Contractor's Name and Address:

Systems & Technology Research
600 West Cummings Park
Woburn MA 01801

Program Name:

Gaia (Global AI Accelerator)

Milestone:

Month 2 Milestone Report: Develop framework for the hybrid models identifying the known and unknown parts along with mathematical approaches

Date of Report:

February 13, 2022

Gaia Description and Overall Goals (from M1 report)

Gaia aims to develop new hybrid AI tools and methods to accelerate Global Climate Models (GCMs) by replacing the standard parametric cloud-resolving physics models with improved Artificial Intelligence (AI) surrogate models that better capture local convection aggregation. By training the AI surrogate off of both GCMs (using CAM and SPCAM cloud physics parameterizations) and a Weather Research and Forecasting - Large-Eddy Simulation (WRF-LES) turbulence-resolving weather model, Gaia will concentrate on three specific subgoals:

a) improve the overall computational efficiency of the GCM;
b) improve its ability to accurately predict self-organizing atmospheric wave phenomena;
c) exploit this improved model to explore previously unobserved regimes and to identify early warning signatures of large scale changes to such self-organizing phenomena (our interpretation of Tipping Points) that could have large downstream global weather and climate impacts.

Gaia specifically focuses on modeling the eastward-moving cloud structures known as Madden Julian Oscillations (MJO) and learning predictive signatures for tipping point changes in these structures and for other convective organizations. For example, it has been hypothesized that sufficient CO_2 forcing could cause the MJO to transition to a "super MJO" in which tropical rainfall aggregates into a large mass and easterly winds weaken or even reverse, with large adverse impacts on ecosystems and populations.

More generally, Gaia exemplifies a systems approach to applying AI and Machine Learning (ML) methods to construct hybrid models with sufficient fidelity to test specific scientific hypotheses and accelerate scientific discovery near the limits of available/affordable computation.

M2 Update: AI Hybrid Model & Initial Training

Figure 1 (from M1 report) shows the basic Gaia framework, where an AI surrogate is progressively trained using training datasets from multiple cloud physics models (CAM4, SPCAM, WRF-LES). The physics model takes state data from 26 altitude layers (humidity, relative humidity, temperature, and wind in zonal U and meridional V directions) for every latitude-longitude voxel and for every time step, and predicts the humidity and temperature tendencies (rate of change) along with precipitation. Much of our architectural framework was completed in January, so this M2 report will focus on our current and rapid progress in generating training data and testing an initial AI model.

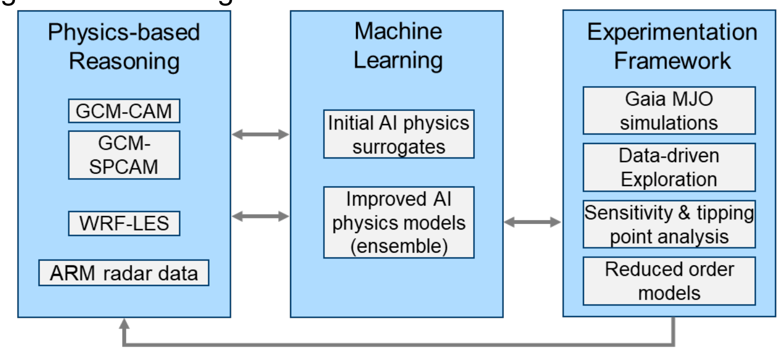

Figure 1. Overall Gaia Model and Integration Architecture

We note that a key assumption in our approach rests on the same assumption that physics-parameterized global climate models rely upon, i.e. that there is sufficient state information captured in the physics parameterization to be usefully predictive, and that this model generalizes across both geography of the voxels and the season. Based on this assumption of universality, our AI strategy is to build up datasets using multi-year runs of GCM-CAM (and later, WRF-LES), and randomly subsample voxels in space and time to obtain quasi-independent input-output pairs for training and validation purposes.

We initially start with a vanilla deep neural network architecture (fully connected, seven layer, 512 neurons wide, using LeakyRelu activations, BatchNorm and Dropout (2) regularization). Input variables are normalized to 0 mean, 1 variance, and output variables are normalized by variance only. The current DNN network has 1.4e6 parameters.

Our initial dataset consists of ~ 1 year of GCM-CAM4 simulations, with data pulled out at 30 minute timesteps over the course of 334 days. The grid is a 2.5-degree grid, 96 latitude x 144 longitude. This gives us 2.2e8 space-time voxels. Input vector length was 130 (26 levels x 5 parameters); output vector length was 53 (26 x 2 tendencies plus precipitation).

Table 1 summarizes the current training approach.

Data	**Metrics**	**Comments**
Timesteps	334 day x 48 timesteps per day	
Grid points	96 latitude by 144 longitude	
Input	26 levels x 5 params	Humidity, rel. humidity, temp, wind in u and v directions
Output	26 levels x 2 tendencies + 1	Humidity, temp tendencies + precipitation
Test interval	Last 3 days of every 30 days = 10% of data	Subsample by 8
Training interval	= Not_test	
Model		
FCN w LeakyRelu	7 layer x 512; 1.4e6 parameters	with BatchNorm, Dropout
Normalization	Variance = 1	
Training		

Optimization	ADAM w. 1e-3 learning rate	
Batch	24x96x144 (I/O) pairs	
Convergence	Convergence after 500 epochs	
Train time	1 epoch ~ 30 sec	(30 x 8 sec without subsampling)
Loss	MSE	Skill = 1-MSE/Var

Table 1. AI Model Parameters

At this stage of our project we are ahead of plan, and initial model results are encouraging. Based on the one year of simulated training data, we demonstrate useful AI predictive skill, shown in Figure 2a for PTTEND (temperature tendency) and in Figure 2b for the least-well predicted variable PTEQ (humidity tendency). Prediction skill for temperature remains fairly good for most altitude levels. Prediction skill for humidity tendency diminishes in the lower latitudes and highest altitude levels, although it must be noted that at these high altitudes the humidity and humidity tendency are close to zero and there is significant computational noise in the simulated data.

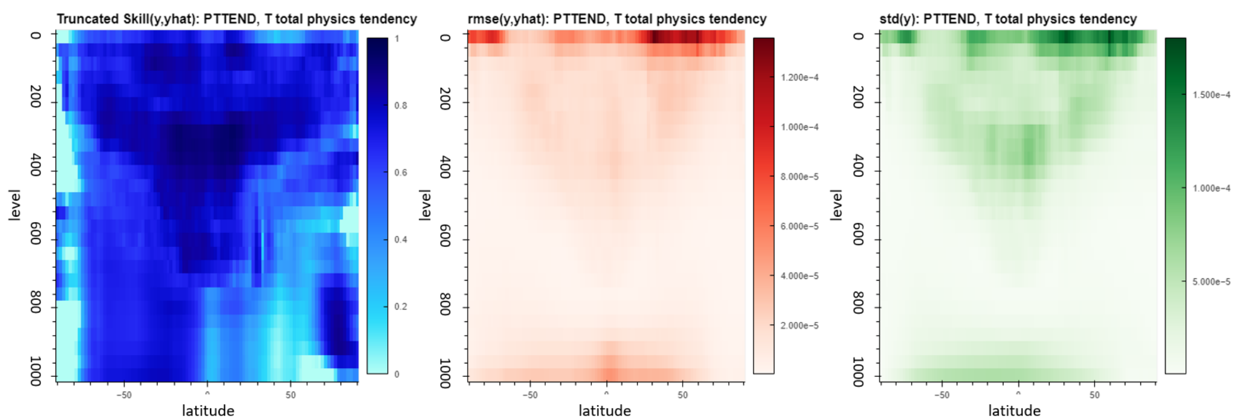

Figure 2a. AI model skill, MSE and STD in predicting temperature tendency vs altitude level and latitude (aggregated over longitudinal coordinates)

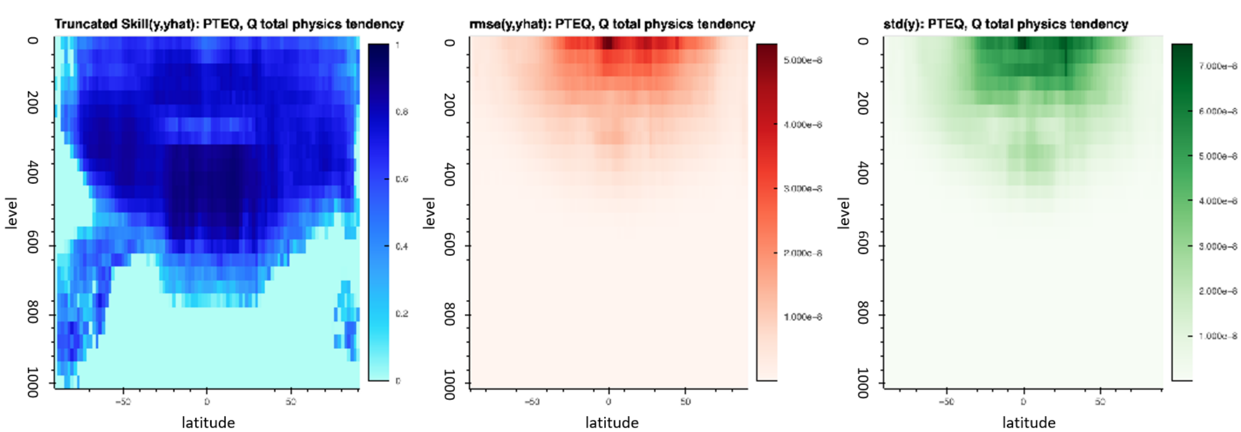

Figure 2b. AI model skill, MSE and STD in predicting humidity tendency vs. altitude level and latitude

Figure 3 shows humidity tendency timesteps (both training data and prediction), showing very good predictive skill for humidity tendency at lower altitudes but noisy predictions at high altitude. It’s possible that the CAM physics model itself (our ground truth) has limited predictive skill in these regions other than to predict zero, and that our model is simply learning computational noise. But the high variance in our AI prediction points towards a normalization or data sampling bias problem with the model not fully understood. The remainder of this M2 report focuses on our attempts to understand and deal with the problem.

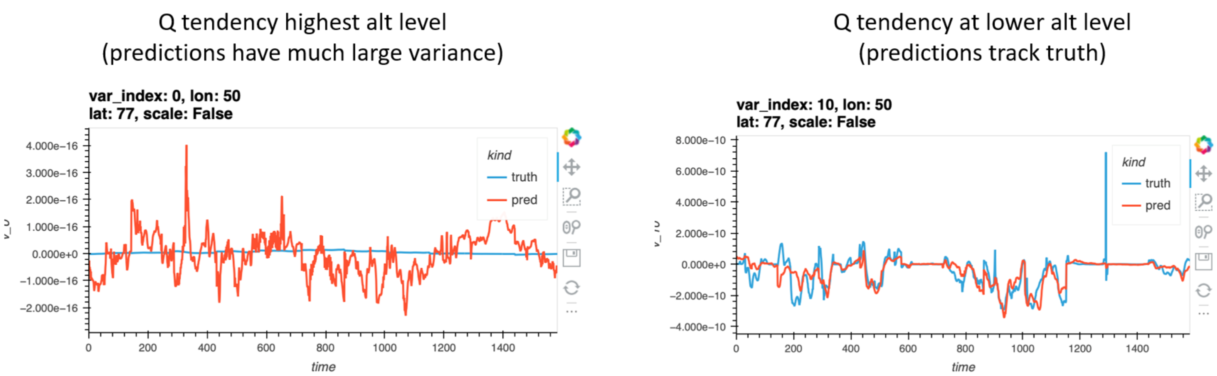

Figure 3. Prediction waveforms for PTEQ (humidity tendency) at highest and lowest altitude levels

Two more charts further illustrate the problem. Figure 4 shows the generalization ability (test skill/training skill) of the model in predicting PTEQ (humidity tendency) for the 26 levels, again showing the problem with the highest altitude levels. Figure 5 shows the ratio of test-to-training variance for Q, also highlighting a potential sampling bias resulting in substantially different training and test variances at high altitudes where the tendencies are near zero.

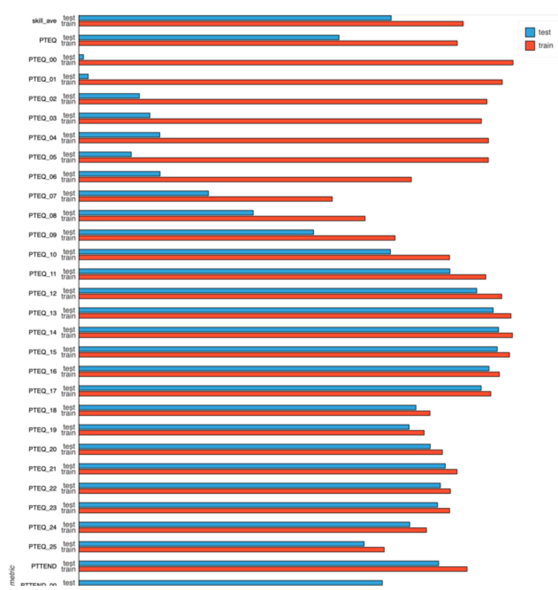

Figure 4. Ratio of PTEQ test/train prediction skill as function of level (levels 0 through 10 correspond to the highest altitude levels and are most problematic for prediction skill)

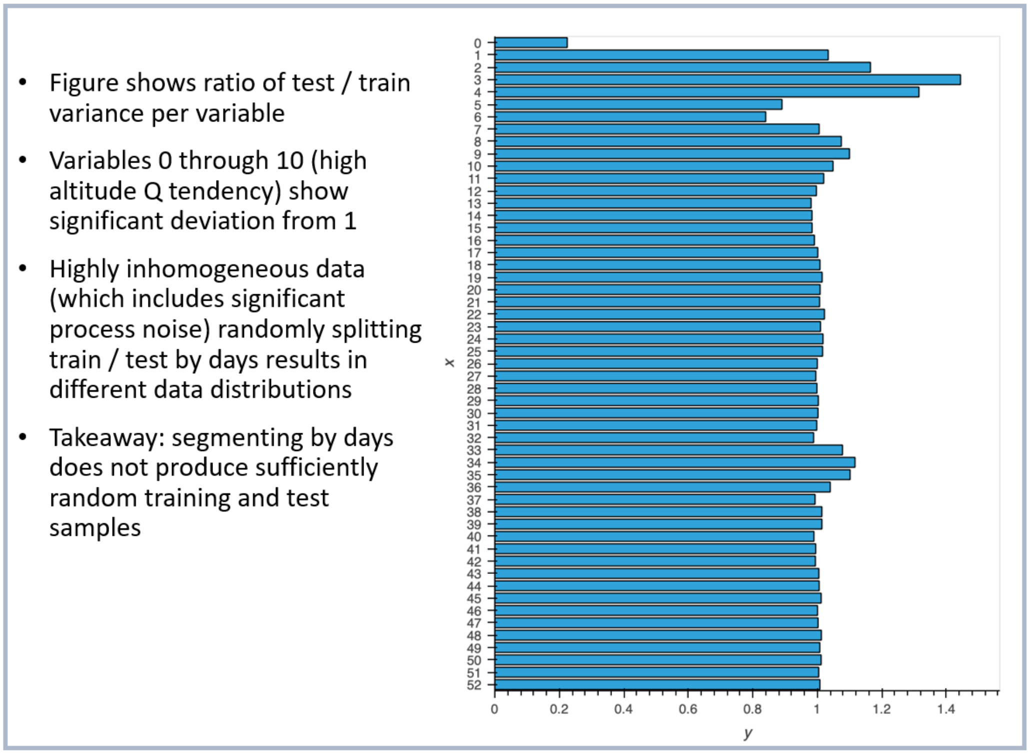

Figure 5. Ratio of test/train PTEQ variance per variable as a function of level

In order to confirm the hypothesis of insufficient randomization in our day-by-day sampling, we trained a new AI model that attempts to achieve predictive skill for the five highest altitudes (levels 0 through 4). Randomly splitting test and training data on day boundaries continued to produce poor test validation scores, as shown in Figure 6. However, randomly pulling timesteps for test and training data pairs without segmenting on day boundaries produced much better skill, possibly indicating sample correlation issues in the limited test dataset.

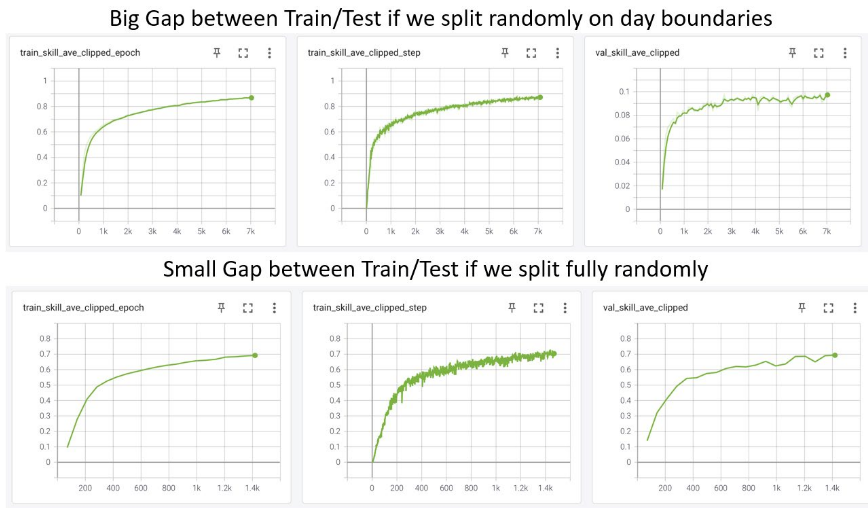

Figure 6. Model trained to predict the five highest levels of Q (humidity), showing effect of randomizing test and train samples on day boundaries

This sampling issue remains puzzling and warrants further investigation, although we expect that adding in the larger simulation runs may alleviate the problem.

For these preliminary data studies, we performed one final check. In order to estimate the intrinsic dimensionality of the model, we ran a spectral PCA (principal components analysis) on the input and output variables. Recall that our input variables form a vector of 130 parameters, and output variables another 53 parameters. PCA was performed looking at both the space of input variables, and the concatenated space of input plus output variables. The results of this analysis are shown in Figure 7, showing a sharp knee at about 100 variables, which one might take as a measure of the intrinsic dimension of the model.

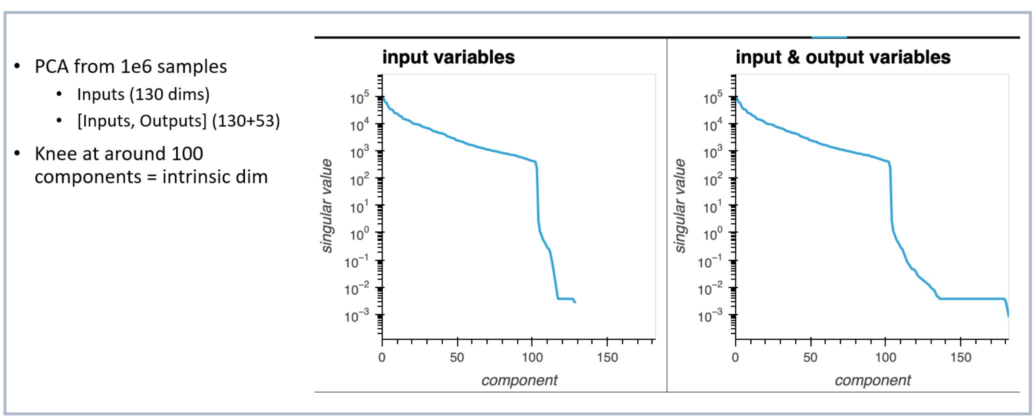

Figure 7. Spectral PCA dimensionality analysis of the data size

Next Steps / Planned Activities

Over the next period we will continue to generate additional training (roughly 10 years of CAM and SPCAM GCM model simulations). We have already completed four years of simulation using SPCAM with CLUBB (unified cloud parameterization) and Morrison microphysics, having a 20-minute timestep and 16 SAM columns, and another four years of simulation using SPCAM with Morrison microphysics but no CLUBB. We note that the GCM with CAM or SPCAM include additional variables that will be incorporated into the AI surrogate, e.g.:

- Sea Level Pressure, SLP
- Solar Insolation, SOLIN
- Sensible Heat Flux, SHFLX
- Surface Latent Heat Flux, LHFLX
- Net Solar Flux at Surface, FSNS
- Net Longwave Flux at Surface, FLNS
- Net Solar Flux at the Top of Model, FSNT
- Net Longwave Flux at the Top of Model, FLNT

While completing the ten year simulations, we will continue to assess our AI models using these additional parameters and the already-simulated four-year datasets.

Two potential issues arise with the augmented training: (1) we will require additional storage for the new data sets (roughly 4TB, for which remote hosting on the cloud will cost ~ $100/mo); (2) we may run into computational bottlenecks and would be interested to explore AWS access since some of the ACTM performer base is already partnering with Amazon on climate modeling.

In addition to revised randomization and test holdout strategies, we plan several additional investigations towards AI surrogate optimization:

- Experiment with alternative re-weighting/normalization schemes for handling high altitude, low latitude PTEQ tendencies.
- Embed additional structure within the AI networks to exploit spatial structure of variables at different altitudes.
- Assess the value of adding output history from the previous timestep; use various spectral analyses and manifold learning approaches to determine the most compact set of history variables required to improve prediction skill and generalizability; to accommodate the previous timestep memory we propose to subsample model outputs, e.g. using timesteps for every 3 hours rather than every 20 minutes, in order to limit file sizes.
- Post-process the SPCAM runs to assess the differences in simulated fields due to the CLUBB parameterization.
- Investigate nonlinear (as opposed to PCA) techniques to quantify model dimensionality, e.g. using diffusion map embedding https://datafold-dev.gitlab.io/datafold/tutorial_03_basic_dmap_scurve.html