Climate Model Diagnostic Analyzer

Seungwon Lee, Lei Pan, Chengxing Zhai, Benyang Tang, and Terry Kubar
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA U.S.A.
Seungwon.Lee@jpl.nasa.gov

Jia Zhang and Wei Wang
Department of Software Engineering
Carnegie Mellon University
Silicon Valley, CA U.S.A.
Jia.Zhang@sv.cmu.edu

Abstract— **The comprehensive and innovative evaluation of climate models with newly available global observations is critically needed for the improvement of climate model current-state representation and future-state predictability. A climate model diagnostic evaluation process requires physics-based multi-variable analyses that typically involve large-volume and heterogeneous datasets, making them both computation- and data-intensive. With an exploratory nature of climate data analyses and an explosive growth of datasets and service tools, scientists are struggling to keep track of their datasets, tools, and execution/study history, let alone sharing them with others. In response, we have developed a cloud-enabled, provenance-supported, web-service system called Climate Model Diagnostic Analyzer (CMDA). CMDA enables 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. At the same time, CMDA provides a crowdsourcing space where scientists can organize their work efficiently and share their work with others. CMDA is empowered by many current state-of-the-art software packages in web service, provenance, and semantic search.**

Keywords—climate data; analytics; model evaluation, online collaborative environment, web services, cloud computing

I. INTRODUCTION

Improving the model representations of the current climate system is essential to enhancing confidence in seasonal, decadal, and long-term climate projections. Both the National Research Council (NRC) Decadal Survey and the latest Intergovernmental Panel on Climate Change (IPCC) Assessment Report stressed the need for comprehensive and innovative evaluations of climate models with the synergistic use of global observations. The traditional approach to climate model evaluation, which compares a single parameter at a time, identifies symptomatic model biases and errors but fails to diagnose the model problems. A new innovative approach needs to be developed to diagnose model biases in contemporary climate models, identify the physical processes responsible for model biases, and incorporate the understanding into new model representations that reduce the model biases.

NASA and NOAA have established a collection of data centers to store the rapidly growing satellite-based and ground-based sensor data and model-generated data. To process such datasets, a number of tools and analytics software have been established. For example, NASA has supported the development of data and information systems, including data processing tools and data-service discovery and publication tools, such as the Earth Observing System Data and Information System (EOSDIS) and NASA Earth Exchange (NEX). However, sharing the tools and knowledge with the community has not been fully explored and realized because of the lack or insufficiency of infrastructure tools to support the tool and knowledge sharing. With an explosive growth of datasets and service tools, Earth scientists are struggling to keep track of their datasets, tools, and execution/study history, let alone sharing them with others. The community is in desperate need of infrastructure tools to support the organization of their work and sharing their knowledge.

In response, we have developed a system called Climate Model Diagnostic Analyzer (CMDA). CMDA enables diagnostic model evaluations with advanced multi-variate statistical and machine learning computing. CMDA provides an online collaborative environment where Earth scientists can easily publish their climate data analytics web services, share them within groups, and find those of others. CMDA currently supports (1) all the datasets from Obs4MIPs and a few ocean datasets from NOAA and Argo, which serve as observation-based reference data for model evaluation, (2) many of CMIP5 model outputs covering a broad range of atmosphere, ocean, and land variables from the CMIP5 specific historical runs, AMIP runs, and RCP 4.5 experiment runs, and (3) ECMWF reanalysis outputs for several environment variables in order to supplement observational datasets. Analysis capabilities currently supported by CMDA are (1) the calculation of annual and seasonal means of physical variables, (2) the calculation of time evolution of the means in any specified geographical region, (3) the calculation of correlation between two variables with a time lag if needed, (4) the calculation of difference between two variables, (5) the conditional sampling of one physical variable with respect to another variable, (6) the random-forest based feature importance ranking of a variable with respect to dependences on other variables, and (7) the regridding of datasets with specified horizontal and vertical resolutions.

In this paper, we describe the innovative methodology that we have developed for diagnostic model evaluations in Section 2. We describe the web service technology that we have applied to build the online collaborative environment infrastructure system in Sections 3 and 4, respectively. We

978-1-4799-9926-2/15/$31.00 ©2015 IEEE

summarize the applications and perceived impact of CMDA in climate projection and climate model evaluation in Section 5.

II. Methodology for Diagnostic Model Evaluations

We have developed a novel methodology to diagnose model biases in contemporary climate models, to identify the physical processes responsible for model biases, and to incorporate the understanding into new model representations that reduce the model biases. The methodology includes (1) conditional sampling method, (2) conditional probability density function, and (3) random forest.

A. Conditional Sampling

Conditional sampling method is originally developed and applied for model evaluations by Bony et al. [1]. It is a novel way to display a physical quantity X according to the values of another physical quantity E and to decompose the errors of the average quantity of X in terms of errors in X, errors in E and errors in the covariance of X and E. The quantity X is called a sampled parameter and the quantity E is called a sampling parameter. By displaying X in terms of E, this method enables to study the inter-relationship between X and E. For example, we may want to display the cloud ice content according to the sea surface temperature. The resulting plot reveals how the cloud ice content values are distributed at different sea surface temperatures, how the two parameters are related, and how well its governing physical process is represented in a model in comparison with an observational reference dataset.

Fig. 1 shows an example plot of the conditional sampling method applied to cloud water content profile sampled by 500 hPa vertical velocity. When the method is applied to model outputs and observation/reanalysis dataset, one can easily identify model biases in representing the inter-relationship between the two physical quantities by comparing the pattern of the conditional sampling plots. The bottom middle panel shows the result with the CloudSat observation and ECMWF reanalysis data, providing a reference for the climate model evaluation.

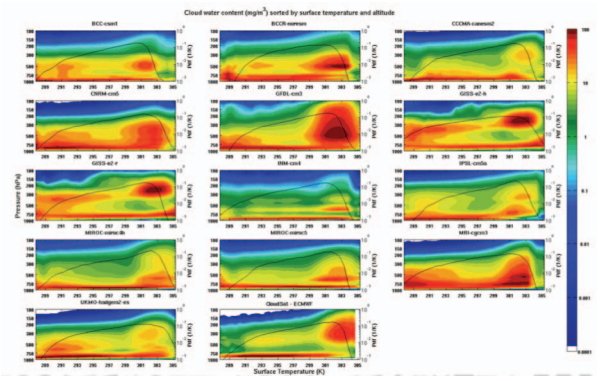

Fig. 1. Example of the conditional sampling method: Cloud water content profiles conditionally sampled with 500 hPa vertical velocity for contemporary climate models and observational datasets. The comparison between the models and observations give insight into how well the models represent physical processes governing the relationship between the two variables.

B. Conditional Probability Density Function

NASA ROSES CMAC and AIST Programs

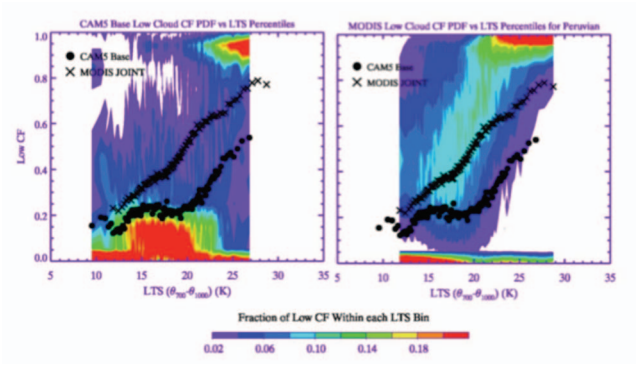

Fig. 2. Example of the conditional probability density function method: Conditional PDFs of low cloud fraction as function of lower tropospheric stability. The left panel is with the CAM5 model, and the right panel is with the MODIS-ECMWF observation/reanalysis dataset. The striking difference in the two plots identifies systematic biases in the CAM5 model in representing the low cloud.

Conditional probability density function is a by-product of the conditional sampling method and is a very insightful method to identify the sources and characteristics of model errors and to diagnose the model error sources. The conditional probability density function (PDF) is defined as the PDF of the sampled parameter X for a given sampling parameter bin e, referred to as $P(x|e)$. The conditional PDF carries considerably more information about the relationship between parameter X and E than the mean and variance of the parameter X for a given bin e.

As an example of the method, the conditional PDFs of low cloud fraction (LCF) as functions of lower tropospheric stability (LTS) can be calculated for a climate model output (CAM5: the NCAR's Community Earth System Model) and an observation/reanalysis dataset (MODIS-ECMWF). The resulting conditional PDFs along with mean values for the model and the observation are shown in Fig. 2. The mean CAM5 model LCF is systematically lower than observations, and the model LCF PDF distributions differ from MODIS. While MODIS indicates a rather smooth transition from low values of cloud fraction to a solid stratocumulus regime with LTS, CAM5 shows a small cloud fraction mode until an unrealistic "step-wise" increase in cloud fraction starting at around the LTS of 20K. These biased features of the PDFs would not have been identified if only the mean values were examined. The scientific interpretation of the result is presented elsewhere [3].

C. Random Forest

Random Forest is an algorithm for classification and regression developed by Leo Breiman in 2001 [2]. Random Forest uses an ensemble of decision trees. Each decision tree is built using a subset of the data samples, and at each split of the tree a random subset of input variables is used to evaluate the split threshold. The final prediction is the voting (for classification) or averaging (for regression) of all the trees. Random forest has good performance in both classification and regression problems, with skills comparable or sometimes superior to those of the support vector machines (depending on datasets). We use the random forest method to measure

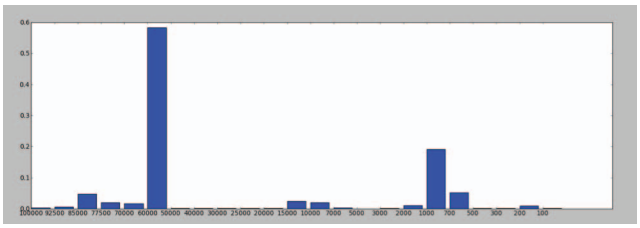

Fig. 3. Example of the random forest method: random forest variable importance score with air temperature vertical profiles as input variables and precipitation as a target variable. The x-axis is the index for the input variables. The y-axis is the input variable importance score.

feature importance in model target variables with respect to model input variables.

In Earth science data analysis, it is often helpful to see the relationships among a set of physical variables. A typical approach is to calculate the correlation coefficient between any two variables. However, the correlation coefficient is limited to a linear relationship. Even a fitting to an assumed function is limited to the assumption made for the fitting. In contrast, the random forest does not assume a linear relationship or other particular relationship between variables. When a target variable and input variables are selected, then a random forest model can be built to predict the target variable from the input variables. Random forest variable importance can be calculated, which ranks the relatedness of the input variables to the target variable.

As an example of random forest application to climate data, we have calculated random forest variable importance with air temperature vertical profiles as input variables and precipitation as a target variable. Fig. 3 shows the results of the variable importance ranking. The x-axis is the index (pressure value) for the air temperature at a different pressure level, and the y-axis is the random forest variable importance score. The result shows that the air temperature at 65000 Pa has the highest score, indicating that the air temperature at 65000 Pa influences the most in predicting the target variable – precipitation. The scientific interpretation and understanding of this kind of results require a more examination and exploration of this method.

III. Web Service Technology

Many of research codes are written in a non-general and non-scalable way, making it difficult to share with others. In addition, the programming languages and libraries used by the code often require a local software installation and environment configuration, making it difficult for others to adopt the tool. In response, we have developed a methodology to transform an existing science application code in various programming languages into a web service A web service approach is chosen because it not only lowers the learning curve and removes the adoption barrier of the tool but also enables instantaneous use compared to offline standalone application, avoiding the hassle of local software installation and environment incompatibility. The web service technology also has a simple and flexible environment with a rich set of open source packages.

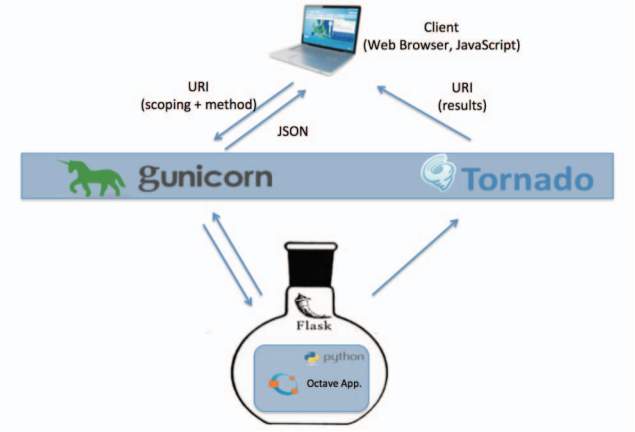

Fig. 4. Schematic diagram of creating a web service in CMDA

The schematic diagram of creating a web service in CMDA with technical components and their relationship with one another is shown in Fig. 4. The screenshots of the CMDA web service web browser interface are shown in Fig. 5.

The following steps are taken in creating a web service.

(1) We wrap an existing science application code with a Python caller. The Python caller treats the application as a child process, prepares all input arguments for the child process, defines where to put the outputs of the child process, spawns off the child process, captures the stdout and stderr of the child process. At the end, the science application looks like a python application.

(2) We use Flask, an open source light-weight web development framework for Python applications, to create an entry point code for a web service. The entry code parses input arguments from a client (a web service), calls a Python application and passes input arguments to the Python application, and retrieves return values from the Python application and pass them to a client. It follows a REST-ful (Representational State Transfer) style, where scoping information (what data to operate) is placed in a URI (Uniform Resource Identifier) while method information (what to do with the data) is conveyed in an HTTP (the

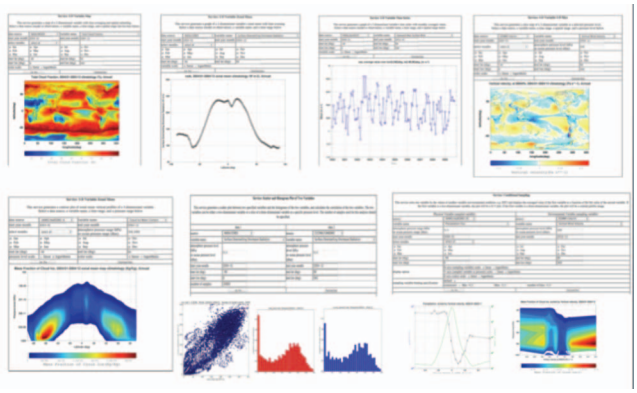

Fig. 5. Screenshots of CMDA service web browser interface

Hypertext Transfer Protocol) method [1,2].

(3) We separate application traffic from static HTTP traffic. We use Gunicorn to provide WSGI service application

traffic for web service scoping and method information, while we use Tornado to provide web service static HTTP traffic for web service results.

(4) We design a web browser interface for a web service and implement it using JavaScript.

IV. Online Collaborative Environment

We are building an online collaborative environment in CMDA, where a community can build, share, search, and recommend web services for climate data analytics and organize their execution history. The key functionalities are CMDA service publication, CMDA service and dataset search and recommendation, and CMDA service execution history management. We are building a centralized CMDA service registry that maintains a set of links for published CMDA services. We are developing a way to automatically create front-end HTML pages for CMDA services.

A system with strong query and recommendation facility requires an underlying semantics model. We are developing both static semantics and behavioral semantics: static semantics describe the functionalities and goals that a CMDA service promises to provide, and behavioral semantics describe the required circumstances when a CMDA service can behave, including input and output parameters, pre- and post-constraints, and historical usage patterns. Based on the service semantics model, we are developing a technique that can automatically extract aforementioned semantic metadata from CMDA services.

In order to support reproducibility, we are developing a provenance model to record and track scientists' activities and behaviors. With the execution history stored in a database, we are developing a system to search executions and to reproduce the results. Fig. 6 shows our current design of the execution search page and the search result page.

We are applying mature semantic web techniques and machine learning techniques to build an intelligent search facility [6]. Furthermore, we are applying the most recent web techniques (including HTML5, JavaScript, Apache Lucene, Play framework) and modern software engineering methodologies (including Extreme Programming, Agile technique, and Scrum) to develop a scalable, extensible, and interoperable online environment [7,8].

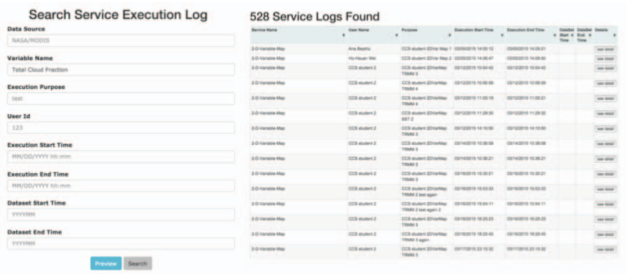

Fig. 6. CMDA service execution history search page and result page.

V. Educational Use of CMDA

CMDA has been used as an educational tool for the summer school organized by JPL's Center for Climate Science in 2014 and will be used again in the summer school in 2015 [9]. The theme of the summer school is using satellite observations to advance climate models, which is well aligned with the main goal and capability of CMDA. The requirements of the educational tool are defined with the interaction with the school organizers, and CMDA is customized to meet the requirements accordingly. Since CMDA is used by 30+ simultaneous users during the school, we have imported CMDA to the Amazon cloud environment. The cloud-enabled CMDA provides each student with an independent computing resource and working environment while the user interaction with the system remains the same through web-browser interfaces.

The summer school in 2014 had five group research topics: (1) surface and deep ocean connections; (2) observed variability of clouds and precipitation; (3) modeled spatial and temporal variability of clouds and precipitation; (4) vegetation phenology and climate controls; and (5) land water storage variability as a function of human and natural controls. CMDA have provided students with 336 climate datasets and 10 analysis tools. The datasets covered are multi-year monthly gridded data from observations, reanalysis runs, and model runs. The analysis tools visualize one variable or two variable relationships and differences in time average, spatial average, correlation, and conditional sampling. A one-hour session for the CMDA introduction was given, and immediately after the introduction the students were able to start using CMDA with a virtual machine assigned to them in Amazon Cloud. The students used CMDA for two practice sessions, which lasted about 5 hours total, and were able to present their results of their group research project. Fig. 7 shows the pictures of the summer school during the CMDA introduction session and the group research sessions.

The summer school served as a valuable test-bed for the CMDA development, preparing CMDA to serve its target community: Earth-science modeling and model-analysis community. In the upcoming summer school in 2015, we are planning to deploy our online collaborative environment

Fig. 7 JPL Center for Climate Sciences summer school in 2014. CMDA provided datasets and tools for students to use for their group research project during the school.

features including dataset search and execution history search. The new features will help students quickly discover and identify datasets needed for their research project and help

students keep track of their previous results obtained by running the CMDA web services and reproduce them if needed.

VI. Conclusions

Rapidly growing datasets and analytics services in Earth Science challenge individual Earth scientists in organizing their work and concurrently challenge the whole community in sharing the datasets and tools and derived knowledge. With the community recognizing the need of infrastructure systems to address those challenges, some systems are under development but with a marginal impact so far in terms of tool adoption by the community and tool functionality. CMDA is designed to address the community need in a lightweight and easy-to-use and easy-to-maintain manner, with a focused domain of climate data analysis. CMDA provides a space where Earth scientists can organize their work efficiently and at the same time, share their work with others. With the projected exponential growth of the datasets and analytics tools, the goal of CMDA is to significantly ease the burden of individual scientists, increase their productivity, and as the whole community, to increase the scientific return of the NASA and NOAA's Earth science investments.

The focused Earth science application of CMDA is climate data analysis for climate projection and climate model evaluation. Recent community reports emphasize the need for the comprehensive and innovative evaluation of climate models with the synergistic use of global observations. Improving the model representations of the climate system is critically needed to order to enhance the fidelity of the models in seasonal, decadal, and long-term climate projections. CMDA enables the important model evaluation activities by providing key resources in terms of datasets, analytics services, computing resources, and at the same time by providing a platform to organize their work and share their work with others.

Acknowledgment

The authors acknowledge support from their institutions and funding sources. A part of the work was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. A part of the work was carried at Carnegie Mellon University – Silicon Valley. We thank the NASA ROSES CMAC and AIST programs for funding this work.

References

[1] Bony, S., J.-L. Dufresne, H. L. Treut, J.-J. Morcrette, and C. Senior, "On dynamic and thermodynamic components of cloud changes," *Clim. Dyn.,* **22**, 71-86, 2004.

[2] Breiman, Leo, 2001: "Random Forests," *Machine Learning* **45** (1): 5–32. doi:10.1023/A:1010933404324.

[3] Kubar, T. L., G. L. Stephens, M. Lebsock, V. E. Larson, and P. A. Bogenschutz, 2014: Regional assessments of low clouds against large-scale stability in CAM5 and CAM-CLUBB using MODIS and ECMWF-Interim Reanalysis Data. *J. Climate*, in press.

[4] Fielding, Roy T. and Taylor, Richard N, "Principled Design of the Modern Web Architecture" (PDF), *ACM Transactions on Internet Technology (TOIT)* (New York: Association for Computing Machinery) **2** (2): 115150, May 2002.

[5] Richardson, Leonard and Ruby, Sam, *RESTful Web Services*, O'Reilly, ISBN 978-0-596-52926-0, May 2007.

[6] G. Bort, "The High Velocity Web Framework for Java and Scala", 2015, accessed, Available from: https://www.playframework.com/.

[7] Apache, "Apache Lucene", accessed on: Jul. 16, 2015, Available from: https://lucene.apache.org.

[8] K. Schwaber and J. Sutherland, "The Scrum Guide", 2011, accessed, Available from: http://www.scrum.org/storage/scrumguides/Scrum_Guide.pdf.

[9] JPL Center for Climate Scineces Summer School Site: http://climatesciences.jpl.nasa.gov/events/summer-school.