

# Climate Analytics as a Service

*John Schnase*

*Office of Computational and Information Sciences and Technology*

*NASA Goddard Space Flight Center*



# Earth Science Data Analytics

Refers to the processes and workflows useful for processing data specifically from the perspective of scientific needs

Includes examining, preparing, reducing, and analyzing large amounts of spatial, temporal, or spectral data encompassing a variety of data types to uncover patterns, correlations, and other information, to better understand our Earth.

- Data Preparation – Preparing heterogeneous data so that they can be jointly analyzed
- Data Reduction – Correcting, ordering, and simplifying data in support of analytic objectives
- Data Analysis – Applying techniques to derive results

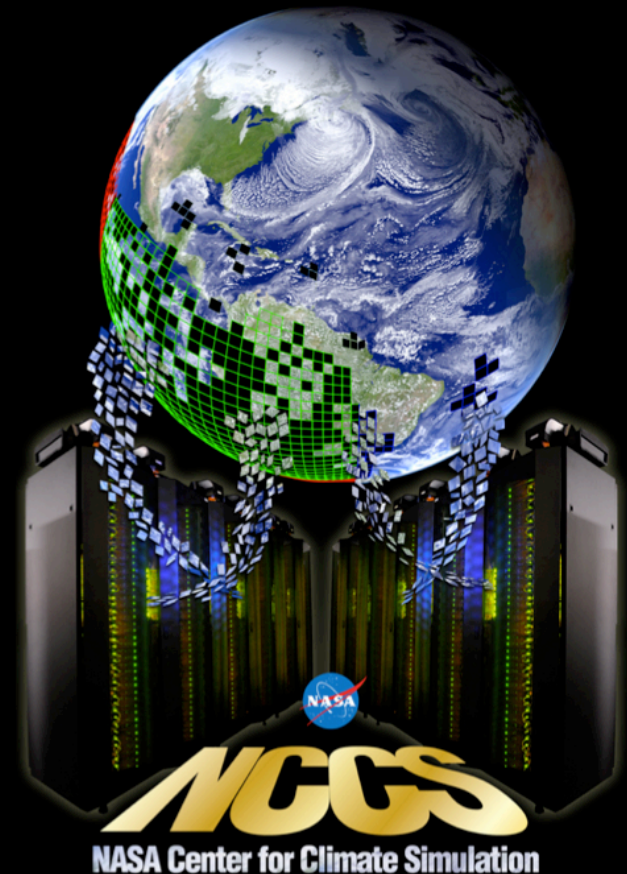


# Climate Science Data Analytics

Refers to the processes and workflows useful for processing data specifically from the perspective of **climate science's** needs

Includes examining, **preparing**, **reducing**, and analyzing **large amounts** of **spatiotemporal** data encompassing a variety of **climate model outputs** to uncover patterns, correlations, and **other information**, to better understand our Earth.

- Data Preparation – Preparing heterogenous **climate model outputs** so that they can be jointly analyzed
- Data Reduction – Correcting, ordering, and **simplifying data in support of analytic objectives** <=
- Data Analysis – **Applying techniques** to derive results of value to a growing community of users and applications within and beyond the climate research community



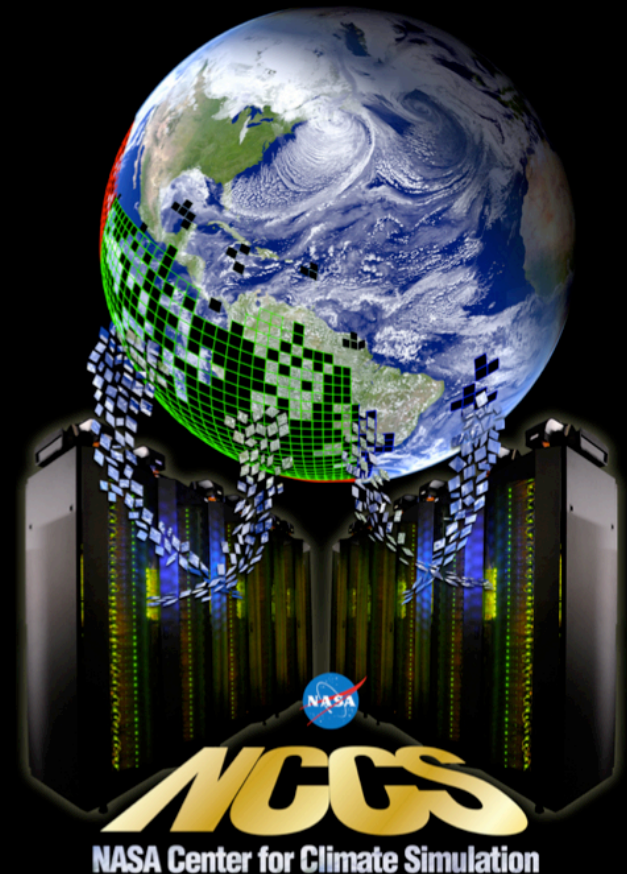


# Climate Analytics as a Service

*How can data analytics best serve the climate science community?*

We started with what we knew and what we knew our customers needed, and we set the following goals:

- Use the analytics tier of the new scientific software stack to enable real and virtual collections of the community's most commonly used information products
- Deliver those products in a highly personalized and tailored form
- Deliver those products fast
- Focus on the Big Data challenge of data assembly ...

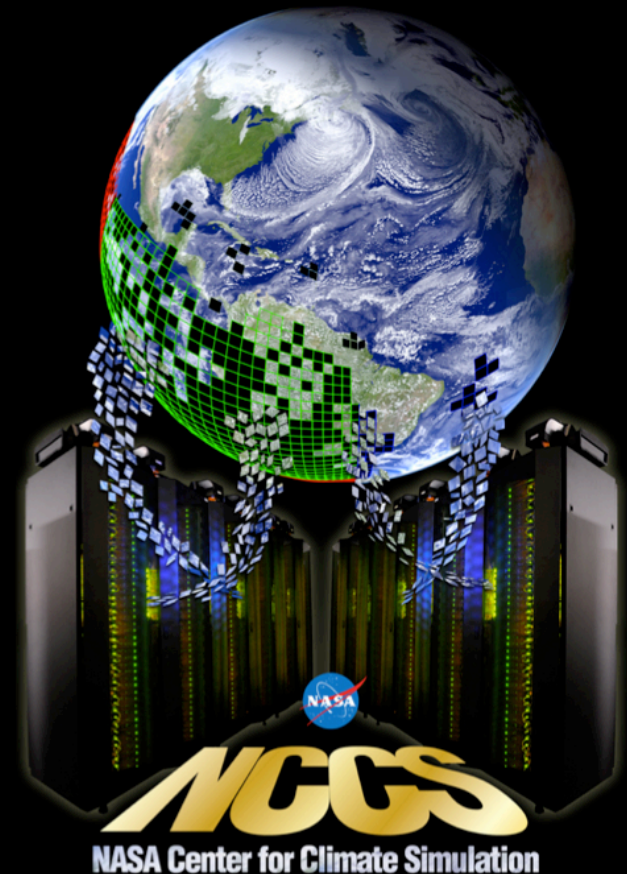




# Climate Analytics as a Service

As a starting place, we chose the following approach:

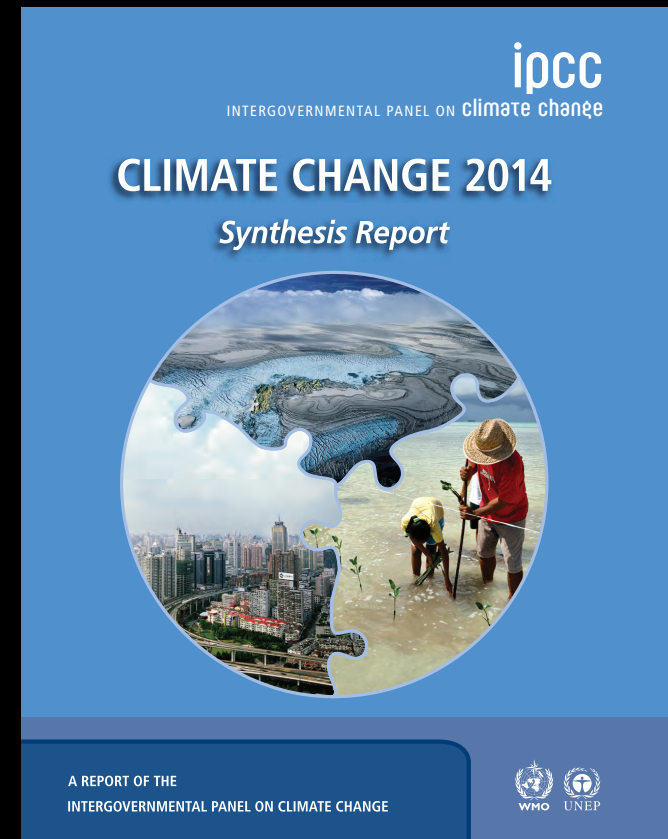
- Do simple things fast  
*(Canonical Ops)*
  - Do work near the data source  
*(Compute-Storage Fabric / Collections Reference Model)*
  - Apply capabilities to an interesting and useful climate science dataset  
*(MERRA Reanalysis)*
  - Create a service that enables community construction of advanced capabilities  
*(CDSlib Python Library)*
- 
- Improve efficiencies in upstream data processes of climate science workflows  
*(Data Wrangling)*
- 
- Create a context for developing more advanced approaches ...



# Rationale

The Intergovernmental Panel on Climate Change (IPCC) is the leading international body for the assessment of climate change

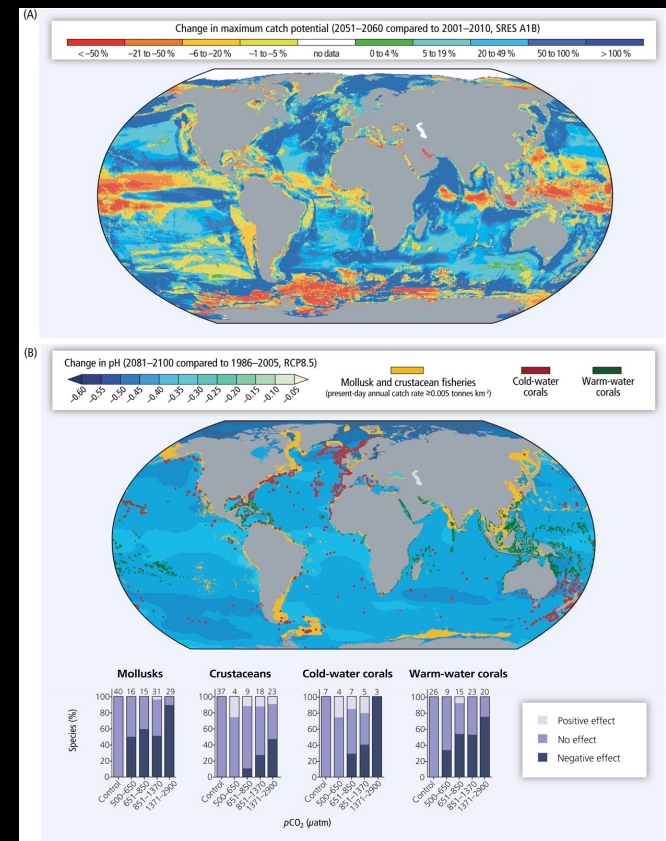
- Open to all members of the United Nations and World Meteorological Organization
- Currently 195 member countries
- Provides world with a clear view of the current state of scientific knowledge through its Assessment Reports
- 2000+ scientists contributed to IPCC's Fifth Assessment Report (AR5)
- IPCC Assessment Reports provide the basis for environmental policy-making throughout the world ...



# Rationale

IPCC's Fifth Assessment Report contained hundreds of findings distributed across 5 printed volumes and 1820 pages, but ...

- There were relatively few classes of findings
  - Statements about past, present, and future values of climate variables
  - Statements about climatologies — the maximum, minimum, and average values of those climate variables over given periods of time
  - Statements about trends — how those variables and climatologies change over time
  - Statements about anomalies — how the values of variables, climatologies, and trends at one time or place might depart from the corresponding variables at another time or place

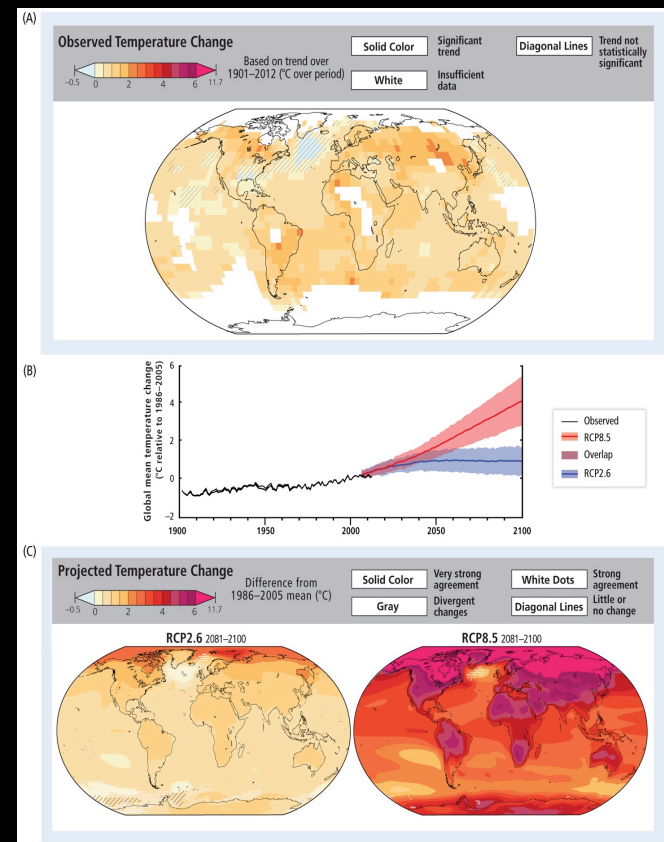
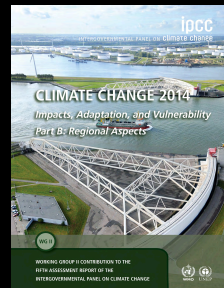
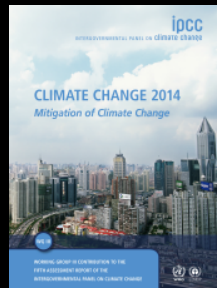
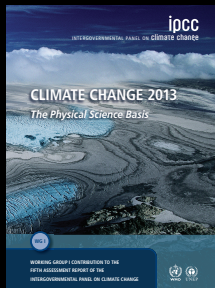




# Rationale

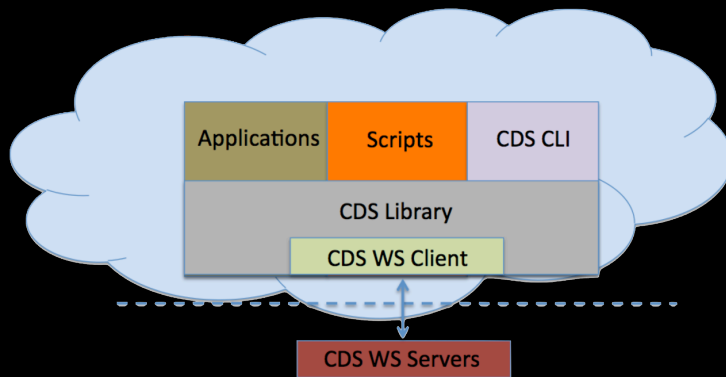
A small collection of data attributes provides the basis for a remarkable amount of intellectual work in the discipline ...

- AR5's published work
  - Contains 4 million words (50,000 unique)
  - Eight words collectively account for 12,000 direct references to climate model data: *maximum, minimum, average, variance, difference, climatology, anomaly, and trend*
  - Over halve of those references are “trend” ...



# MERRA Analytic Services

# Climate Data Services Python Library



# Data

Relevance  
Co-location

*Data have to be significant, complex,  
and physically co-located for CAaaS  
to be useful right now ...*

# Exposure

## Convenience Extensibility

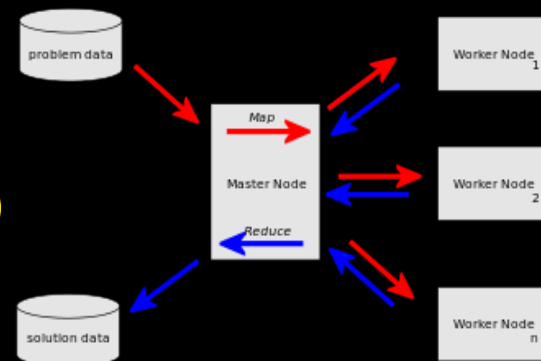
*Capabilities need to be easy to use and facilitate community engagement and adaptive construction ...*

*What are the critical  
Climate Analytics as a Service  
elements?*

# Compute-Storage Fabric

High-performance analytics  
Canonical operations

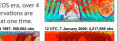
*Data can't move, analyses need horsepower, and leverage requires something akin to an analytical assembly language ...*



# MapReduce

### 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.



### Conventional data & Satellite retrievals

Observation Type	Instrument	Resolution	Frequency	Lat	Lon	Alt
Temperature	AMSU	2500x2500	10000	90S	90N	0-1000
Humidity	AMSU	2500x2500	10000	90S	90N	0-1000
Clouds	AMSU	2500x2500	10000	90S	90N	0-1000
Surface temperature	AMSU	2500x2500	10000	90S	90N	0-1000
Surface humidity	AMSU	2500x2500	10000	90S	90N	0-1000
Surface wind	AMSU	2500x2500	10000	90S	90N	0-1000
Surface pressure	AMSU	2500x2500	10000	90S	90N	0-1000
Surface albedo	AMSU	2500x2500	10000	90S	90N	0-1000
Surface ice extent	AMSU	2500x2500	10000	90S	90N	0-1000
Surface snow extent	AMSU	2500x2500	10000	90S	90N	0-1000
Surface vegetation	AMSU	2500x2500	10000	90S	90N	0-1000
Surface soil moisture	AMSU	2500x2500	10000	90S	90N	0-1000
Surface ocean color	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice extent	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice concentration	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice thickness	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice velocity	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice age	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice volume	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice density	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice salinity	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice temperature	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice pressure	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice height	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice area	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice perimeter	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice shape	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice color	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice texture	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice roughness	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice smoothness	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice wetness	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice dryness	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice stickiness	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice brittleness	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice flexibility	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice rigidity	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice elasticity	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice plasticity	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice viscosity	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice fluidity	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice malleability	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice ductility	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice tenacity	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice toughness	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice hardness	AMSU	2500x2500	10000	90S	90N	0-1000
Surface sea ice softness	AMSU	2500x2500	10000	90S	90N	0-1000

## MERRA2 Reanalysis

# The MERRA Reanalysis

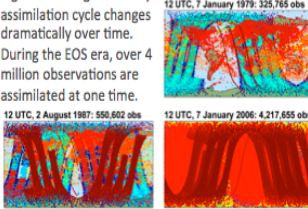
## Modern Era-Retrospective Analysis for Research and Applications

- 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:  $1/2^\circ$  latitude  $\times$   $2/3^\circ$  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 ...

**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.

12 UTC, 7 January 1979: 325,765 obs  
12 UTC, 2 August 1987: 550,602 obs  
12 UTC, 7 January 2006: 4,217,655 obs



**Conventional data & Satellite retrievals**

Data Source/Type	Period	Data Supplier
Radiosondes	1970 – present	NCEP
PIBAL winds	1970 – present	NCEP
Wind profiles	1992/5/14 – present	UCAR
Conventional, ASDAR and MDCKR aircraft rep.	1970 – present	NCEP
Drosondes	1970 – present	NCEP
PAOB	1978 – 2010/8	NCEP
GVS, METEOSAT, cloud drift IR & visible winds	1977 – present	NCEP
GOES cloud drift winds	1997 – present	NCEP
EOS/Terra/MODIS winds	2002/7/01 – present	NCEP
EOS/Aqua/MODIS winds	2003/9/01 – present	NCEP
Surface ship and buoy observations	1977 – present	NCEP
Surface land observations	1970 – present	NCEP
SSM/I V6 wind speed	1987/7 – present	RSS
SSM/I rain rate	1987/7 – present	GSFC
TMI rain rate	1997/12 – present	GSFC
QuikSCAT surface winds	1999/7 – 2009/9	JPL
ERS-1 surface winds	1991/8/5 – 1996/5/21	CERSAT
ERS-2 surface winds	1996/3/19 – 2001/1/17	CERSAT
SeaWiFS ocean (V6 retrievals)	1978/10 – present	GSFC

**Satellite radiance data**

Data Source/Type	Period	Data Provider
TOVS (TIROS N, N-6, N-7, N-8)	1978/10/30 – 1985/01/01	NCAR
(AT)OVS (N-9, N-10, N-11, N-12)	1985/01/01 – 1997/07/14	NESDIS/NCAR
ATOVS (N-14, N-15, N-16, N-17, N-18)	1995/01/19 – present	NESDIS
EOS/Aqua	2002/10 – present	NESDIS
SSM/I V6 (F08, F10, F11, F13, F14, F15)	1987/7 – present	RSS
GOES Sounder $T_6$	2001/01 – present	NCEP

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 was conducted at the NASA Center for Climate Simulation (NCCS).

*The GMAO works to maximize the impact of satellite observations in the analysis and prediction of climate and weather through integrated Earth system modeling and data assimilation.*

**GLOBAL MODELING AND  
ASSIMILATION OFFICE**

Code 610.1

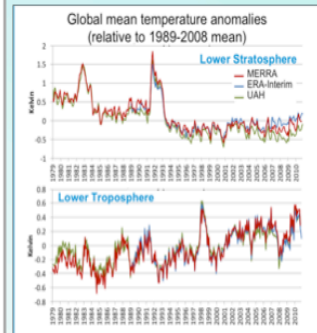
NASA/Goddard Space Flight Center

Greenbelt, MD 20771

<http://gmao.gsfc.nasa.gov>

## MERRA

The Modern-Era  
Retrospective analysis  
for Research and  
Applications



Global Modeling  
and  
Assimilation Office

Goddard Space Flight Center



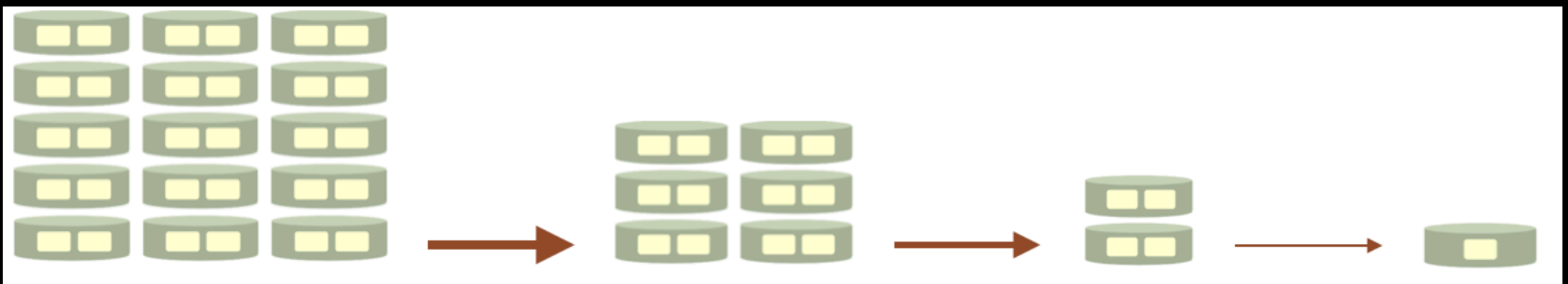


# Canonical Ops

- Our “canonical operations” — or *microservices* — take as input (1) a variable name, (2) a spatial extent, and (3) a temporal extent
- Then (1) perform arithmetic operations, (2) extract spatiotemporal subsets, and (3) model data by computing the descriptive statistics of limits, central tendency, and dispersion in the spatial and temporal domains, e.g.:

$$result \leftarrow average(var, (t_0, t_1), ((x_0, y_0, z_0), (x_1, y_1, z_1))),$$

- Canonical ops can be combined under programmatic control to create higher-order products such as climatologies, trends, anomalies, and tailored workflows
- Built-in canonical ops improve the efficiency of data reduction and assembly by stratifying work along the workflow chain ...



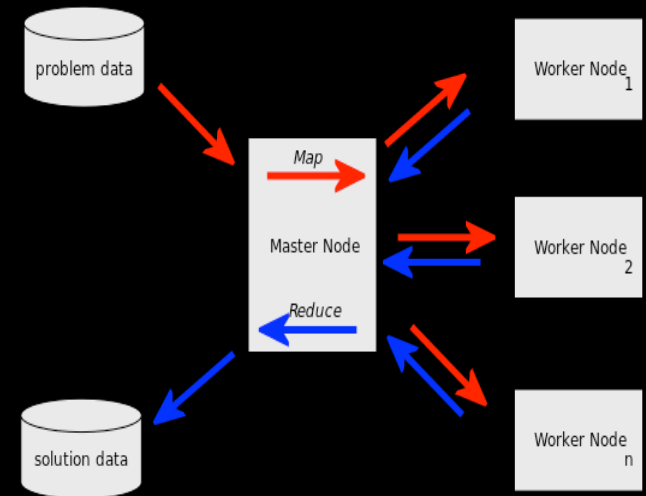
- *Large amounts of unstructured data*
- *Simple, common, general-purpose operations*

- *Highly structured, tailored, reduced, refined analytic products*
- *Specialized tools, models, operations*

# MapReduce

We use MapReduce to implement the canonical ops ...

- 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.



# Collections Reference Model

High-performance canonical ops enable virtual collections of realizable objects ...

**Collection:** MERRA.inst3\_3d\_asm\_Cp (3 Hr)

**Variables (14):** Temperature (T)  
Geopotential Height(H)  
Relative Humidity(RH)  
Specific Humidity (QV)  
Eastward Wind(U)  
Northward Wind (V)  
Vertical Pressure Velocity (OMEGA)  
Ozone Mixing Ratio(O3)  
Ertel Portential Vorticity (EPV)  
Cloud Liquid Water Mixing Ratio(QL)  
Cloud Ice Mixing Ratio (QI)  
Sea-level Pressure (SLP)  
Surface Pressure (PS)  
Surface Geopotential (PHIS)

**Time Span:** 19790101–Present

**Time Interval:**

Year	(1979, 1980, ..., 2015)	(Mx, Mn, Av, Va)
Season	(JJA, SON, ..., MAM)	(Mx, Mn, Av, Va)
Month	(Jan, Feb, ..., Dec)	(Mx, Mn, Av, Va)
Week	(1, 2, ..., 4)	(Mx, Mn, Av, Va)
Day	(Mo, Tu, ..., Su)	(Mx, Mn, Av, Va)
6 Hr	(1, 2, ..., 4)	(Mx, Mn, Av, Va)
3 Hr	(1, 2, ..., 8)	(Mx, Mn, Av, Va)
1 Hr	(1, 2, ..., 24)	(Mx, Mn, Av, Va)

**Spatial Coverage:** Global

**Horizontal Resolution:** 1.25 x 1.25 (deg lat x deg lon)

**Vertical Levels(42):** 1000–975x25/650–100x50/70–10x10/7–1x1/0.7,0.5,0.4,0.3,0.1 (hPa)

*“Big Data” Implications ...*

*This makes possible a converged approach to analytics and archive management.*

*If you know an object is computable, discovery becomes an assertion rather than a query.*

*Queries cause results to be created, rather than found and retrieved ...*

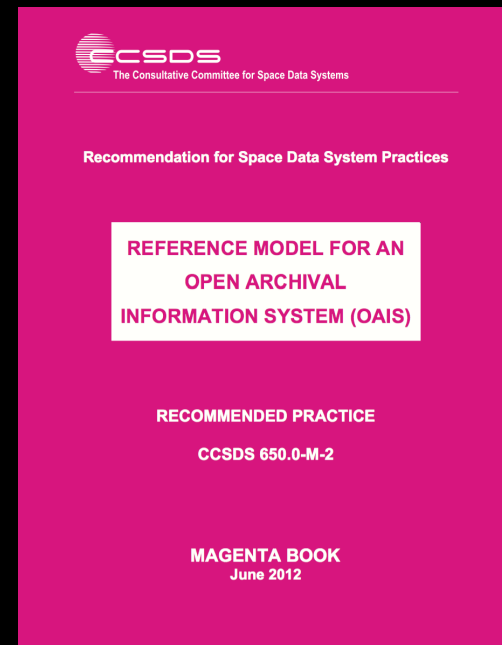
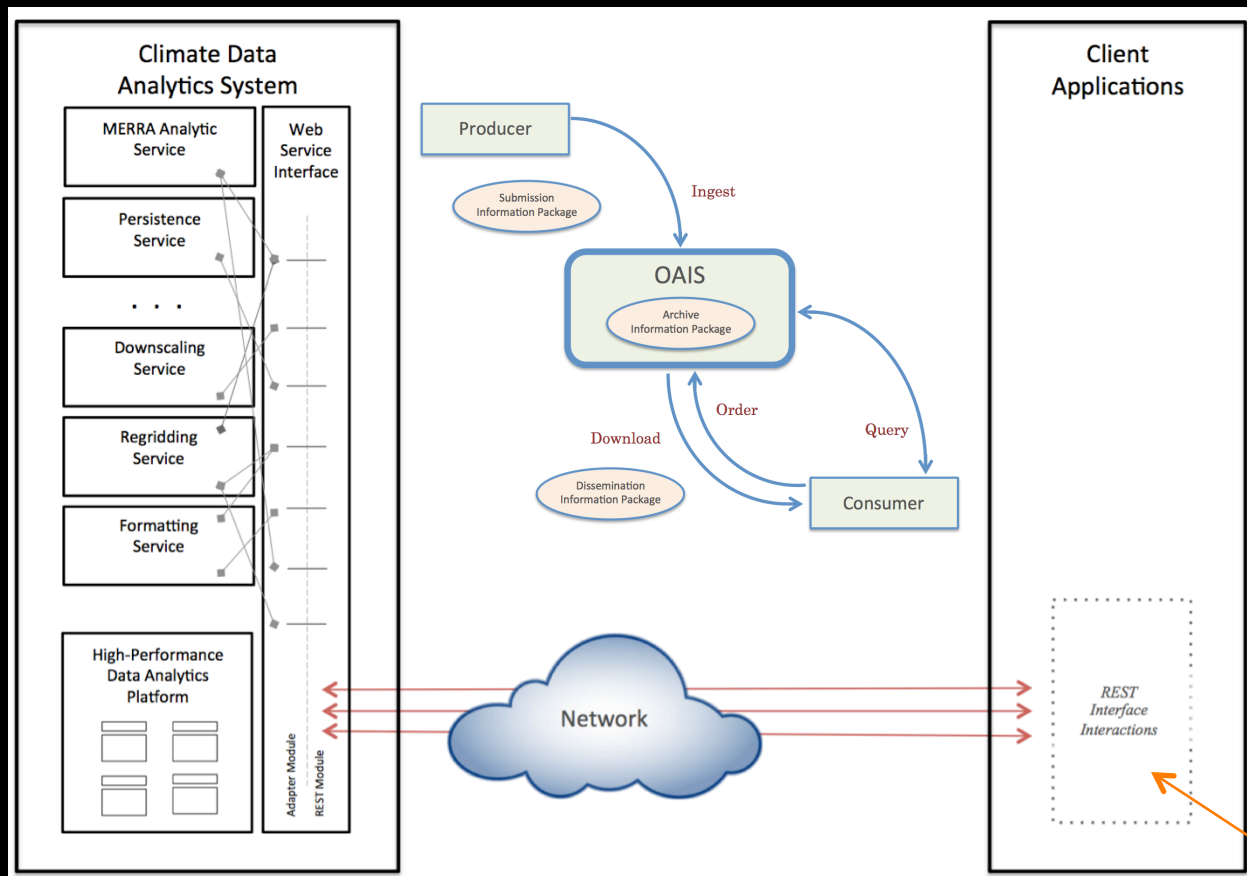
	Base Collection	(Real)
	Realizable Objects	(Virtual)
	Not computed	



# Web Service API

- Capabilities exposed through two interfaces:
  - RESTful Web service API

- Based on the dataflow interactions of the Open Archival Information Systems (OAIS) Reference Model
- Positioned to integrate analytics into digital preservation systems

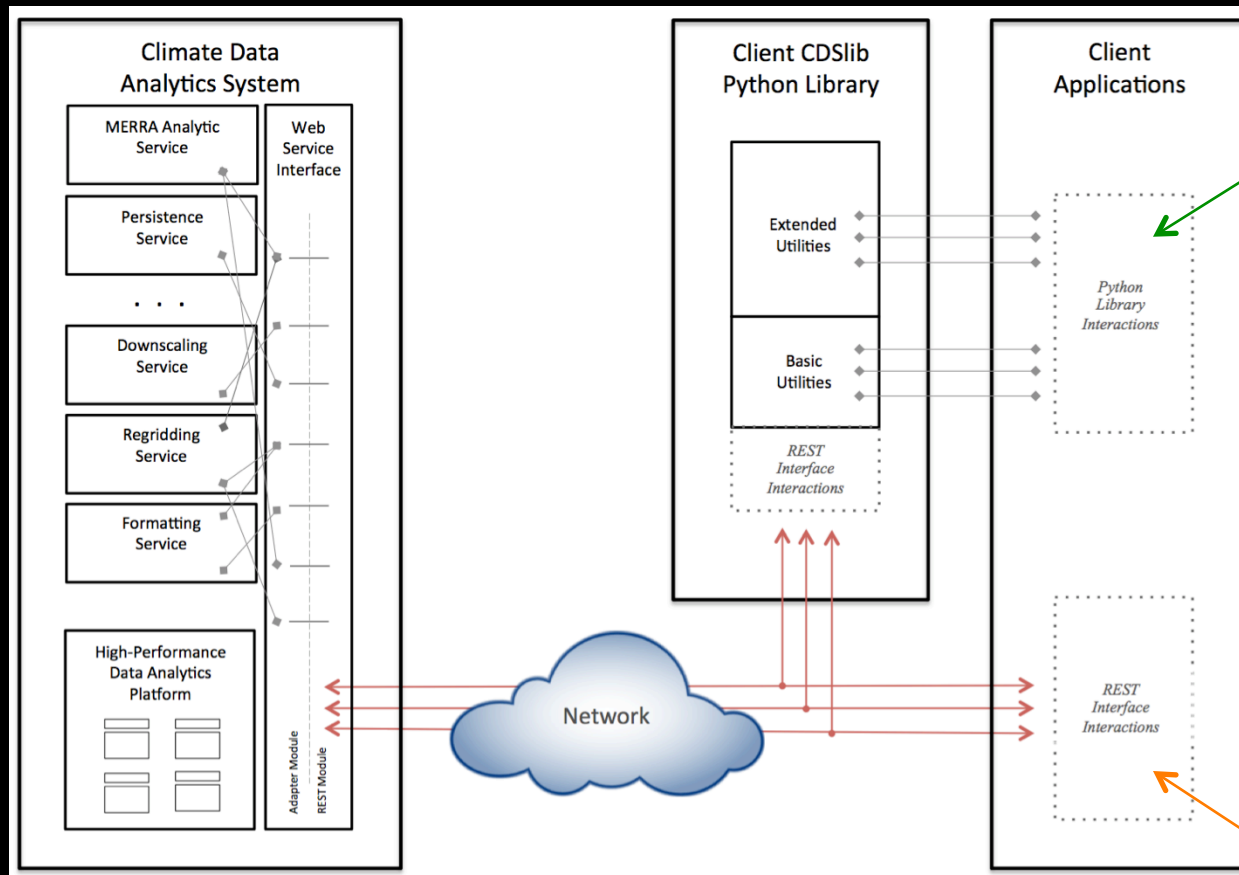


Example Web service call ...

[http://skyportal.sci.gsfc.nasa.gov/cds/mas/res/order.php?service=M2AS&service\\_request=GetVariableByCollection\\_Operation\\_TimeRange\\_SpatialExtent\\_VerticalExtent&job\\_name=res\\_Sprint\\_1&collection=instM\\_3d\\_ana\\_Np&request=GetVariableBy\\_TimeRange\\_SpatialExtent\\_VerticalExtent&variable\\_list=T&operation=avg&time\\_span=201101-201102&time\\_interval=fullinterval\(fullspan\)&spatial\\_extent=-125,24,-66,50&spatial\\_resolution=native&vertical\\_extent=fullextent](http://skyportal.sci.gsfc.nasa.gov/cds/mas/res/order.php?service=M2AS&service_request=GetVariableByCollection_Operation_TimeRange_SpatialExtent_VerticalExtent&job_name=res_Sprint_1&collection=instM_3d_ana_Np&request=GetVariableBy_TimeRange_SpatialExtent_VerticalExtent&variable_list=T&operation=avg&time_span=201101-201102&time_interval=fullinterval(fullspan)&spatial_extent=-125,24,-66,50&spatial_resolution=native&vertical_extent=fullextent)

# Python Library

- Capabilities exposed through two interfaces:
  - RESTful Web service API
  - Client Climate Data Service Python Library (CDSlib)



## Example Python library call ...

```
# Average North America Temp 201101-201102
variable_list = "T"
service = "M2AS"
collection = "instM_3d_ana_Np"
time_span = "time_span=198601-199512"
time_interval = "time_interval=fullinterval(fullspan)"
spatial_extent = "spatial_extent=-125,24,-66,50"
spatial_resolution = "spatial_resolution=native"
vertical_extent = "vertical_extent=fullextent"
```

## # Wei Experiment Script

```
input = prepSprint("./wei_parameters")
output = "./out"
```

```
cds_lib.avg(service, input, output)
```

## Example Web service call ...

```
http://skyportal.sci.gsfc.nasa.gov/cds/mas/res/order.php?
service=M2AS&service_request=GetVariableByCollection
_Operation_TimeRange_SpatialExtent_VerticalExtent&&j
ob_name=res_Sprint_1&collection=instM_3d_ana_Np&re
quest=GetVariableBy_TimeRange_SpatialExtent_Vertical
Extent&variable_list=T&operation=avg&time_span=20110
1-201102&time_interval=fullinterval(fullspan)&spatial_ext
ent=-125,24,-66,50&spatial_resolution=native&vertical_ex
tent=fullextent
```

# Single Reanalysis Estimation of the Contribution of Irrigation to Precipitation

*J. Wei study, tailored climatologies use case*

## Study Areas

- Nile Valley
- North China
- California Central Valley
- Northern India/Pakistan

## Other Requirements

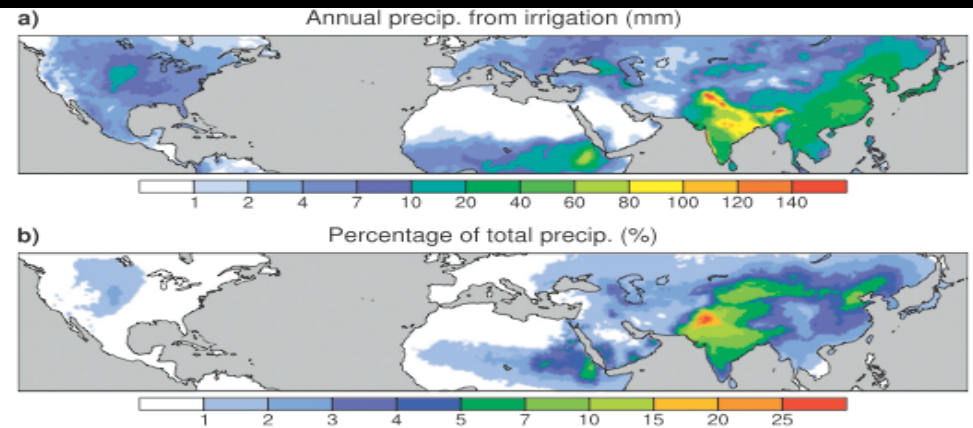
- 1979 – 2002
- 6-hr time steps
- 18 atmospheric levels

## Variables Needed

- Humidity
- Wind speed
- Temperature

## Data Wrangled:

- $23 \times 365 \times 4 \times 4 \times 18 \times 3$   
= **7,253,280 layers** ...



FEBRUARY 2013

WEI ET AL.

275

## Where Does the Irrigation Water Go? An Estimate of the Contribution of Irrigation to Precipitation Using MERRA

JIANGFENG WEI\*

Center for Ocean–Land–Atmosphere Studies, Calverton, Maryland

PAUL A. DIRMEYER

Department of Atmospheric, Oceanic and Earth Sciences, George Mason University, Fairfax, Virginia, and Center for Ocean–Land–Atmosphere Studies, Calverton, Maryland

DOMINIK WISSER

Department of Physical Geography, Utrecht University, Utrecht, Netherlands

MICHAEL G. BOSILOVICH

Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, Maryland

DAVID M. MOCKO

SAIC and Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, Maryland

(Manuscript received 24 May 2012, in final form 21 September 2012)

### ABSTRACT

Irrigation is an important human activity that may impact local and regional climate, but current climate model simulations and data assimilation systems generally do not explicitly include it. The European Centre for Medium-Range Weather Forecasts (ECMWF) Interim Re-Analysis (ERA-Interim) shows more irrigation signal in surface evapotranspiration (ET) than the Modern-Era Retrospective Analysis for Research and Applications (MERRA) because ERA-Interim adjusts soil moisture according to the observed surface temperature and humidity while MERRA has no explicit consideration of irrigation at the surface. But, when compared with the results from a hydrological model with detailed considerations of agriculture, the ET from both reanalyses show large deficiencies in capturing the impact of irrigation. Here, a back-trajectory method is used to estimate the contribution of irrigation to precipitation over local and surrounding regions, using MERRA with observation-based corrections and added irrigation-caused ET increase from the hydrological model. Results show substantial contributions of irrigation to precipitation over heavily irrigated regions in Asia, but the precipitation increase is much less than the ET increase over most areas, indicating that irrigation could lead to water deficits over these regions. For the same increase in ET, precipitation increases are larger over wetter areas where convection is more easily triggered, but the percentage increase in precipitation is similar for different areas. There are substantial regional differences in the patterns of irrigation impact, but, for all the studied regions, the highest percentage contribution to precipitation is over local land.

\* Current affiliation: Jackson School of Geosciences, The University of Texas at Austin, Austin, Texas.

Corresponding author address: Jiangfeng Wei, Jackson School of Geosciences, The University of Texas at Austin, 2275 Speedway C9000, Austin, TX 78712.  
E-mail: jwei@utexas.edu

DOI: 10.1175/JHM-D-12-079.1

© 2013 American Meteorological Society

### 1. Introduction

Irrigation is an important human activity that has the potential to impact local and regional climate through the hydrological cycle and surface energy balance (e.g., Chase et al. 1999; Pielke et al. 2011). About two-thirds of the global freshwater withdrawals from surface and underground are used for agriculture (Shiklomanov 2000),



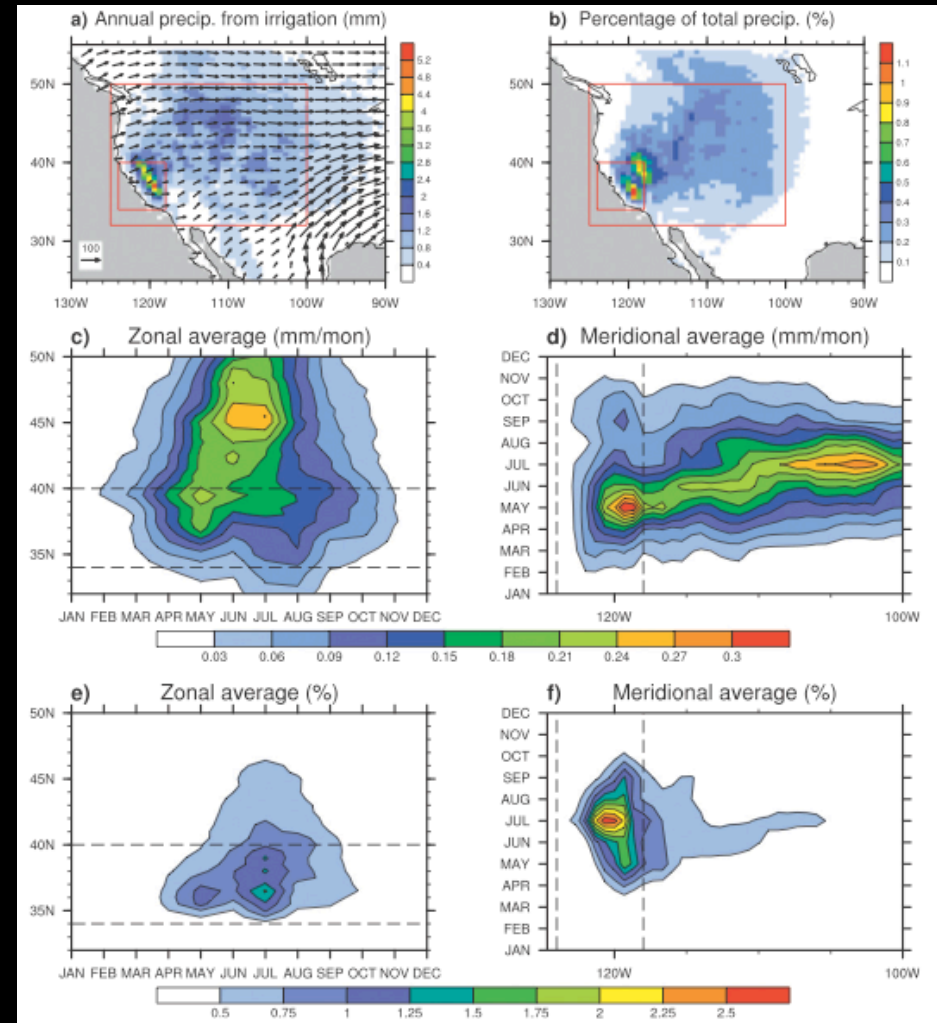
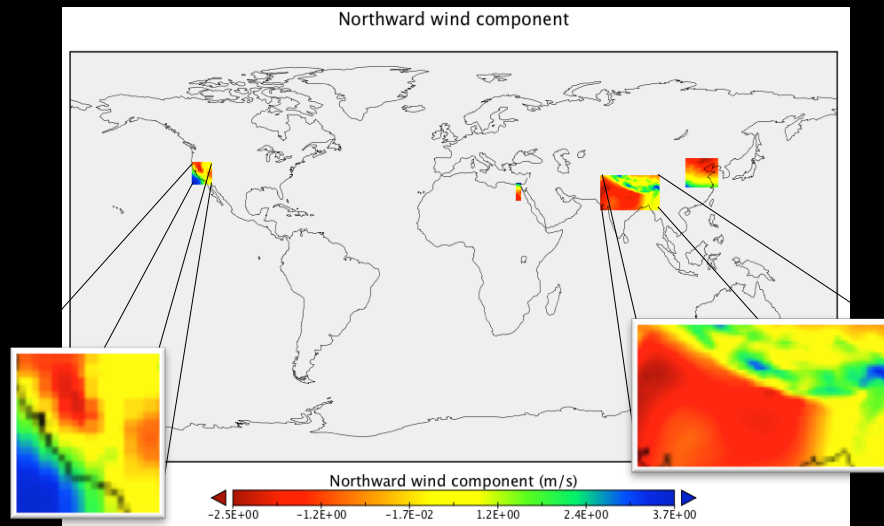
# Single Reanalysis Estimation of the Contribution of Irrigation to Precipitation

## Traditional Approach

- 8.4 TB moved from archive (days)
- Clipping / averaging (days – weeks)

## With MERRA/AS and CDSlib ...

- Clipping / averaging (2.5 minutes)
- 500 MB of final product moved to local workstation in minutes



# Single Reanalysis Intercomparison of Global Temperature

*D. Nadeau demo, simple anomaly use case*

## 2011 Monthly Temperature Anomaly

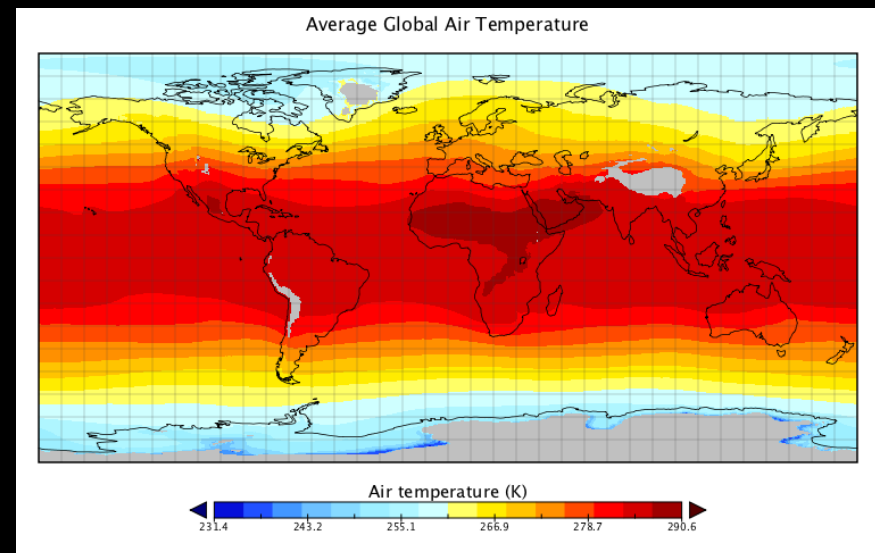
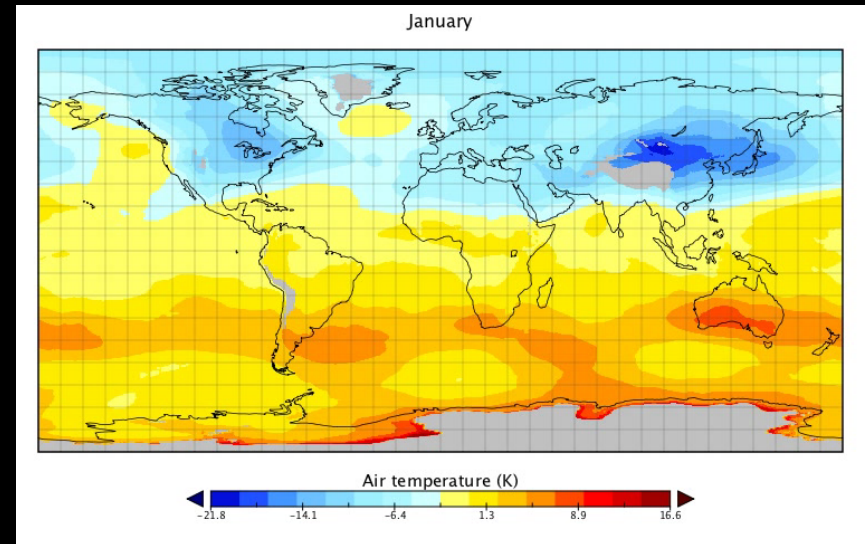
- Coverage: Global
- Period: 1 month
- Collection: instM\_3d\_ana\_Np
- Time span: January — December 2011
- Levels: 1 – 42 (0.1 hPa – 1000 hPa)

## Traditional Approach

- Find and order from archive (hours – days)
- Transfer ~10 GB (hours)
- Client-side clip/compute, GrADS (days)

## With MERRA/AS and the CDSlib ...

- One line in a python script
- **3 minutes** run time
- Final product ~0.5 GB



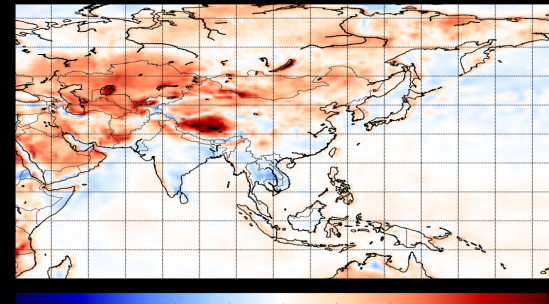
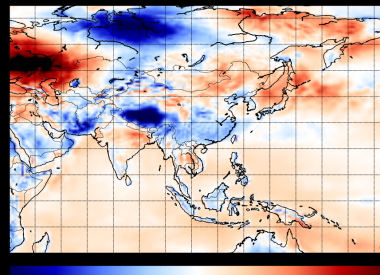
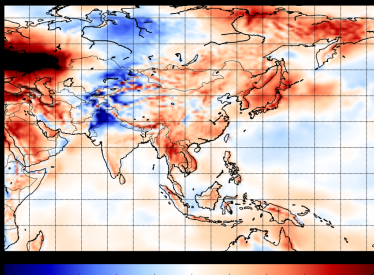
# Ensemble Reanalysis

- Our next step: MERRA/AS => Reanalysis Ensemble Service (RES)
  - RES will support data from six major reanalysis collections, a core suite of commonly-used ensemble operations, and accessibility through a RESTful web services API and a client CDSlib Python library
  - Will improve access to multiple reanalyses and enable reanalysis intercomparison, ensembled products with reduced error and bias, and support uncertainty quantification
- Participating Reanalyses
  - Modern-Era Retrospective Analysis for Research and Applications (MERRA-2)\*
  - ECMWF Interim Reanalysis (ERA-Interim)
  - NOAA NCEP Climate Forecast System Reanalysis (CFSR)
  - NOAA ESRL 20th Century Reanalysis (20CR)
  - Japanese 25-Year Reanalysis (JRA-25)
  - Japanese 50-Year Reanalysis (JRA-55)
- Outputs
  - Single source and ensembled analytic results
  - Uncertainty quantification packages
    - Related observational data
    - Other comparative data (e.g. Bioclim, Worldclim, etc.)

# Multiple Reanalysis Intercomparison of Global Precipitation

## *D. Nadeau experiment, 2014 AGU ensemble anomaly use case*

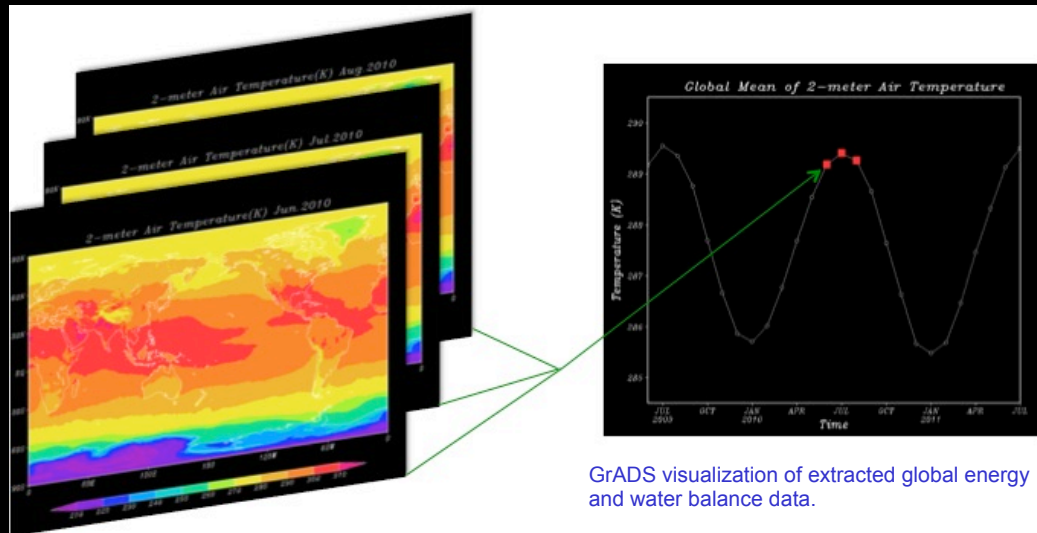
- Workflow uses prototype RES and CDSlib V2.0 to:
  - Calculate the average precipitation from 3 collections (MERRA, CFSR, ECMWF) for a long-term (34-year) temporal span (1979-2013) across the entire globe at the Earth's surface.
  - Calculate the average precipitation from the same 3 collections for a yearly (2010) temporal span across the entire globe at the Earth's surface.
  - Normalize the results (i.e., re-grid)
  - Calculate the ensemble average across the normalized results for a).
  - Calculate the difference (anomaly) between the re-gridded 2010 average and the overall ensemble average (all collections) of the 3 collections.
- Traditional data assembly time 2 days
- Reanalysis Ensemble Service data assembly workflow takes **50 – 90 seconds** ...



# Multiple Reanalysis Intercomparison of Global Energy and Water Balance

*M. Bosilovich research, complex ensemble anomaly use case (in progress ...)*

- Requires MERRA, MERRA2, JRA55, CFSR, ERA-20C, NOAA-20CR reanalyses and multiple observations for 1979-2014
- Calculate global and regional average radiations and water cycle components
- Calculate 12-month moving average to remove high frequency variability
- Calculate ensemble average, anomaly, and trend analysis on time series

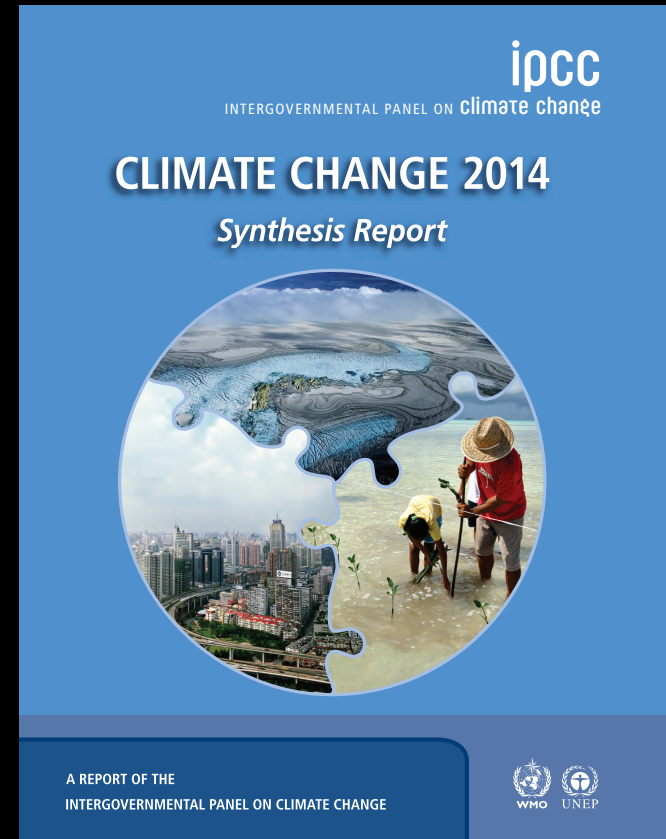




# Closing Thought

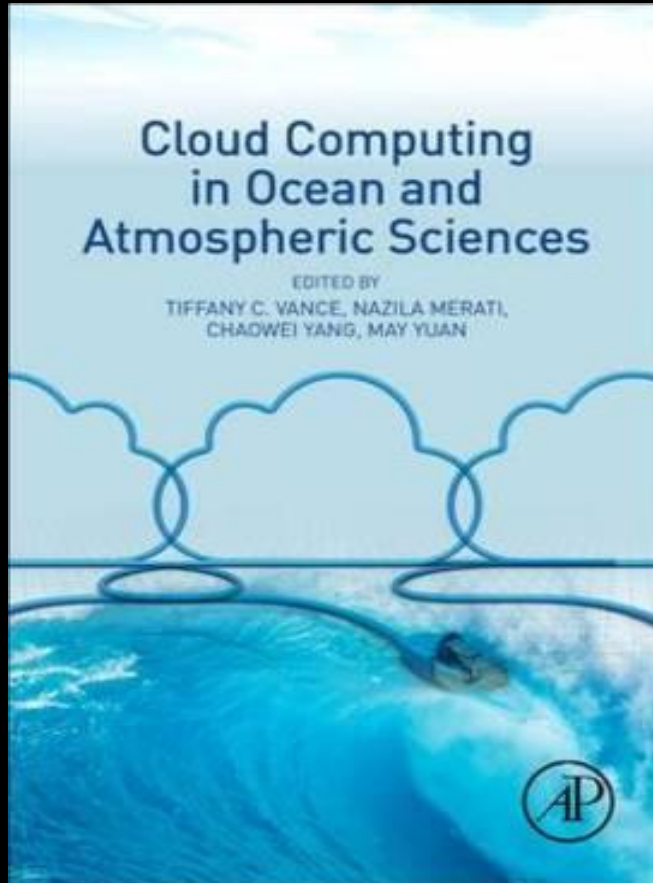
Simple and fast can make a difference ...

- It is estimated that climate scientists spend 50% – 80% of their time gathering and preparing data for study (Lohr, 2014)
- Our examples show a 3 orders-of-magnitude reduction in the time it takes to assemble data for many common tasks
- AR5 represents 6 years of research by a global community of over 2000 scientists
- If you assume that 10% of a 2000-hr work-year is spent directly dealing with data, and half that time is spent with data assembly, that amounts to 135 person-years of work
- Reducing that time by  $10^3$  yields an aggregate effort of less than 2 months ...



Lohr, S., August 17, 2014. For Big-Data Scientists, “Janitor Work” Is Key Hurdle to Insights. The New York Times. Technology Section [http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html?\\_r=2](http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html?_r=2).

# For more information ...



Schnase, J.L., 2016. Climate Analytics as a Service.  
In: Vance, T.C., Merati, N., Yang, C., Yuan, M. (Eds.),  
*Cloud Computing in Ocean and Atmospheric Sciences*.  
Academic Press, pp. 187-219. ISBN: 9780128031926,  
<http://dx.doi.org/10.1016/B978-0-12-803192-6.00011-6>

## CHAPTER 11

### Climate Analytics as a Service

J.L. Schnase

NASA Goddard Space Flight Center, Greenbelt, MD, USA

#### INTRODUCTION

Cloud technologies provide an unprecedented opportunity to expand the power and influence of computing in daily life. So far, those opportunities have unfolded in largely ad hoc ways, resulting in a creative chaos that at times can be confusing. Over the years, we have become comfortable with a classic von Neumann perspective on what constitutes a computer. We share mental models and patterns of thinking about how computers are built and how they behave—what in broad terms computing technologies can do, how they do it, and what they cannot do. But when it comes to cloud computing, those patterns have not been established—save one: the concept of *service*. We have developed a shared notion that cloud technologies in an essential way provide the basis for services. By definition, cloud capabilities reside there, not here—the action of helping is conveyed to the user: the user is served. Hence, terms such as Software-as-a-Service and Platform-as-a-Service have become common parlance in the world of cloud computing.

In our efforts to deal with the big data problems of climate science, we are trying to take a deeper dive into our understanding of cloud-computing services. To begin, we ask the fundamental question: What is it that needs to be served? Our answer is *analytics*. But analytics served in a particular way. For now at least, we believe it would be productive to focus on the basics—do simple things well and very fast. We need to garner the agile high-performance computing and storage resources of cloud computing to address climate science's big data problems in a new way—in a way that melds knowledge creation with data creation, curation, discovery, and workflow orchestration. This chapter is an effort to advance the cause.

Our story begins with the observation that big data challenges are generally approached in one of two ways. Sometimes they are viewed as a problem of large-scale data management, in which solutions are offered through an array of traditional storage and database theories and

*Cloud Computing in Ocean and Atmospheric Sciences*  
ISBN 978-0-12-803192-6  
<http://dx.doi.org/10.1016/B978-0-12-803192-6.00011-6>

Copyright © 2016 Elsevier Inc.  
All rights reserved.

187

Cloud Computing in Ocean and Atmospheric Sciences, First Edition, 2016, 187-219

# BDTF Questions

(1) What are the processes for planning for future (5-10 years) capabilities of your service? How and from whom do you gather input for this planning process and where does input typically come from? What new features have highest priority?

*Our Climate Informatics group is an applied R&D unit. When our technologies reach an appropriate TRL level, they are transferred to the NCCS for operational deployment. In our planning, we garner input from NASA climate scientists primarily through collaboration with the GMAO. We receive input from the extended community of climate scientists through our work with the Earth System Grid Federation and collaborating IPCC scientists. Goddard's Office of Patent Council and our Innovative Technology Partnership Office helps us interface with customers in the private sector. We work closely with NASA's Applied Sciences Program to identify other science application needs. Expanding work into the area of machine-learning-based analytics and knowledge discovery is an emerging priority.*

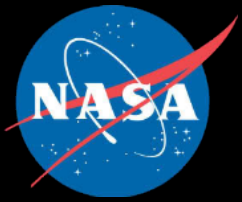
(2) What feature(s) of your service would you like to stop performing? How do you gather input for making such decisions and where does input typically come from? What is preventing you from stopping?

*Given the applied R&D nature of our group, these questions are largely non-applicable. We work closely with NASA's High-End Computing program, the academic research community, and the private sector to identify opportunities for new directions and the overall trajectory of our activities. We seek funding through competitive NASA ROSES venues, a process that sometimes helps us decide what work to discontinue.*

(3) What steps you are taking to make your data interoperable with allied data sets from other data sites in and out of NASA? How do you find allied data sets and what criteria make data sets candidates for enabling interoperability?

*We conform to the major data standards in play within the climate science community. Data held in the Reanalysis Ensemble Service are Climate Model Intercomparison Project 5/6 (CMIP5/6) compliant. Where standards are needed, we often contribute to the development of those standards. We collaborate with and provide leadership to the Earth System Grid Federation (ESGF) Compute Working Team (CWT) in their efforts to develop standards for interoperability among participating nodes of the ESGF. Our web services interfaces are designed to accommodate the CWT's emerging ESGF extensions to the Open Geospatial Consortium's (OGC) Web Processing Service (WPS) standard. Candidate datasets for inclusion in our work are identified through conversations with NASA HQ and interested groups of climate scientists. With the data sets we work with, a high level of interoperability is achieved by being integrated into our ecosystem of analytics-based climate data services.*





# Climate Analytics as a Service

*John Schnase*

*Office of Computational and Information Sciences and Technology*

*NASA Goddard Space Flight Center*

