Big Data Challenges at the Community Coordinated Modeling Center (CCMC)

Tina Tsui[1], Polymnia Manessis[2], Matthew Lesko[3], Damian Barrous-Dume[3], Karen Scheiber[2], Edgar Russell[4], Philip Poole[2], Maria Kuznetsova[1], Leila Mays[1]
(1) NASA GSFC, Community Coordinated Modeling Center, Greenbelt MD, United States, (2) Community Coordinated Modeling Center, NASA GSFC, Adnet, Greenbelt MD, United States, (3) Community Coordinated Modeling Center, NASA GSFC, Navteca, Greenbelt MD, United States, (4) Community Coordinated Modeling Center, NASA GSFC, Eclipse Technical Systems, Greenbelt MD, United States

Abstract

Like other research centers, the Community Coordinated Modeling Center (CCMC, https://ccmc.gsfc.nasa.gov) at NASA Goddard Space Flight Center (GSFC) is also experiencing the big data challenges. CCMC hosts over 80 space weather models for Runs On Request (ROR), Continuous Runs and Instant Runs simulation services for the research community. In addition, CCMC has started to support simulation output onboarding in response to the Open Science initiative. Overall, we have accumulated over petabytes of simulation output data and are rapidly growing.

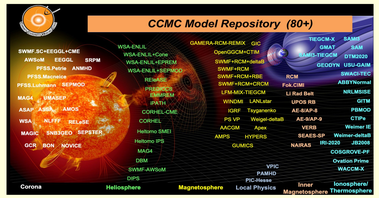
In this presentation, we will discuss our data and storage challenges. We will present our attempts to address our challenges and any associated lessons learned. CCMC uses Apache Airflow to ensure data transfer is consistent. We will give a brief overview on how we leverage Apache Airflow to enhance our environment.

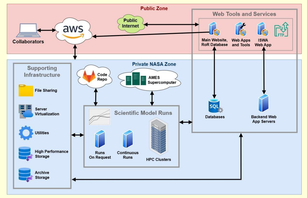
About the CCMC

Community Coordinated Modeling Center (CCMC) is a multi-agency partnership to enable, support, and perform the research and development for next generation space science and space weather models.

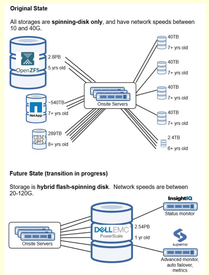
The main goals are:
- facilitate space science and space weather research and model development,
- support development and deployment of new operational space weather capabilities,
- improve quality and usefulness of simulation results and increase scientific return

CCMC hosts more than 80 space science and space weather models developed by the international research community. This unique and expanding collection provides associated Runs On Request (ROR), Continuous Runs and Instant Runs simulation services to any researchers interested in running and evaluating those models.





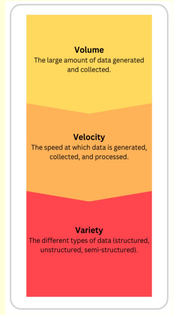
CCMC Hybrid Infrastructure



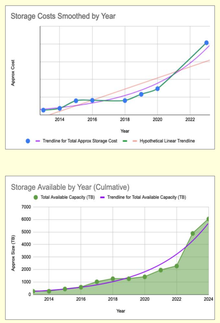
CCMC infrastructure consists of our usage at NASA Ames Supercomputing, Amazon Web Services (AWS) cloud provided by NASA Science Managed Cloud Environment (SMCE), and CCMC owned and operated environments. We have multiple on premise storage devices, with various storage capacities, shareable to all CCMC servers. Since 2023, we have been modernizing our on-site hardware via refreshing existing equipment with more powerful and modern technology. We deployed Dell PowerScale storage to consolidate all local storages.

The Reality of Big Data: Ongoing Challenges We Face

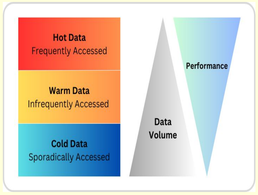
Big Data



More Data, More Costs



Data Tiering



CCMC stores downloads from various observational data and generates model simulations for Runs On Request (ROR), Continuous Runs and model validations. We regularly store simulation output from collaborators for models that's not yet available at CCMC, too. We also started to onboard simulation output from researchers in response to the Open Science initiative. In addition, we are looking into ways to provide CCMC visualization services for model runs that was done by other collaboration groups.

To ensure reproducibility, CCMC keeps all data and **NOTHING** is deleted. As the data volume increase, the storage cost increases, whether its on-premise or in the cloud. For on-premises storage increase requirement, it requires us to purchase additional hardware (expansion vs. new cluster), might need to negotiate additional facility power and/or clear out physical space for the new cluster. As for cloud, little to no provision is needed, just need to pay for the additional storage costs.

Data Tiering to the rescue?

Data tiering refers to organizing and storing based on their category. It involves categorizing data into different tiers based on their importance (example: is this data part of publication?) and usage (example: how often has this data been requested?). Common (industry) tiers categories: hot, warm, and cold; CCMC uses high, medium and low value categories to identify our data. Once identified, it allows us to cost-effectively store the data into the different storage technologies or tiers. We have been discussing various alternative solutions for cold storages. We also have our staff identifying known ROR simulation runs that were used in publications. We still have many uncategorized ROR runs and can't do it alone. We need our scientists and/or the research community to help us identify the dataset so we can reduce the storage cost. Please let us know if you use our dataset/service in your publication!

Overcoming Data Transfer Challenges

Managing multiple file transfers pipeline across different systems and/or environments on a daily basis is an intricate task. We, like other organizations, face the following challenges:

scalability
schedule and scale complex pipeline easily

automation
automatically re-run the pipeline after failure

monitoring/ observability
monitor pipeline using logs and dashboards in real-time

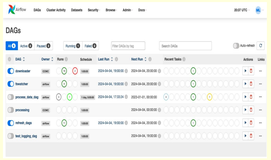
flexibility
change the workflows with minimal effort

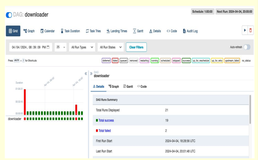
Apache Airflow has empower our data transfer strategy.

Enhancing CCMC Operation with Apache Airflow

Apache Airflow is an open source tool for creating, scheduling and monitoring workflow or pipelines. It provides a central control over data transfers and manage all scheduled and event-based tasks in our hybrid environment, which **ensures reliable transfers**. With the single pane of glass, one can quickly resolves data issues. One can **easily manage essential project tasks in real-time** from any location within our various environments.

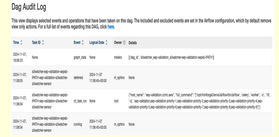

Airflow environment dashboard


list of directed acyclic graph (DAG) in the environment


status of DAG span across time

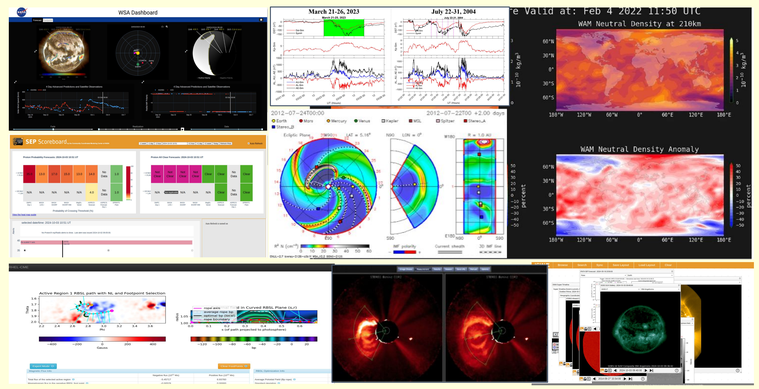

graph view of DAG's dependencies and their current status


code that generates the DAG


log events related to a DAG

Advancing CCMC services

- Enhanced **monitoring and error detection** for improved reliability.
- **Standardized metadata** capture to build a metadata database for tiering and improved DataOps.
- Support for complex workflows and data transformations.
- **Improved data reliability and accessibility** for researchers and forecasters for all of the services we offer.



Simulation Output Onboarding Procedures (https://ccmc.gsfc.nasa.gov/simulation-onboarding/)

For onboarding a model at the CCMC, please see our 'Model Onboarding' instead. These procedures are for folks who want to deliver their **full resolution simulation(s) outputs only** (not the model) to the CCMC to be used by the community via the CCMC visualization services (e.g. interactive web visualization, Kamodo, etc.).

1 Pre-onboarding

If your project is at the proposal stage, we ask you to fill out a short pre-onboarding questionnaire.

Note: Feel free to skip any question related to onboarding a model. There is a specific section on the form for simulation results.

2 Preparation

Originator of the simulation(s) output: when you are ready to deliver the output to the CCMC, please contacts the CCMC to receive and complete this questionnaire [MS Word or PDF]. The model information will be added to the CCMC Metadata Registry, which automatically populate the CCMC Model Catalog on the CCMC website. Even though the model is not hosted at the CCMC, such model information is essential to help users in understanding and using the simulation(s) outputs effectively.

The originator of the outputs agrees to provide model output reader and/or interpolation/visualization routine/source code to the CCMC. Such information/source code is needed to incorporate the simulations outputs into the CCMC visualization services.

3 Implementation

If needed, the originator of the simulation(s) outputs agrees to work collaboratively with the CCMC, which might involve installing/testing/running their provided output reader/interpolation/visualization code on the CCMC AWS Cloud environment.

The CCMC will add the functionality on the CCMC visualization services allowing users to use and visualize the provided simulation(s) outputs.

4 Testing

During testing phase, the originator of the simulation(s) outputs agrees to test and validate any visualization and/or interpolation results generated by the CCMC visualization services via the provided outputs.

5 Public Release

Once testing phase is complete and with agreement from the originator of the simulation(s) outputs, the CCMC will advertise the availability of the simulation(s) outputs on the CCMC website.

Visit Us!

https://ccmc.gsfc.nasa.gov