Meeting the Big Data Challenges of Climate Science through Cloud–Enabled Climate Analytics–as–a–Service

MERRA Analytic Services
John Schnase
Office of Computational and Information Sciences and Technology
NASA Godard Space Flight Center

High-Performance Science Cloud
Dan Duffy
NASA Center for Climate Simulation
NASA Godard Space Flight Center
The discovery and communication of meaningful patterns in data.

There is extensive use of mathematics and statistics, the use of descriptive techniques, and predictive models to gain valuable knowledge from data …

… for example …

Business Analytics, Customer Analytics, Market Analytics, Fraud Analytics, Risk Analytics, Human Capital Analytics, Operations Analytics, Business Analytics, Customer Analytics, Market Analytics, Sales Analytics, Customer Services Analytics, Banking Analytics, Communications Analytics, Health Analytics, Insurance Analytics, Public Service Analytics, Retail Analytics, Learning Analytics, Web Analytics, Predictive Analytics, Prescriptive Analytics, Climate Analytics, and Analytics Analytics.
The analysis of large quantities of data to extract previously unknown interesting patterns.

There is extensive use of mathematics and statistics, the use of descriptive techniques, and predictive models to gain valuable knowledge from data …

… for example …

Business Data Mining, Customer Data Mining, Market Data Mining, Fraud Data Mining, Risk Data Mining, Human Capital Data Mining, Operations Data Mining, Business Data Mining, Customer Data Mining, Market Data Mining, Sales Data Mining, Customer Services Data Mining, Banking Data Mining, Communications Data Mining, Health Data Mining, Insurance Data Mining, Public Service Data Mining, Retail Data Mining, Learning Data Mining, Web Data Mining, Predictive Data Mining, Prescriptive Data Mining, Climate Data Mining, and Data Mining Data Mining.
“Analytics”

Hadoop, MapReduce, Cluster Computing, Big Data, Unstructured Data, Event Processing, Visualization …

Database, Artificial Intelligence, Bayesian Neural Networks, Genetic Algorithms, Machine Learning …

It’s all about scalability. At scale, even simple things become hard, even simple things become useful …
We’re working on the technology framework for climate analytics.

Right now, our analytics are simple ...
Climate science is a Big Data domain.
How big?

- MERRA Reanalysis Collection ~200 TB
- Total data holdings of the NASA Center for Climate Simulation (NCCS) is ~45 PB
- Intergovernmental Panel on Climate Change Fifth Assessment Report ~5 PB
- Intergovernmental Panel on Climate Change Sixth Assessment Report ~100 PB
Think friction and resonance ...

Data bigness depends on ease of use for the type of questions being asked ...

... and a particular technology may or may not help.

Query: “GSFC IS&T”

Google: 13,800 results in 0.37 secs.
Outlook: No results, about as fast.
Think **friction** and **resonance** ...  

Data bigness depends on ease of use for the type of questions being asked ...  

... and a particular technology may or may not help.

Google: 255,000,000 results in 0.36 secs.

Outlook: No results, about as fast.  
(You have to select the folder to search!)
Think friction and resonance ...

Data bigness depends on ease of use for the type of questions being asked ...

Successful interactions with data result when a resonance relationship sets up between data, technology, and use ...

Query: “garage”

Google: 255,000,000 results in 0.36 secs.

Outlook: No results, about as fast.
(You have to select the folder to search!)

Note to Microsoft — I want to know where it is, not where it’s not …
What are the critical resonance elements for climate analytics?

**High-Performance Compute/Storage Fabric**

Storage-proximal analytics
Canonical operations

*Data can’t move, analyses need horsepower, and leverage requires something akin to an analytical assembly language …*

**Data**

Relevance
Collocation

*Data have to be significant, sufficiently complex, and physically or logically co-located to be interesting and useful …*

**Exposure**

Convenience
Extensible

*Capabilities need to be easy to use and facilitate community engagement and adaptive construction …*
Climate Analytics-as-a-Service

MERRA Reanalysis

MERRA
The Modern Era
Retrospective-analysis
for Research and
Applications

MERRA products are available online through the Goddard Earth Sciences Data and Information Services Center:
http://disc.sci.gsfc.nasa.gov/mdisc/data-holdings

Find more information on MERRA at http://gmao.gsfc.nasa.gov/merra
MERRA products are available online through the Goddard Earth Sciences Data and Information Services Center:
http://disc.sci.gsfc.nasa.gov/mdisc/data-holdings

MERRA
The Modern Era
Retrospective Analysis
for Research and Applications

The GMAO works to maximize the impact of satellite observations in climate, weather and atmospheric composition prediction using comprehensive global models of the Earth's atmosphere.

Precipitation
15S - 15N
compared against GPCP
Satellite radiance
data
Conventional data & Satellite retrievals

Data Assimilated for MERRA

The volume of data ingested during a 6-hourly assimilation cycle changes dramatically over time. During the EOS era, over 4 million observations are assimilated at one time.

Climate Analytics-as-a-Service

High-Performance Compute/Storage Fabric

Storage-proximal analytics
Canonical operations

Data can’t move, analyses need horsepower, and leverage requires something akin to an analytical assembly language ...

MERRA Analytic Services

Data
Relevance
Collocation

Data have to be significant, sufficiently complex, and physically or logically co-located to be interesting and useful...

Exposure
Convenience
Extensible

Capabilities need to be easy to use and facilitate community engagement and adaptive construction...

Climate Data Services API
Modern Era–Retrospective Analysis for Research and Applications

- **Source:** Global Modeling and Assimilation Office (GMAO)

- **Input:** 114 observation types (land, sea, air, space) into “frozen” numerical model. (~4 million observations/day)

- **Output:** a global temporally and spatially consistent synthesis of 26 key climate variables. (~418 under the hood.)

- **Spatial resolution:** \(1/2^\circ \times 2/3^\circ\) longitude \(\times 42\) vertical levels extending through the stratosphere.

- **Temporal resolution:** 6-hours for three-dimensional, full spatial resolution, extending from 1979–Present.

- **~ 200 TB, but MERRA II is on the way ...**
MERRA Analytic Services

MapReduce

- MapReduce is a framework for processing parallelizable problems across huge datasets using a large number of computers.
- Computational processing can occur on data stored either in a filesystem (unstructured) or in a database (structured).
- MapReduce can take advantage of locality of data, processing data on or near the storage assets to decrease transmission of data.
- "Map" step: The master node takes the input, divides it into smaller sub-problems, and distributes them to worker nodes. A worker node may do this again in turn, leading to a multi-level tree structure. The worker node processes the smaller problem, and passes the answer back to its master node.
- "Reduce" step: The master node then collects the answers to all the sub-problems and combines them to form the output – the answer to the problem it was originally trying to solve.

 Canonical Ops Library

- We’re also creating a small set of canonical near-storage, early-stage analytical operations that represent a common starting point in many analysis workflows in many domains. For example, \( \text{avg}, \text{max}, \text{min}, \text{var}, \text{sum}, \text{count} \) operations of the general form:

\[
\text{result} \leq \text{avg}(\text{var}, (t_0,t_1), ((x_0,y_0,z_0),(x_1,y_1,z_1))),
\]

that return, in this example, the average of a variable when given a variable name, temporal extent, and spatial extent ...

- Averages over time, space, and elevation can be performed now for all MERRA variables.

Hadoop File System Organization

- Total size of the native, compressed NetCDF MERRA collection in a standard filesystem ~80 TB.
- Native MERRA files are sequenced and ingested into the Hadoop cluster in triplicated 640 MB blocks.
- Total size of MERRA/AS HDFS repository ~480 TB.

5621 lines of MapReduce code behind avg operation ...

MERRA Analytic Services

Cluster / Node Configuration

- 36 node Dell cluster, 576 Intel 2.6 GHz SandyBridge cores, 1300 TB raw storage, 1250 GB RAM, 11.7 TF theoretical peak compute capacity.
- FDR Infiniband network with peak TCP/IP speeds >20 Gbps.

Much of the MapReduce work has been building the code ecosystem to manage multidimensional binary NetCDF files ...
Climate Data Services API

CDS Client Stack

• The MERRA/AS project has been the starting point for development of the NASA Climate Data Services (CDS) Application Programming Interface (API).

• The CDS client stack can be distributed as a software package or used to build a cloud service (SaaS) or distributable cloud image.

• This approach to API design focuses on the specific analytic requirements of the climate sciences and marries the language and abstractions of collections management (OAIS) with those of high-performance analytics (MapReduce) ...

CDS References

Ingest – Submit/register a Submission Information Package (SIP).
Query – Retrieve data from a pre-determined service request (synchronous).
Order – Request data from a pre-determined service request (asynchronous).
Download – Retrieve a Dissemination Information Package (DIP).
Status – Track progress of service activity.
Execute – Initiate a service-definable extension. Allows for parameterized growth without API change.

CDS Library

Class CDSLibrary(object):

def order(self, service, params):
    cds_ws.order(service, params)

def avg(self, service, params, destination):
    sessionId = cds_ws.order(service, params)
    response = cds_ws.status(service, sessionId)
    …… Loop until result is available
    cds_ws.download(service, sessionId, destination)

CDS CLI

Welcome to the NASA GSFC CDSIO Climate Data Services (cds). Type help or ? to list commands.

$./code/README.txt

CDS Applications

- Make a request to the CDS API for data
- Use the service to download the data
- Process the data using the desired methods

CDS Scripts

#!/usr/bin/env python

import time

from cdl import CDSAPI
from we_input import WEInput
we_exp = WEInput()

# The rest of the file is run by the Python interpreter.
__doc__ = """This string is treated as the module docstring."""

service = we_exp.getService()
catalog = we_exp.getInput()
destination = we_exp.getDestination()
cds_lib = CDSAPI()
logger = cds_lib.getLogger()
start_time = time.time()

logger.debug("Generating: cd_avg_temp")
input = cds_lib.encode(catalog["cd_avg_temp_dictionary"])
cds_lib.avgs(service, input, destination)

exit()
Air Temperature, Precipitation / Avg, Max, Min / 1979–2014 / monthly means, 3–hourly

- **Traditional:** Find and order from archive (hrs?)
  Transfer ~100 GB (~1 hr, depending)
  Client-side clip/compute using GrADS
  1–1.5 days
  Server-side clipping using OPeNDAP
  (single stream op, time ??, > 2 mos)

- **MERRA/AS:** Server-side clip/compute (~24 hrs)
  Transfer final product ~1.5 GB

Takes about as long, but the scientist is free to work on other things …

So What? Where’s the Resonance?

![Arctic-Boreal Vulnerability Experiment](image)

**So What?** The rapid and widespread thawing of permafrost soils is altering the carbon cycle by releasing large quantities of carbon dioxide and methane to the atmosphere. The study of these changes is crucial for understanding the impacts of climate change.

**Where’s the Resonance?** The resonance in this context refers to the broader implications of the permafrost thawing, such as changes in ecosystem function, impacts on indigenous communities, and feedbacks to the global climate system.
Wei team used MERRA data to study four intensively irrigated regions: northern India/Pakistan, the North China Plain, the California Central Valley, and the Nile Valley.

Seasonal rates of evapotranspiration with and without irrigation over the studied areas were then compared to assess the impact of irrigation.

The data required for these calculations include precipitation, evapotranspiration, temperature, humidity, and wind at different tropospheric levels at six-hourly time steps from 1979 to 2002.

This early-stage data reduction—average values for environmental variables over specific spatiotemporal extents—is the type of data assembly that historically has been performed on the scientist’s workstation after transfers from public archives of large blocks of data.

Wei Experiment
An Estimation of the Contribution of Irrigation to Precipitation Using MERRA

Wei, et al.
- ~8.4 TB transferred from archive to local workstation (weeks)
- Clipping, averaging performed by Fortran program on local workstation (days)

MERRA/AS (Time trials in progress ...)
- Clipping, averaging performed by MERRA/AS (~28 hrs)
- ~35 GB of final product moved to local workstation

- Significant time savings in data wrangling,
- rapid screening over monthly means files takes minutes, and
- there’s a possibility of folding Dr. Wei’s modeling algorithm back into the CDS API ...
Rehabilitation Capability Convergence for Ecosystem Recovery

An Automated Burned Area Emergency Response Decision Support System for Post-fire Rehabilitation Management of Savanna Ecosystems in the Western US

Keith T. Weber
GIS Training and Research Center
Idaho State University

John L. Schnase\textsuperscript{1}, Molly E. Brown\textsuperscript{2}, and Mark Carroll\textsuperscript{2}
\textsuperscript{1}Office of Computational and Information Sciences and Technology
\textsuperscript{2}Biospheric Science Branch
NASA Godard Space Flight Center
• After a major wildfire, law requires that the federal land management agencies certify a comprehensive plan for public safety, burned area stabilization, resource protection, and site recovery.

• These BAER plans are due within 14 days of containment of a major wildfire and become the guiding document for managing the activities and budgets for all subsequent remediation efforts.

• Post-fire rehabilitation planning is a data-intensive process and requires better access to new types of data products …

e.g MERRA, SMAP, …
RECOVER

• RECOVER is a site-specific decision support system bringing together all the information necessary for post-fire rehabilitation decision-making.

• Designed in close collaboration with the US Department of Interior Bureau of Land Management (BLM) and Idaho Department of Lands (IDL).

• Uses rapid resource allocation capabilities of cloud computing to automatically gather data from various web services.
  • Earth observational data
  • Derived decision products
  • Historic biophysical layers

• **Automated data assembly** provides operational partners a complete and ready-to-use analysis environment customized for target wildfires.

• RECOVER is transforming this information-intensive process by reducing from days to a matter of minutes the time required to assemble and deliver crucial wildfire-related data.
For YouTube demonstrations, please see:

http://www.youtube.com/watch?v=LQKi3Ac7yNU  RECOVER Server
http://www.youtube.com/watch?v=SGhpDypVE  RECOVER Client
More than a dozen agency collaborators participated in the Phase 1 feasibility study.

The system was used in Idaho in six actual fires in the 2013 fire season.

More than two dozen data layers assembled on average in 60 minutes.

~ 90 sec. to automatically gather 20+ layers
~ 60 min. to manually assemble the remaining specialized, site-specific layers

<table>
<thead>
<tr>
<th>Fire</th>
<th>Start Date</th>
<th>End Date</th>
<th>Acres Burned</th>
<th>RECOVER Response Time (min)</th>
<th>RECOVER Client URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crystal</td>
<td>15-Aug-06</td>
<td>31-Aug-06</td>
<td>220,000</td>
<td>N/A</td>
<td><a href="http://naip.giscenter.isu.edu/recover/CrystalFire">http://naip.giscenter.isu.edu/recover/CrystalFire</a></td>
</tr>
<tr>
<td>Charlotte</td>
<td>2-Jul-12</td>
<td>10-Jul-12</td>
<td>1,029</td>
<td>N/A</td>
<td><a href="http://naip.giscenter.isu.edu/recover/CharlotteFire">http://naip.giscenter.isu.edu/recover/CharlotteFire</a></td>
</tr>
<tr>
<td>2 ½ Mile</td>
<td>2-Jul-13</td>
<td>3-Jul-13</td>
<td>924</td>
<td>30</td>
<td><a href="http://naip.giscenter.isu.edu/recover/2nHalfMileFire">http://naip.giscenter.isu.edu/recover/2nHalfMileFire</a></td>
</tr>
<tr>
<td>Mabey</td>
<td>8-Aug-13</td>
<td>19-Aug-13</td>
<td>1,142</td>
<td>120</td>
<td><a href="http://naip.giscenter.isu.edu/recover/MabeyFire">http://naip.giscenter.isu.edu/recover/MabeyFire</a></td>
</tr>
<tr>
<td>Incendiary Creek</td>
<td>18-Aug-13</td>
<td>n/a</td>
<td>1,100</td>
<td>90</td>
<td><a href="http://naip.giscenter.isu.edu/recover/IncendiaryFire">http://naip.giscenter.isu.edu/recover/IncendiaryFire</a></td>
</tr>
<tr>
<td>Ridgetop</td>
<td>28-Jul-12</td>
<td>n/a</td>
<td>16,616</td>
<td>4</td>
<td><a href="http://naip.giscenter.isu.edu/recover/Ridgetop_v2fire/">http://naip.giscenter.isu.edu/recover/Ridgetop_v2fire/</a></td>
</tr>
</tbody>
</table>
Nadeau’s Standardized Temperature Anomaly ...

- **Period:** 1 month
- **Collection:** instM_3d_ana_Np
- **Time span:** January — December 2011
- **Coverage:** Global
- **Levels:** 1 — 42 (0.1 hPa — 1000 hPa)

- **Traditional:** Find and order from archive (hrs?)
  Transfer ~10 GB (~15 min, depending)
  Client-side clip/compute using GrADS
  1-1.5 days

- **MERRA/AS:** One line in a python script *
  3 minute run time
  Final product ~0.5 GB

*Will be added to CDS Library...
Climate Analytics—as—a—Service

Who’s interested?

• Energy
• Education
• Agriculture
• Climate analytics
• Insurance industry
• Department of Interior
• The White House (climate.gov)
Next steps

- Beta testing, add other reanalyses
- Operational deployment via Climate Data Services …
Meeting the Big Data Challenges of Climate Science through Cloud-Enabled Climate Analytics-as-a-Service

MERRA Analytic Services

John Schnase
Office of Computational and Information Sciences and Technology
NASA Godard Space Flight Center

High-Performance Science Cloud

Dan Duffy
NASA Center for Climate Simulation
NASA Godard Space Flight Center