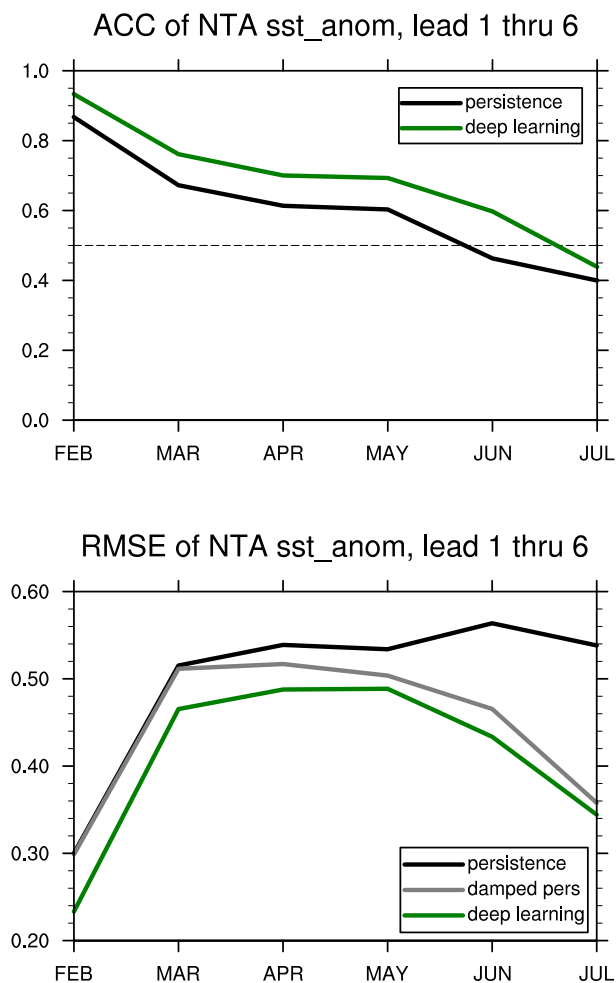

RMSE of NTA sst_anom, lead 1 thru 6

Figure 2: Anomaly correlation coefficient (ACC) and RMSE in degrees Celsius of deep learning approach vs. persistence model and damped persistence model on the Northern Tropical Atlantic from 1958-2017. The climate change trend has been removed from the data. All predictions are made from February 1st.

(i.e. farther into the future). However, the NN is no better than the damped persistence approach on the historical ATL3 data (Figure 3), reflecting the challenge in modeling this region.

Figure 4 compares the sea-surface temperature prediction skill of the NTA model with a range of other approaches. In addition to the persistence forecast, we compare to a linear inverse model (LIM) and GCM-based predictions. Linear inverse modeling is a technique that assumes that the evolution of a system can be approximated by a linear operator with white noise forcing. In practice, the linear operator is typically calculated in principal component space using multivariate regression at a fixed time lag (Penland and Sardeshmukh 1995). LIMs are usually derived from observational data but here we use a LIM derived from the output of the CanESM2 GCM. The other forecast models are GCM based, i.e. they use complex atmosphere-ocean models initialized

with observations to predict the evolution of the system. The GCM forecast models include the SINTEX-F, a prediction model used at the Japan Agency for Marine-Earth Science and Technology (Luo et al. 2005), and 8 models from various forecast centers that participated in the Climate-system Historical Forecast Project (Tompkins et al. 2017); see also (Kirtman and Pirani 2009). These GCM forecast models were selected to illustrate the performance of complex prediction systems. The performance of the NN is competitive with these state-of-the-art methods.

Conclusion

We demonstrate the use of deep learning for forecasting monthly sea surface temperature variations in the Atlantic Ocean with a lead time of 1-6 months, a problem known to be significantly harder than forecasting the ENSO in the Pacific. Training on CanESM2 climate model data and testing on historical data, the deep learning approach performs as well as the best GCM physics models on the northern tropical atlantic region with much less computation. However, on the equatorial Atlantic, our model does no better than a simple damped persistence model.

In this work we restricted ourselves to only training on GCM simulation data at a fixed grid size, and thus we only expect the model to perform as well as the simulation it was trained on. We expect the NN approach to do *better* if it is given a chance to learn from historical data, since then it could learn to correct for deficiencies in the GCM. Fine tuning the model on historical data is an opportunity for future work, although there is a significant danger of overfitting given the limited amount of historical data.

Acknowledgments

References

de Bezenac, E.; Pajot, A.; and Gallinari, P. 2019. Deep learning for physical processes: Incorporating prior scientific knowledge. *Journal of Statistical Mechanics: Theory and Experiment* 2019(12):124009.

Ham, Y.-G.; Kim, J.-H.; and Luo, J.-J. 2019. Deep learning for multi-year enso forecasts. *Nature* 573(7775):568–572.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

ACC of ATL3 sst_anom, lead 1 thru 6
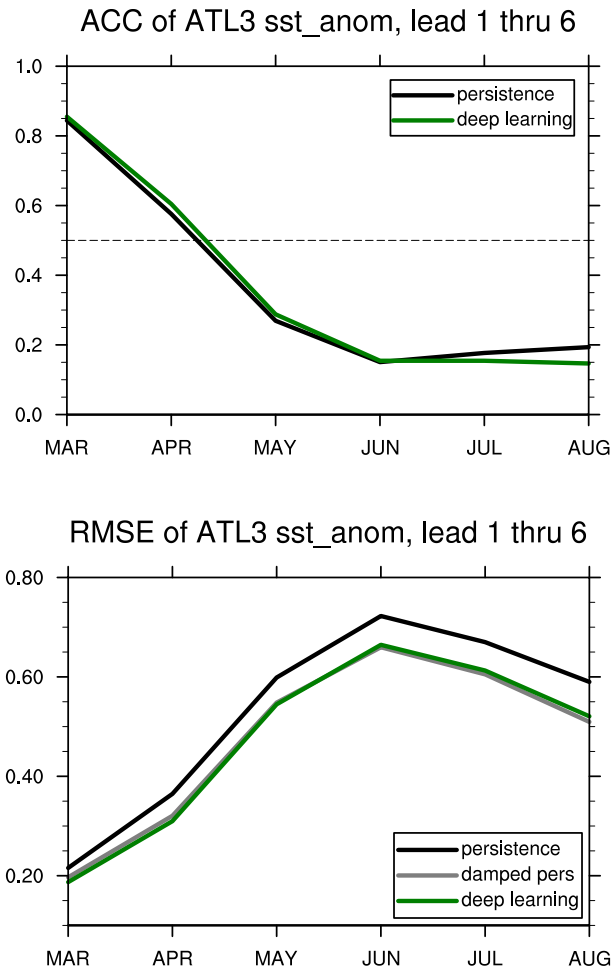
RMSE of ATL3 sst_anom, lead 1 thru 6

Figure 3: Anomaly correlation coefficient (ACC) and RMSE in degrees Celsius of deep learning approach vs. persistence model and damped persistence model on the equitorial Atlantic from 1958-2017. The climate change trend has been removed from the data. All predictions are made from February 1st.

Hertel, L.; Collado, J.; Sadowski, P.; and Baldi, P. 2018. Sherpa: Hyperparameter optimization for machine learning models.

Kirtman, B., and Pirani, A. 2009. The state of the art of seasonal prediction: Outcomes and recommendations from the first world climate research program workshop on seasonal prediction. *Bulletin of the American Meteorological Society*.

Luo, J.-J.; Masson, S.; Behera, S.; Shingu, S.; and Yamagata, T. 2005. Seasonal climate predictability in a coupled oagcm using a different approach for ensemble forecasts. *Journal of climate* 18(21):4474–4497.

Penland, C., and Sardeshmukh, P. D. 1995. The optimal growth of tropical sea surface temperature anomalies. *Journal of climate* 8(8):1999–2024.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Tompkins, A. M.; Ortiz De Zárate, M. I.; Saurral, R. I.; Vera, C.; Saulo, C.; Merryfield, W. J.; Sigmond, M.; Lee, W.-S.; Baehr, J.; Braun, A.; et al. 2017. The climate-system historical forecast project: Providing open access to seasonal forecast ensembles from centers around the globe. *Bulletin of the American Meteorological Society* 98(11):2293–2301.

Wang, R.; Kashinath, K.; Mustafa, M.; Albert, A.; and Yu, R. 2019. Towards physics-informed deep learning for turbulent flow prediction. *arXiv preprint arXiv:1911.08655*.
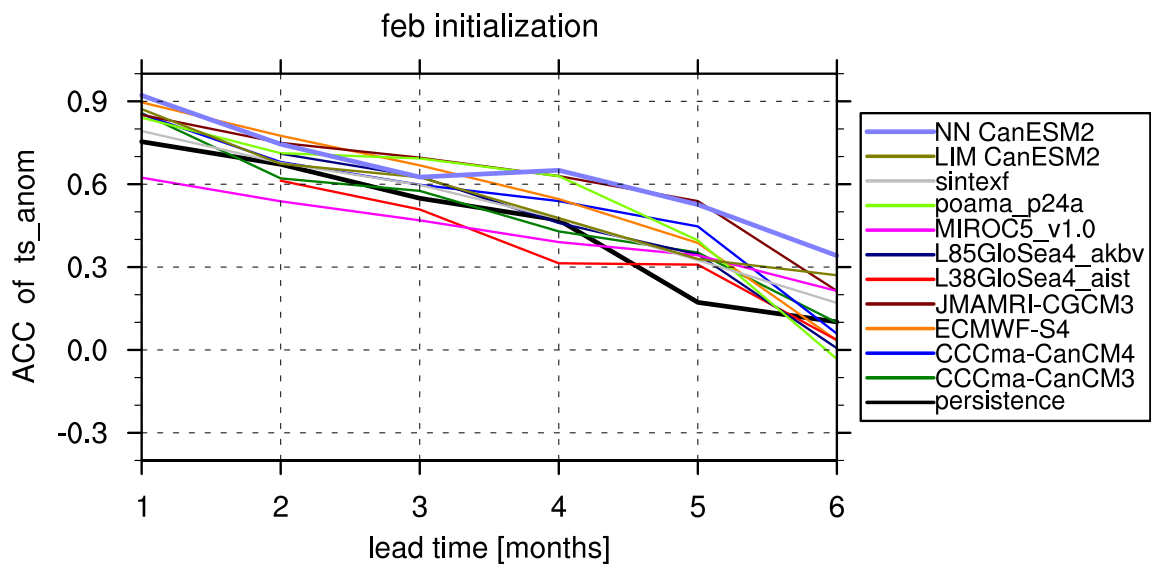
Figure 4: Comparison of deep neural network model (NN CanEMS2, blue line) and dynamical models from the CHFP inter-comparison for years 1982-2010 in the NTA region. The performance of the linear inverse model trained on CanESM2 output (LIM CanESM2) is shown in yellow-green. All predictions are initialized on February 1. The metric shown is the anomaly correlation coefficient (ACC), which is calculated as the correlation of predicted and observed deviations (anomalies) from the mean climate.