Gaia (Global AI Accelerator) Month 1 Milestone Report

January 13, 2022

Gaia Team: STR, University of New South Wales (with Professor Steven Sherwood), & Professor Yannis Kevrekidis (JHU)

M1 Topic: “Report identifying constituents of the hybrid models along with planned datasets and the problems and targeted effects to be investigated”

Gaia Description and Overall Goals

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 CO2 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.

Model Constituents & Integration

Figure 1 shows the basic Gaia model. The improved AI physics models are learned in an iterative process. First, we run two variants of the GCM with differing atmospheric models (NCAR CAM5 and NCAR SPCAM), spanning a wide range of environmental conditions. These runs will provide a large number of atmospheric physics input-output pairs to initially train a deep network surrogate model with broad coverage across the same range of environmental conditions. Next, we improve this AI surrogate by retraining it on data from curated model runs of a fine-grained WRF-LES model, with boundary conditions constrained to match the GCM. It is expected that the latent representation acquired by the deep network in the first step helps ensure rapid transfer to the WRF-LES trained surrogate, analogous to how broadly trained models like ImageNet-trained Deep Neural Networks (DNNs) learn to recognize new object types from limited new training data.

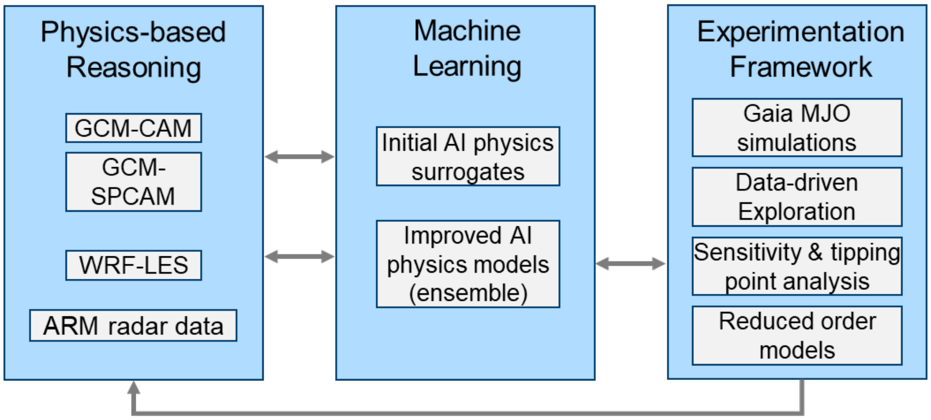


Figure 1. Overall Gaia Model and Integration Architecture

The improved LES-trained physics model then replaces the GCM's original parametric models of atmospheric physics, resulting in the Gaia hybrid model that captures convection aggregation at scales from 200m and up. The Gaia model should also run significantly faster than the original GCM model (and far faster than any WRF-LES model). This improved hybrid model now serves as a basis for experimentation to characterize MJO tipping points. Given the time series information from these runs, we deploy dynamical systems-based techniques (coarse-grained bifurcation detection) as well as manifold learning techniques (for the parameterization of coarse basin boundaries) to identify high value environmental regimes for further model exploration. Having identified these high value regimes, GCM and LES models can be re-invoked to generate additional training data to help ensure Gaia accuracy overall, and especially as we edge closer to a tipping point.

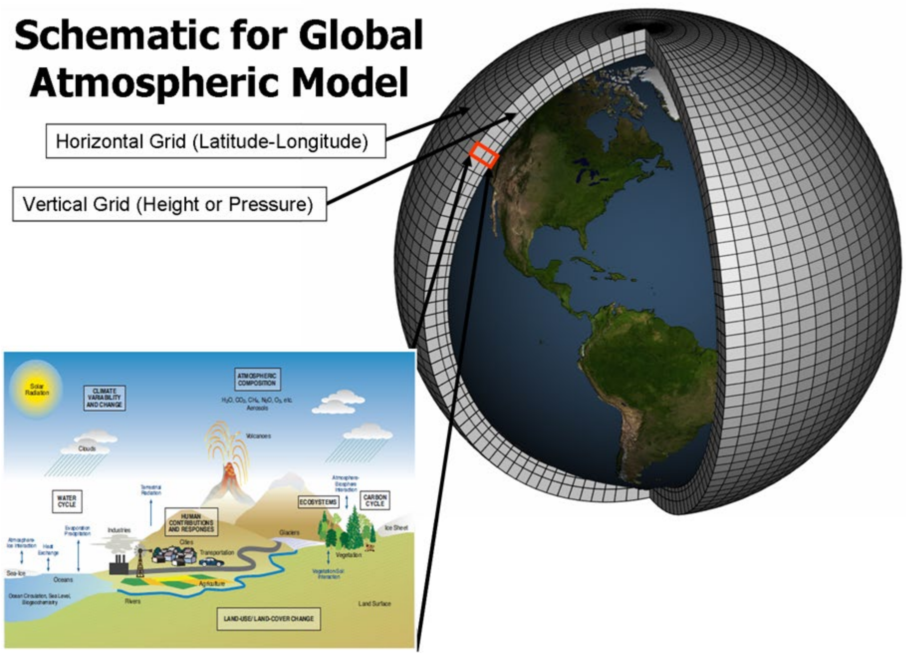

Schematic for Global Atmospheric Model
Horizontal Grid (Latitude-Longitude)
Vertical Grid (Height or Pressure)
Solar Radiation
CLIMATE VARIABILITY AND CHANGE
ATMOSPHERIC COMPOSITION
H₂O, CO₂, CH₄, N₂O, O₃, etc.
Aerosols
Volcanoes
Clouds
WATER CYCLE
Terrestrial Radiation
HUMAN CONTRIBUTIONS AND RESPONSES
ECOSYSTEMS
Atmosphere-Biosphere Interaction
CARBON CYCLE
Atmosphere-Ice Interaction
Heat Exchange
Evaporation Precipitation
Industries
Cities
Transportation
Glaciers
Ice Sheet
Vegetation
Sea-Ice
Oceans
Agriculture
Vegetation-Soil Interaction
Ocean Circulation, Sea Level, Biogeochemistry
Rivers
Land Surface
LAND-USE/ LAND-COVER CHANGE

Figure 2. GCM modeling

The initial CAM5 and SPCAM GCM models will be run for a time window of five years, based on both present conditions as well as a forced 4K temperature rise, for a total of ten years simulation time for each GCM model. The physics surrogates predict sub-grid tendencies d**x**/dt from **x** (the mean state within a GCM column). The state vector **x** includes temperature, humidity, and long/lat wind conditions. A full table of input-output variables is shown in Table 1. In contrast to current GCM physics models, we propose to retain convective systems memory by also including d**x**/dt from the previous GCM time step in the state description. The GCMs will generate roughly 500 million I/O training pairs per GCM, which will be down-selected by a factor of ~ 10 for training purposes.

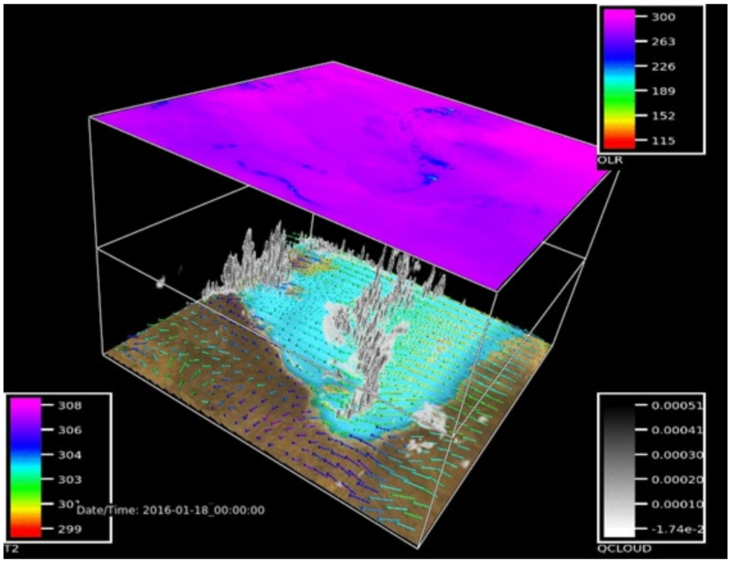


Figure 3. LES Model

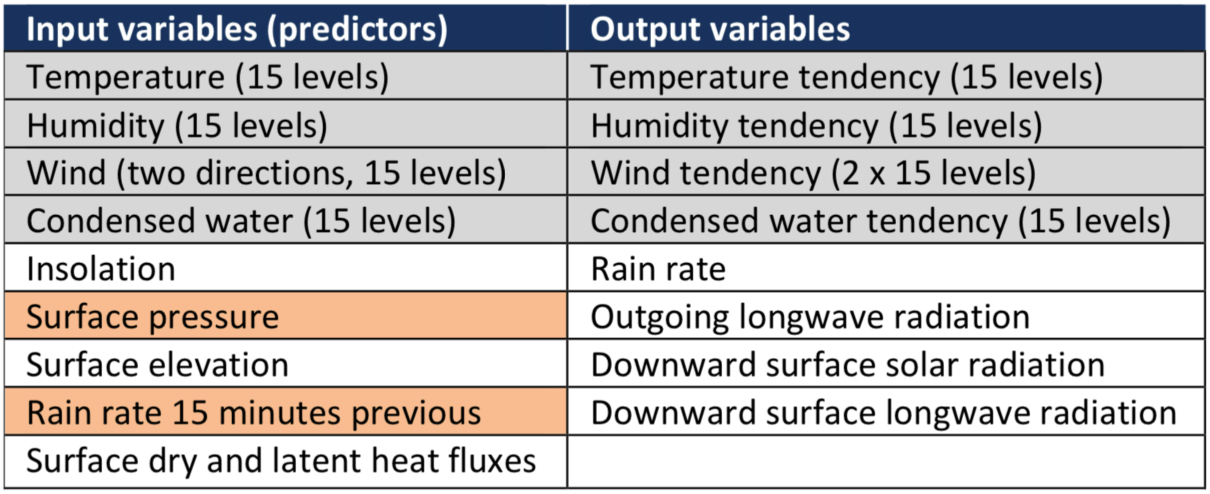
We propose to use deviations between CAM5 and SPCAM models to help in the selection of WRF-LES training instances to run. Each of these runs (up to 2000 training instances total) will be run on 200 m grids within a 20 km x 20 km domain (using periodic boundary conditions). These runs will be spun up by running for one week at 2 km resolution, and then continued for 3 – 5 days at 200 m resolution, initialized with the re-gridded final 2 km state. The 200 m resolution is chosen to capture key boundary-layer dynamics (not captured in prior studies) while still retaining computational affordability, as informed by initial testing at UNSW. For these runs, incoming sunlight, surface elevation, pressure, and fluxes, will be prescribed as time-evolving LES boundary conditions (but not prior rain rate, which will have been predicted by the LES), while the LES sounding will be nudged toward the input sounding to keep the state variables close to the global model (this nudging is a standard feature in WRF and in our approach represents large-scale forcing). The resulting training data will consist of LES-generated inputs and outputs, at an interval matching the CAM5 time step of 15 minutes, horizontally averaged and mapped onto the same vertical grid.

Planned Datasets

Input and output variables generated by the above models (CAM5, SPCAM, and WRF-LES), capture the key drivers and consequences of local convection, and are shown in Table 1. The gray boxes show sounding variables, applied to LES by nudging. Orange-colored variables are of special interest – our expectation is that these variables will especially benefit by adding memory to the model in the form of additional state information at the prior time step. Additional memory representations will also be assessed via traditional, as well as nonlinear data-driven dimensional analysis techniques.

We have modified CAM5 to output these input/output pairs (still in the process of modifying CAM5 and SPCAM to output wind tendencies).

Table 1. Input/Output variables used to train the surrogate DNN model



Input variables (predictors)	Output variables
Temperature (15 levels)	Temperature tendency (15 levels)
Humidity (15 levels)	Humidity tendency (15 levels)
Wind (two directions, 15 levels)	Wind tendency (2 x 15 levels)
Condensed water (15 levels)	Condensed water tendency (15 levels)
Insolation	Rain rate
Surface pressure	Outgoing longwave radiation
Surface elevation	Downward surface solar radiation
Rain rate 15 minutes previous	Downward surface longwave radiation
Surface dry and latent heat fluxes	

Key Challenges and Risks

One potential risk of our approach is that SPCAM and LES models will exhibit different biases (e.g. too warm or too cold at different altitude levels) that may be nontrivial relative to the system variability. To mitigate this risk, we will obtain ERA5 reanalysis and satellite data from the same time period to assess bias severity, and if necessary, explore methods for adaptive bias correction in our training.

We are depending on the SPCAM structural behavior to be good enough to be “corrected” using O(1000) LES retraining points. This may place additional constraints on the types of DNN architectures to be employed, and will likely require the enforcing of additional physical constraints (e.g. conservation of energy) in the architectures as well.

The proposed Gaia hybrid model is being optimized for tropical convection and tropical midlatitude interactions key to the targeted tipping point analysis, so its unlikely that it will generalize to non-tropical environments without additional training.

A key challenge is that we may have little ground truth existing near a potential tipping point. We believe our targeted and interactive data generation/sampling will help mitigate this challenge.