**CyberWay Rescheduled Monthly Telecon Minutes**

2:00-3:00 PM Apr 16 2018

website: <http://cube.csiss.gmu.edu/CyberWay>

github: <https://github.com/CSISS/CyberWay>

**1. Roll Call of Participants**

Ben Cash, Juozas Gaigalas, Ziheng Sun, Sheng-hung Wang, Lei Xue, Eugene Yu, Chen Zhang

**2. Agenda**

Discuss the architecture design and sub-group progress of CyberWay.

**3. Subgroup status & action item report**

*3.1 GMU CSISS*

Dr. Ziheng Sun reported the progress of CSISS. In the past month, we started to work on three items:

1) Data searcher: cataloging the data from Thredds Data Servers (TDS) and some other standards (REST, API, OAI-PMH, CSW and OpenSearch) (UCAR RDA supports all of them). The cataloging need several innovations on how to quickly harvesting new records and searching the entire metadata database.

2) Data processing functionality: We are working on wrapping and providing general tools for model intercomparision and validation, such as interpolation, regridding, reformatting, reprojecting the data in different types. We try to make them reusable in the variety of scenarios while dealing with all the already-using or going-to-use data sources.

3) Resource integration: We are trying to settle down a plan to integrate all the datasets, metadata, functionalities into one interface and make it easy and intuitive for scientists to use. The scientists are not supposed to face a brandnew unknown tool, but a familiar and efficient tool. That is what the integration development is all about.

*3.2 OSU*

Dr. Sheng-hung Wang reported the landscape of ASR products in RDA. The historical products are archived on tapes and there is a system to automatically mount and load the data. The header of the ASR product identifier is DS3631. If someone wants to query the data, he has to input the header plus observation date and the system will automatically load the corresponding tape and read the content out. Sheng-hung also mentioned the side-by-side comparison is the mostly-accepted way for result-result or result-observation comparison in science. He provided an experimental website ([link](http://www2.mmm.ucar.edu/rt/amps/)) as reference. CSISS can learn from the system and understand how scientists want to look at the data. Regarding the data transferring, most scientists are using the scripts provided by RDA, which is a composition of wget commands and aims at small amount of data downloading. For large amount of data, direct downloading is usually adopted.

Juozas please set up a FTP server for Sheng-hung to upload their sample data according to we can start developing the data comparison interface. I guess most of them are in NetCDF and GRIB, just like COLA data, and the system should be reusable on any data in any common formats.

*3.3 GMU COLA*

Dr. Ben Cash reported his current work in teleconnection models and the scripts are being prepared and everything goes well. Ben also met with CSISS teams to discuss the models and how the workflow looks like face-to-face in the meeting room of commerce building in March 15, 2018. Ben introduced the datasets of COLA, and also the datasets that he is using the research, including Model NMME North Aemerica MultiModel Ensemble, Global Modeling and Assimilation Office dataset, CPC Global unified cauge-based analysis of daily precipitation, and Extended reconstructed sea surface temperature (ERSST) v5.

Ben said the comfortable way for intercomparison is a three panel view including two panels side by side for comparison and a plot of difference map. To generate the difference map, a lot of unifying processes are needed. Juozas has shown Ben the system of workflows (<https://www.myexperiment.org/workflows>) and it is doubtful if the schema would work in the atmospheric science domain. It is very interesting but not optimistic to persude the scientists to actually adopt it. The deep-level reasons need be studied and exposed to figure out a good enough model comparison and sharing approach are that scientists have no reason to refuse to migrate.

As for big data processing, we at least should work on some theory researches, because generating 100TB on demand is always a challenging thing on either hardware or software aspects. This project may turn out to on-demand generate smaller size of results to reduce the system burden, after being published online and open for tens of thousands of users. COLA has around 1.5 to 2 PB data which is even too big for Google Earth Engine. We need think about our physical limits and how we can achieve the most at our hardware condition.

**4. Next Agenda**

Continue the discussion on the four cases in the proposal. Specificly:

1) the integration of Bcube and CyberConnector and other building blocks

2) the way to remotely access and efficiently visualize COLA/ASR datasets

3) the streamline process for COLA teleconnection and reanalysis models

4) the streamlined process for ASR model

5) interface design for intercomparison and model validation

6) We also need think about the possibility of creating a federation of CyberConnector, which means the CyberConnector can be treated as a tool and scientists can install it on their own servers to run. The federation of CyberConnectors will partially solve the big-data-processing-n-transferring problem.