ClimateProposal

Apache Open Climate Workbench, tool for scalable comparison of remote sensing observations to climate model outputs, regionally and globally.

Abstract

The Apache Open Climate Workbench proposal desires to contribute an existing community of software related to the analysis and evaluation of climate models, and related to the use of remote sensing data in that process.

Specifically, we will bring a fundamental software toolkit for analysis and evaluation of climate model output against remote sensing data. The toolkit is called the Regional Climate Model Evaluation System (RCMES). RCMES provides two fundamental components for the easy, intuitive comparison of climate model output against remote sensing data. The first component called RCMED (for "Regional Climate Model Evaluation Database") is a scalable cloud database that decimates remote sensing data and renalysis data related to climate using Apache OODT extractors, Apache Tika, etc. These transformations make traditionally heterogeneous upstream remote sensing data and climate model output homogeneous and unify them into a data point model of the form (lat, lng, time, value, height) on a per parameter basis. Latitude (lat) and Longitude (lng) are in WGS84 format, but can be reformatted on the fly. time is in ISO 8601 format, a string sortable format independent of underlying store. value carries with it units, related to interpretation and height allows for different values for different atmospheric vertical levels. All of RCMES is built on Apache OODT, Apache Sqoop/Apache Hadoop and Apache Hive, along with hooks to PostGIS and MySQL (traditional relational databases). The second component of the system, RCMET (for "Regional Climate Model Evaluation Toolkit") provides facilities for connecting to RCMED, dynamically obtaining remote sensing data for a space/time region of interest, grabbing associated model output (that the user brings, or from the Earth System Grid Federation) of the same form, and then regridding the remote sensing data to be on the model output grid, or the model output to be on the remote sensing data grid. The regridded data spatially is then temporally regridded using techniques including seasonal cycle compositing (e.g., all summer months, all Januaries, etc.), or by daily, monthly, etc. The uniform model output and remote sensing data are then analyzed using pluggable metrics, e.g., Probability Distribution Functions (PDFs), Root Mean Squared Error (RMSE), Bias, and other (possibly user-defined) techniques, computing an analyzed comparison or evaluation. This evaluation is then visualized by plugging in to the NCAR NCL library for producing static plots (histograms, time series, etc.)

We also have performed a great deal of work in packaging RCMES to make the system easy to deploy. We have working Virtual Machines (VMWare VMX and Virtual Box OVA compatible formats) and we also have an installer built on Python Buildout (http://buildout.org/) called "Easy RCMET" for dynamically constructing the RCMET toolkit.

RCMES is currently supporting a number of recognized climate projects of (inter-)national significance. In particular, RCMES is supporting the U.S. National Climate Assessment (NCA) activities on behalf of NASA's contribution to the NCA; is working with the North American Regional Climate Change Assessment Program (NARCCAP); and is also working with the International Coordinated Regional Downscaling Experiment (CORDEX).

Proposal

We propose to transition the RCMES software community, which includes developers of the RCMET and RCMED software, along with users of RCMES in the CORDEX project across a variety of academic institutions, scientists helping to improve the RCMES metrics, and visualizations, and regridding algorithms, packagers making RCMES easier to install, and scientists helping to lead some of these international projects that are already using RCMES.

We have been working on the RCMES project since 2009 funded initially by the American Recovery and Reinvestment Act (ARRA) project out at NASA, and then branching out into other sources of support and sustainability (NASA; NSF, etc. – see the acknowledgements section on the RCMES website for a full list of supporting U.S. and international partners).

With the existing RCMES community at Apache, we will also work to encourage other climate software projects e.g., Open Climate GIS, elements of the Earth System Grid Federation, other NASA climate projects funded under the Computational Modeling, Algorithms and Cyberinfrastructure (CMAC) to contribute to the Open Climate Workbench here at Apache.

RCMED is a Big Data project that combines several underlying Apache software – OODT, Tika, Hadoop, HIVE, and Sqoop – and other related data management software. Its primary language is Java; RCMET, on the other hand, is a Python API, associated set of classes (framework), set of Python Bottle Web services, and a PHP "Wizard"-based User Interface that leverages Apache OODT Balance.

Background

Bringing RCMES to Apache was the brain-child of Chris Mattmann, based on his solid experience with Apache OODT and bringing it to the ASF. Chris worked for a year to get the support of the JPL community including approvals from the Software Release authority at JPL to release the software.

The initial code drop will include the RCMES SVN repository from JPL including prior revisions. We anticipate also including a smaller package, CDX, which contains some useful facilities for regridding, and command line tools for manipulating large datasets, and working with OPeNDAP, etc.

After the code drop, we will work with our developers, users, documentors, and other members of the team to teach those unfamiliar with the Apache way how it works around here at Apache. 30% of the community from RCMES includes those intimately familiar with Apache including 6 ASF members – the other 70% include a range of scientific code developers, climate scientists that use RCMES, program officers that will help make documentation and slides for the code, and advocate for it in the community. Their experience with Apache ranges from using various ASF products, to contributing patches to them, to not using any ASF software at all.

With this diversity, we anticipate that while everything may not just work turnkey out of the box, this represents a unique opportunity to demonstrate Apache to the international community and to show the benefits of its community and social models. That said, we also have a lot of ASF experience to make sure everyone learns the Apache way.

Rationale

We are bringing RCMES to Apache for a few reasons. First, we feel that it will immediately enable our collaborators across a number of institutions both nationally and internationally have the opportunity to work on a common software base, and to improve it with contributions from their own sites. Currently these are difficult to negotiate now because of varied legal and contribution frameworks – Apache allows us to simplify this to a unified one. Second, using the ASF's world-wide mirroring system, we will be able to deliver climate software broadly to the community as we release it, rather than sneaker netting the software around or establishing our own point release infrastructure.

Bringing this project to Apache also immediately thrusts the ASF into the thriving ecosystem of the Coordinated Regional Downscaling Experiment (CORDEX), the US National Cimate Assessment, the North American Regional Climate Change Assessment Program (the US contribution to CORDEX) and into relevance for upcoming Intergovernmental Panel on Climate Change (IPCC) assessment activities at a number of different institutions. We also seek to help lead and encourage du jour standard development rather than top down level dictating of standards for climate software and the ASF will provide us a means for that.

Initial Goals

The initial goals of the proposed project are:

- Stand up a sustaining Apache-based community around the JPL RCMES codebase.
- Active relationships and possible cooperation with related projects and communities, including end user and scientific communities, CORDEX, NARCCAP, US NCA, IPCC, ESG, etc.
- Active relationships and possible cooperation with existing Apache communities, e.g., OODT, Hadoop/HIVE, Sqoop, Tika, SIS, etc.
- Initial Apache release.
- Leverage Apache Open Climate Workbench in climate activities at NASA, in the international community as mentioned above, and beyond.
- Vetting all software licenses and making sure IP is clear (software grant from JPL forthcoming).

Current Status

Meritocracy

30% of the proposed initial committers are familiar with the meritocracy principles of Apache. As stated above this includes 6 ASF members. Of the mentorship list, we have included Chris Douglas, a PMC member from Hadoop and ASF member to help guide the community. Chris M. and Chris D. have guided a number of projects through the Incubator over the years. The other mentor includes Paul Ramirez, who has experience with the Incubator – he was a mentor for Apache Any23, and also was one of the PPMC members and eventual mentor for Apache SIS. The 70% of proposed initial committers that aren't as familiar with Apache have a broad range of experience in other open source projects, and have a deep respect and affinity for the foundation and the work that gets done here. The more experience ASF mentors and project members will help to guide them.

Community

There is an existing, established community of developers and users of this projet. This includes established communities including the Coordinated Regional Downscaling Experiment, the U.S. National Climate Assessment (NCA), the North American Regional Climate Change Assessment Program (NARCCAP), and more. The Coordinated Regional Climate Downscaling Experiment (CORDEX, http://wcrp.ipsl.jussieu.fr/cordex/about.html) is a world wide effort of coordination of regional climate downscaling (RCD) experiments driven by the World Climate Research Program (WRCP, http://www.wcrp-climate.org/index.shtml). Recently, a large number of RCD projects have been carried out on a large parts of the world. To maximize the benefits of these research activities the WCRP designed a framework (Giorgi, WMO-Bulletin, 2009) focused on "quality-control [of] data sets of RCD-based information for the recent historical past and 21st century projections, covering the majority of populated land regions on the globe". CORDEX defined different control domains (up to 10, http://cordex.dmi.dk/joomla/) for almost all the populated regions of the world in a way to standardize the experiments and make them comparable. A key region focused on Africa was also designated as the top priority by WRCP. CORDEX also provides a a series of conventions and list of variables that have to be followed by any project that wants to contribute to the experiment. Each CORDEX region has a coordinator and regional and international periodic meetings are scheduled in a way to ensure the global well being. NARCCAP is the U.S. contribution to CORDEX. From the US National Climate Assessment site, work is "being conducted under the auspices of the Global Change Research Act of 1990. The GCRA requires a report to the President and the Congress every four years that integrates, evaluates, and interprets the findings of the U.S. Global Change Research Program (USGCRP); analyzes the effects of global change on the natural environment, agriculture, energy production and use, land and water resources, transportation, human health and welfare, human social systems, and biological diversity; and analyzes current trends in global change, both human-induced and natural, and projects major trends for the subsequent 25 to 100 years."

Apache Open Climate Workbench will support all of these communities above, with an eye towards being a general purpose climate evaluation toolkit for model output and remote sensing data.

Core Developers

The initial set of developers comes from various NASA centers (JPL, and Goddard Space Flight Center), NASA HQ, various Universities participating in CORDEX (Cape Town, University of New South Wales), the Indian Institute of Tropical Meteorology, the Free Univ. Berlin), the University of California Los Angeles, and Howard University. As mentioned previously several of our developers are Apache veterans and understand how it works around here and for those that don't, they will have great mentorship.

Alignment

Our proposed effort aligns with the U.S. National Climate Assessment, the CORDEX effort, other efforts, including the Earth System Grid Federation, other climate software including the Open Climate GIS toolkit, other science portals for climate including the Climate Information Portal (CIP) at the University of Cape Town, and other related projects.

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

Known Risks

Orphaned products

Our project has a history of funding support from JPL, NASA (Applications program/ARRA, NCA, AIST), NSF (ExArch project), international investment from collaborators, and from other funding sources. The funding sources are all target future deliverables and activities, so there is little chance this software and community will be orphaned.

Inexperience with Open Source

All the initial developers have worked on open source before – 30% of the proposed initial community are experience with the ASF, and are PMC members and committers on ASF project including 6 ASF members. Our mentors are all ASF members, and we welcome any interest from additional Apache mentors in the effort. Those 70% of our project that aren't Apache committers, PMC members, or members will benefit from the leadership of the other 30% of the project.

Homogenous Developers

The initial developers come from a variety of backgrounds and with a variety of needs for the proposed framework. Everyone is used to communicating on mailing lists as the project spans timezones, international institutions and centers of excellence for climate science.

Reliance on Salaried Developers

All of the proposed initial developers are paid to work on this or related projects, but the proposed project is not the primary task for anyone.

Relationships with Other Apache Products

As mentioned above, RCMES and the Apache Open Climate Workbench already depend on Apache OODT for facade interfaces to underlying data warehouses for storing remote sensing data; and for metadata extraction and transformation. The software also uses Apache Tika for this (through a transitive dependency from OODT). In addition, we have hooks to Apache Hadoop/HIVE, as well as dependencies on Apache Sqoop for dumping out remote sensing data from MySQL and into HIVE.

A Excessive Fascination with the Apache Brand

All of us are familiar with Apache and have a respect for its brand and community. Chris Mattmann is a big proponent of Apache's sustainability factor – and it's ability to grow software communities, in an institution, or funding source neutral manner. All of the community have an extreme respect for Apache, including those in our communities who aren't necessarily trained computer scientists, but are Scientists (big "S", e.g., land, physical, Earth/Climate scientists).

Documentation

The initial RCMES code base will come from the internal JPL Subversion repository. The http://rcmes.jpl.nasa.gov at http://www.jpl.nasa.gov/ has documentation on the existing software, including links to funding support, communities, and other projects. We will continue to maintain that site at JPL, which is part of the reason for rebranding the project here at Apache with a new name to not interfere with the existing RCMES one that has a following. In addition, we hope to evolve RCMES@JPL to have increasing levels of dependency on Apache Open Climate Workbench, so that we can incrementally transition with little impact to existing customers.

In addition, JPL's Climate Data eXchange (CDX) website also has documentation on the existing software.

Initial Source

The project will start with seed code donated by NASA JPL via Mattmann and the rest of the initial committers, which consists of the Regional Climate Model Evaluation System (RCMES) toolkit, and the Climate Data eXchange (CDX) software. This will include the core Python API for RCMET, the RCMED OODT catalog project (which stores remote sensing data to MySQL/PostGIS, and HIVE), and the RCMED extractors for various climate formats. The source will also include Easy-RCMET, the Python Buildout for RCMET. In addition, we will bring along the CDX toolkit, which includes a CDX client package that performs subsetting, access, regridding of climate data; and also includes a Python Buildout installer of its own called Uber CDX.

Source and Intellectual Property Submission Plan

All seed code and other contributions will be handled through the normal Apache contribution process. Mattmann has been authorized by NASA JPL to lead the contribution of RCMES and CDX into the Incubator via his existing Apache CLA, and a Software Grant to be provided.

We will also contact other related efforts for possible cooperation and contributions.

External Dependencies

Our project depends on a number of external libraries with various licensing conditions. An initial list of such dependencies is shown below.

Library	License

NCAR NCL	MIT compat
PyNIO	MIT compat
PyNGL	MIT compat
Matplotlib	Modified PSF license
Scipy	MIT compat
NumPy	MIT compat
HDF5	BSD
NetCDF	MIT

Cryptography

The project itself will not use cryptography, but it is possible that some of the external software libraries will include cryptographic code to handle features present in various science data formats. If we need to provide an export control statement regarding cryptographic code per Apache policy, we will follow a similar approach by Mattmann in Apache Nutch and by Jukka Zitting lead this effort in Apache Tika. Mattmann is familiar with this process.

Required Resources

Mailing lists

- dev@climate.incubator.apache.org
- commits@climate.incubator.apache.org
- private@climate.incubator.apache.org

Subversion Directory

- https://svn.apache.org/repos/asf/incubator/climate

Issue Tracking

- JIRA CLIMATE (CLIMATE)

Other Resources

- CLIMATE Wiki http://cwiki.apache.org/CLIMATE
- Review Board instance - CLIMATE
- Jenkins instance - CLIMATE

Initial Committers

Name	Email	Affiliation	CLA
Chris A. Mattmann	mattmann at apache dot org	NASA Jet Propulsion Laboratory	yes
Cameron E. Goodale	goodale at apache dot org	NASA Jet Propulsion Laboratory	yes
Paul Ramirez	pramirez at apache dog org	NASA Jet Propulsion Laboratory	yes
Andrew F. Hart	ahart at apache dot org	NASA Jet Propulsion Laboratory	yes
Jinwon Kim	jkim at atmos dot ucla dot edu	UCLA Joint Institute for Regional Earth System Science and Engineering	no
Duane Waliser	duane dot waliser at jpl dot nasa dot gov	NASA Jet Propulsion Laboratory	no
Huikyo Lee	Huikyo dot Lee at jpl dot nasa dot gov	NASA Jet Propulsion Laboratory	no
Paul Loikith	Paul dot C dot Loikith at jpl dot nasa dot gov	NASA Jet Propulsion Laboratory	no
Daniel J. Crichton	crichton at apache dot org	NASA Jet Propulsion Laboratory	yes
Kim Whitehall	Kim dot D dot Whitehall at jpl dot nasa dot gov	Howard University	no
Paul Zimdars	pzimdars at apache dot org	NASA Jet Propulsion Laboratory	yes
Chris Jack	cjack at csag dot uct dot ac dot za	University of Cape Town	no
Bruce Hewitson	hewitson at csag dot uct dot ac dot za	University of Cape Town	no
Lluis Fita Borrell	l dot fitaborrell at unsw dot edu dot au	University of New South Wales	yes
Jason Evans	jason dot evans at unsw dot edu dot au	University of New South Wales	no
Estani Gonzalez	estanislao dot gonzalez at met dot fu-berlin dot de	Free University Berlin	yes
Luca Cinquini	luca dot cinquini at jpl dot nasa dot gov	NASA Jet Propulsion Laboratory	yes
J. Sanjay	sanjay at tropmet dot res dot in	Indian Institute of Tropical Meteorology	yes
M. V. S. Rama Rao	ramarao at tropmet dot res dot in	Indian Institute of Tropical Meteorology	yes
Tsengdar Lee	tsengdar dot j dot lee at nasa dot gov	NASA HQ	no

Laura Carriere	laura dot carriere at nasa dot gov	NASA Goddard Space Flight Center	no
Denis Nadeau	denis dot nadeau at nasa dot gov	NASA Goddard Space Flight Center	no
Michael Joyce	Michael dot J dot Joyce at jpl dot nasa dot gov	NASA Jet Propulsion Laboratory	yes
Shakeh Khudikyan	Shakeh dot E dot Khudikyan at jpl dot nasa dot gov	NASA Jet Propulsion Laboratory	yes
Maziyar Boustani	Maziyar dot Boustani at jpl dot nasa dot gov	NASA Jet Propulsion Laboratory	no
Suresh Marru	smarru at apache dot org	Indiana University	yes

Sponsors

Champion

- Chris Mattmann (mattmann at apache dot org)

Nominated Mentors

- Chris A. Mattmann (mattmann at apache dot org)
- Chris Douglas (cdouglas at apache dot org)
- Paul Ramirez (pramirez at apache dot org)
- Nick Kew (niq at apache dot org)
- Suresh Marru (smarru at apache dot org)

Sponsoring Entity

- Apache Incubator