The Community Diagnostics Package: A Common Scaffold for Diagnostics

Analytics and Informatics Management Systems (AIMS) team white paper for unifying diagnostics frameworks

Sam Fries, Zeshawn Shaheen, Ji-Woo Lee, Charles Doutriaux, Dean N. Williams

Executive Summary

Modelers and researchers use diagnostics packages to provide quick insight into the effects of model parameters and new development. The various packages currently available in the community perform many similar tasks, and face many similar challenges. The Lawrence Livermore National Laboratory (LLNL) Analytics and Informatics Management Systems (AIMS) team is working on the Community Diagnostics Package (CDP), a unifying framework that solves many of the basic infrastructure issues faced by these packages. It provides an un-opinionated platform that handles intricate but common tasks (parallelization, logging, configuration management, output handling, provenance capture, etc.), an object-oriented application programming interface (API) for development of new metrics, and integration into a wide suite of tools to allow for simpler workflows.

Overview

The LLNL AIMS team is currently involved in three diagnostics projects: the Program for Climate Model Diagnosis and Intercomparison (PCMDI) Metrics Package (PMP), the Accelerated Climate Model for Energy (ACME) Diagnostics Package, and the Atmospheric Radiation Measurement (ARM) Diagnostics. From our work on these, we have identified many parts that are basically identical, but implemented in completely different ways: logging, parallelization, configuration management, output handling, provenance capture, variable derivation, input/output streams, and a variety of others. The code written to handle these features tends to require extensive effort to maintain, and when written badly can hamper the effectiveness of the package substantially. They are the kinds of features that are the domain of computer scientists and software engineers, rather than climate researchers or model developers. Scientific code is frequently created for a single use and narrowly focused on that goal, but code that is woven throughout the entire package must be much more flexible than that to consider various formats, spatiotemporal resolution, or coverage of individual datasets. To this end, we propose to create a framework that handles these features for researchers, allowing them to abstract data management from the scientific part of their job.

This library, the Community Diagnostics Package (CDP), would provide an object-oriented API for managing the various tasks associated with running a suite of diagnostics. It would provide simple plug-and-play components to power each of the various steps (reading parameters, loading data, scheduling multiprocessor execution, computing the metrics, tracking the provenance), which the diagnostics developers could use to choose the most effective options for their specific needs (using SLURM to schedule the jobs on a cluster or supercomputer, or using a message passing interface (MPI) and other parallel processing systems directly, or doing a simple process pool on a single node). All of these components are planned to be modular in design, so different implementations of the same component can be interchangeable with minimal change for existing code. Scientists would be able to implement custom metrics by simply defining in a minimalist fashion what inputs are required (e.g., the root mean square error (RMSE) would require the difference of the model and observations), and then writing a simple computation function to actually do the calculation of the metric on those values, which would be provided by CDP automatically. Once a user defines the metric, it can be used right away as part of a custom diagnostics set, simply plugged in to an existing diagnostics suite and use it right away, or contributed back to the diagnostics suite for community use.

In support of the Earth System Grid Federation (ESGF), we created a library that provides a system for indexing the output of a diagnostics package and automatically generating a user-friendly web interface for looking through the results locally (a "viewer"). This library integrates with a web service to provide a simple workflow for uploading diagnostics results to a server and sharing them with collaborators. This library was integrated with the ACME Diagnostics Package, and has had extremely positive feedback from users; however, the capabilities provided are fairly basic. The viewer generated consists largely of a series of tables with some small enhancements, but users have expressed interest in a more interactive system. In particular, support for clickable graphics has been requests, to enable the navigation system shown in Figure 1.

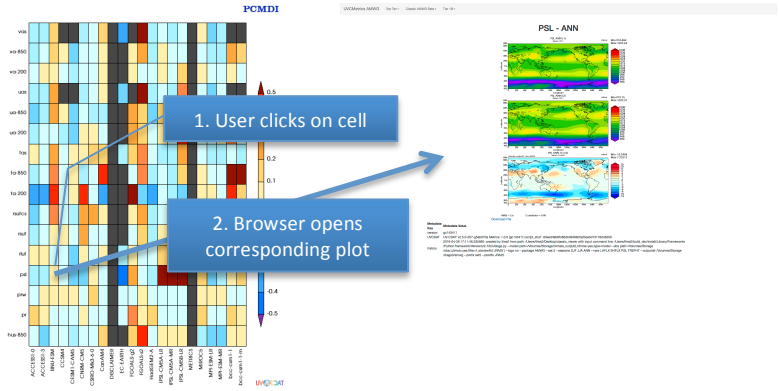

Figure 1: Interactive navigation concept for "Portrait Plot" (left image from PCMDI Metrics Package, right image from ACME Diagnostics Package).

By including the viewer library, the CDP will allow diagnostics packages to have a viewing framework for sharing their results, rather than requiring each package to create viewing options from scratch, or requiring users to compose ad-hoc solutions. The viewer library will also interact with the provenance capture component of the CDP. Each diagnostics run will generate provenance data that will track data used for input, version of the package was used, who ran the diagnostics, etc. This information would then be bundled with the output automatically, available within the web service as well as in the local viewer.

Proposed Work

The plan for development is to build out a skeleton version of CDP, and then begin integrating support for PMP as a proof-of-concept while fleshing out the library. We'll maintain a library of essential metrics (such as RMSE, zonal means, bias, etc.) to allow diagnostics packages to reuse the same code, and set it up so users, researchers, and modelers can submit new metrics for community use. We'll also enhance the viewer software to improve interaction and exploration of diagnostics output, and better integrate the web viewer service with ESGF. In the future, we will look at integrating CDP with other diagnostics packages, such as ACME, the ARM Diagnostics, and the Climate Variability Diagnostics Package (CVDP) from NCAR.