ClimateModelDiagnosticAnalyzerProposal

Apache ClimateModelDiagnosticAnalyzer Proposal

Abstract

The Climate Model Diagnostic Analyzer (CMDA) provides web services for multi-aspect physics-based and phenomenon-oriented climate model performance evaluation and diagnosis through the comprehensive and synergistic use of multiple observational data, reanalysis data, and model outputs.

Proposal

The proposed web-based tools let users display, analyze, and download earth science data interactively. These tools help scientists quickly examine data to identify specific features, e.g., trends, geographical distributions, etc., and determine whether a further study is needed. All of the tools are designed and implemented to be general so that data from models, observation, and reanalysis are processed and displayed in a unified way to facilitate fair comparisons. The services prepare and display data as a colored map or an X-Y plot and allow users to download the analyzed data. Basic visual capabilities include 1) displaying two-dimensional variable as a map, zonal mean, and time series 2) displaying three-dimensional variable's zonal mean, a two-dimensional slice at a specific altitude, and a vertical profile. General analysis can be done using the difference, scatter plot, and conditional sampling services. All the tools support display options for using linear or logarithmic scales and allow users to specify a temporal range and months in a year. The source/input datasets for these tools are CMIP5 model outputs, Obs4MIP observational datasets, and ECMWF reanalysis datasets. They are stored on the server and are selectable by a user through the web services.

Service descriptions

1. **Two dimensional variable services**

- Map of two-dimensional variable: This services displays a two dimensional variable as a colored longitude and latitude map with values represented by a color scheme. Longitude and latitude ranges can be specified to magnify a specific region.
- Two dimensional variable zonal mean: This service plots the zonal mean value of a two-dimensional variable as a function of the latitude in terms of an X-Y plot.
- Two dimensional variable time series: This service displays the average of a two-dimensional variable over the specific region as function of time as an X-Y plot.

2. **Three dimensional variable services**

- Map of a two dimensional slice of a three-dimensional variable: This service displays a two-dimensional slice of a three-dimensional variable at a specific altitude as a colored longitude and latitude map with values represented by a color scheme.
- Three dimensional zonal mean: Zonal mean of the specified three-dimensional variable is computed and displayed as a colored altitude-latitude map.
- Vertical profile of a three-dimensional variable: Compute the area weighted average of a three-dimensional variable over the specified region and display the average as function of pressure level (altitude) as an X-Y plot.

3. **General services**

- Difference of two variables: This service displays the differences between the two variables, which can be either a two dimensional variable or a slice of a three-dimensional variable at a specified altitude as colored longitude and latitude maps
- Scatter and histogram plots of two variables: This service displays the scatter plot (X-Y plot) between two specified variables and the histograms of the two variables. The number of samples can be specified and the correlation is computed. The two variables can be either a two-dimensional variable or a slice of a three-dimensional variable at a specific altitude.
- Conditional sampling: This service lets user to sort a physical quantity of two or dimensions according to the values of another variable (environmental condition, e.g. SST) which may be a two-dimensional variable or a slice of a three-dimensional variable at a specific altitude. For a two dimensional quantity, the plot is displayed an X-Y plot, and for a two-dimensional quantity, plot is displayed as a colored-map.

Background and Rationale

The latest Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report stressed the need for the comprehensive and innovative evaluation of climate models with newly available global observations. The traditional approach to climate model evaluation, which is the comparison of a single parameter at a time, identifies symptomatic model biases and errors but fails to diagnose the model problems. The model diagnosis process requires physics-based multi-variable comparisons, which typically involve large-volume and heterogeneous datasets, and computationally demanding and data-intensive operations. We propose to develop a computationally efficient information system to enable the physics-based multi-variable model performance evaluations and diagnoses through the comprehensive and synergistic use of multiple observational data, reanalysis data, and model outputs.

Satellite observations have been widely used in model-data inter-comparisons and model evaluation studies. These studies normally involve the comparison of a single parameter at a time using a time and space average. For example, modeling cloud-related processes in global climate models requires cloud parameterizations that provide quantitative rules for expressing the location, frequency of occurrence, and intensity of the clouds in terms of multiple large-scale model-resolved parameters such as temperature, pressure, humidity, and wind. One can evaluate the performance of the cloud parameterization by comparing the cloud water content with satellite data and can identify symptomatic model biases or errors. However, in order to understand the cause of the biases and errors, one has to simultaneously investigate several parameters that are integrated in the cloud parameterization.

Such studies, aimed at a multi-parameter model diagnosis, require locating, understanding, and manipulating multi-source observation datasets, model outputs, and (re)analysis outputs that are physically distributed, massive in volume, heterogeneous in format, and provide little information on data quality and production legacy. Additionally, these studies involve various data preparation and processing steps that can easily become computationally demanding since many datasets have to be combined and processed simultaneously. It is notorious that scientists spend more than 60% of their research time on just preparing the dataset before it can be analyzed for their research.

To address these challenges, we propose to build Climate Model Diagnostic Analyzer (CMDA) that will enable a streamlined and structured preparation of multiple large-volume and heterogeneous datasets, and provide a computationally efficient approach to processing the datasets for model diagnosis. We will leverage the existing information technologies and scientific tools that we developed in our current NASA ROSES COUND, MAP, and AIST projects. We will utilize the open-source Web-service technology. We will make CMDA complementary to other climate model analysis tools currently available to the research community (e.g., PCMDI's CDAT and NCAR's CCMVal) by focusing on the missing capabilities such as conditional sampling, and probability distribution function and cluster analysis of multiple-instrument datasets. The users will be able to use a web browser to interface with CMDA.

Current Status

The current version of ClimateModelDiagnosticAnalyzer was developed by a team at The Jet Propulsion Laboratory (JPL). The project was initiated as a NASA-sponsored project (ROSES-CMAC) in 2011.

Meritocracy

The current developers are not familiar with meritocratic open source development at Apache, but would like to encourage this style of development for the project.

Community

While ClimateModelDiagnosticAnalyzer started as a JPL research project, it has been used in The 2014 Caltech Summer School sponsored by the JPL Center for Climate Sciences. Some 23 students from different institutions over the world participated. We deployed the tool to the Amazon Cloud and let every student each has his or her own virtual machine. Students gave positive feedback mostly on the usability and speed of our web services. We also collected a number of enhancement requests. We seek to further grow the developer and user communities using the Apache open source venue. During incubation we will explicitly seek increased academic collaborations (e.g., with The Carnegie Mellon University) as well as industrial participation.

One instance of our web services can be found at: http://cmacws4.jpl.nasa.gov:8080/cmac/

Core Developers

The core developers of the project are JPL scientists and software developers.

Alignment

Apache is the most natural home for taking the ClimateModelDiagnosticAnalyzer project forward. It is well-aligned with some Apache projects such as Apache Open Climate Workbench. ClimateModelDiagnosticAnalyzer also seeks to achieve an Apache-style development model; it is seeking a broader community of contributors and users in order to achieve its full potential and value to the Climate Science and Big Data community.

There are also a number of dependencies that will be mentioned below in the Relationships with Other Apache products section.

Known Risks

Orphaned products

Given the current level of intellectual investment in ClimateModelDiagnosticAnalyzer, the risk of the project being abandoned is very small. The Carnegie Mellon University and JPL are collaborating (2014-2015) to build a service for climate analytics workflow recommendation using fund from NASA. A two-year NASA AIST project (2015-2016) will soon start to add diagnostic analysis methodologies such as conditional sampling method, conditional probability density function, data co-location, and random forest. We will also infuse the provenance technology into CMDA so that the history of the data products and workflows will be automatically collected and saved. This information will also be indexed so that the products and workflows can be searchable by the community of climate scientists and students.

Inexperience with Open Source

The current developers of ClimateModelDiagnosticAnalyzer are inexperienced with Open Source. However, our Champion Chris Mattmann is experienced (Champions of ApacheOpenClimateWorkbench and AsterixDB) and will be working closely with us, also as the Chief Architect of our JPL section.

Relationships with Other Apache Products

Clearly there is a direct relationship between this project and the Apache Open Climate Workbench already a top level Apache project and also brought to the ASF by its Champion (and ours) Chris Mattmann. We plan on directly collaborating with the Open Climate Workbench community via our Champion and we also welcome ASF mentors familiar with the OCW project to help mentor our project. In addition our team is extremely welcoming of ASF projects and if there are synergies with them we invite participation in the proposal and in the discussion.

Homogeneous Developers

The current community is within JPL but we would like to increase the heterogeneity.

Reliance on Salaried Developers

The initial committers are full-time JPL staff from 2013 to 2014. The other committers from 2014 to 2015 are a mix of CMU faculty, students and JPL staff.

An Excessive Fascination with the Apache Brand

We believe in the processes, systems, and framework Apache has put in place. Apache is also known to foster a great community around their projects and provide exposure. While brand is important, our fascination with it is not excessive. We believe that the ASF is the right home for ClimateModelDiagnosticAnalyzer and that having ClimateModelDiagnosticAnalyzer inside of the ASF will lead to a better long-term outcome for the Climate Science and Big Data community.

Documentation

The ClimateModelDiagnosticAnalyzer services and documentation can be found at: http://cmacws4.jpl.nasa.gov:8080/cmac/.

Initial Source

Current source resides in ...

External Dependencies

ClimateModelDiagnosticAnalyzer depends on a number of open source projects:

- Flask
- Gunicorn
- Tornado Web Server
- GNU octave
- epd python
- NOAA ferret
- GNU plot

Required Resources

Developer and user mailing lists

- private@cmda.incubator.apache.org (with moderated subscriptions)
- commits@cmda.incubator.apache.org
- dev@cmda.incubator.apache.org
- users@cmda.incubator.apache.org

A git repository

https://git-wip-us.apache.org/repos/asf/incubator-cmda.git

A JIRA issue tracker

https://issues.apache.org/jira/browse/CMDA

Initial Committers

The following is a list of the planned initial Apache committers (the active subset of the committers for the current repository at Google code).

- Seungwon Lee (seungwon.lee@jpl.nasa.gov)
- Lei Pan (lei.pan@jpl.nasa.gov)
- Chengxing Zhai (chengxing.zhai@jpl.nasa.gov)
- Benyang Tang (benyang.tang@jpl.nasa.gov)
- Jia Zhang (jia.zhang@sv.cmu.edu)
- Wei Wang (wei.wang@sv.cmu.edu)
- Chris Lee (chris.lee@sv.cmu.edu)
- Xing Wei (xing.wei@sv.cmu.edu)

Affiliations

JPL

- Seungwon Lee
- Lei Pan

- Chengxing Zhai
- Benyang Tang

CMU

- Jia Zhang
- Wei Wang
- Chris Lee
- Xing Wei

Sponsors

NASA

Champion

Chris Mattmann (NASA/JPL)

Nominated Mentors

Greg Reddin

Chris Mattmann

Michael Joyce

James Carman

Kim Whitehall

Sponsoring Entity

The Apache Incubator