Towards a Regional AI-Driven Digital Twin Forecast Model

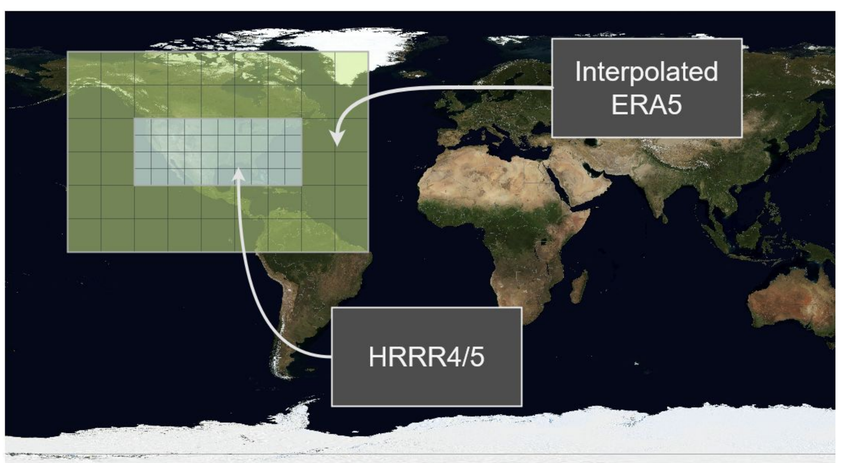
Sophia Hamer[1], Milton Halem[1], Jennifer Sleeman[1,2]
[1] University of Maryland Baltimore County, [2] Johns Hopkins University, Applied Physics Lab

Introduction

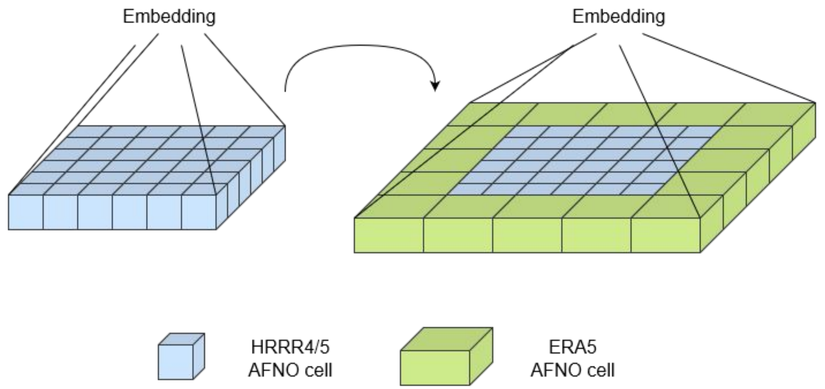
We propose a modification to nVidia's FourCastNet weather forecasting model to enable forecasting over specific regions instead of the entire globe. This enables the use of higher resolution regional data instead of the global ERA5 data used in its original model. In the case of HRRR4/5 data over CONUS, this brings the resolution from 30 km to 3 km

Background

FourCastNet divides the globe into a grid pattern with each cell containing several layers of weather data (wind speed, etc). For each time step, it applies an Fourier transform to each cell and predicts what the Fourier coefficients are in the next time step and then applies a convolutional layer for mixing. Because it predicts over the entire globe, it does not consider boundary conditions at the edge of the grid, which a regional model would need.



Approach



We plan to incorporate nesting to generate boundary data for regional FCN models. A global forecast model is first run to generate lower resolution data outside the boundaries of the regional model, then the regional model is run using the interpolated lower resolution predictions/data as boundary conditions for embedding.

Hardware

FCN was trained on a cluster of 64 nVidia A100s with 64 GB VRAM each, but we have reduced the amount of resources required for training on the ERA5 dataset to less than 16GB of VRAM per GPU. The HRRR4/5 dataset is of significantly higher resolution, but covers a smaller area. In order for our embedding approach to work, we use a subset of variables of the HRRR4/5 dataset equal to the variables used in FCN's pretrained ERA5 dataset. The memory cost is almost identical (<1GB difference), so any system capable of training on the ERA5 dataset can do so with HRRR4/5.

A Digital Twin Forecast Model

Our project's goal is to train FCN with 10 years of HRRR5 data (2014-2024) with hourly forecast timesteps at 3km resolution as part of a digital twin framework that includes integration with NUWRF and SFire under the Fire-tech grant. Further goals also include the integration of CHEM, producing a complete regional forecast model driven by satellite, aircraft, and ground based measurement stations. Due to the relative speed of running inference on a trained model (about 20 minutes for a year-long forecast), forecasts can be produced and updated quickly with new data.

Mobile Forecasting

Inference is relatively cheap to calculate in AI-driven forecast models - initial experiments with FCN were done using a laptop without GPU acceleration, which was able to generate a weeklong forecast in about 40 minutes. Laptop GPUs with sufficient VRAM are able to run inference using the graphics hardware in even less time. We intend to develop a mobile oriented inference model powered by FCN that can be run in the field. Our proposal involves training a lower resolution model and using lower precision data representations to push the inference model to smaller memory footprint and lower computational requirements. Ideally, we want to explore running the inference engine on a tablet or mobile device than can perform OSSE experiments in the field.