1. Use Cases

5.0 Template

a. **Use Case Title**
b. Author/Company/Email - author of the use case
c. Actors/Stakeholders/Project URL and their roles and responsibilities - Who is interested in the project (perhaps providing guidance) and who will benefit from the project (recipients, users, beneficiaries, etc.)
d. Use Case Goal - What is the goal of the Earth science data analytics?:
1. To calibrate data
2. To validate data (quality) (i.e., validation or quality determination is the result; does not have to be via data intercomparison)
3. To perform course data reduction (e.g., subsetting, data mining)
4. To intercompare data (i.e., any data intercomparison; Could be used to better define validation/quality)
5. To derive new data product
6. To tease out information from data
7. To glean knowledge from data and information
8. To forecast/predict phenomena (i.e., Special kind of conclusion)
9. To derive conclusions (i.e., that do not easily fall into another type)
10. To derive analytics tools
11. To recover/rescue data
e. Use Case Description - This yields more details regarding how data analytics is utilized. Examples:
1. Datasets used, when applicable
2. Detailed analysis tasks - e.g., Combining products; Reducing data to just what is needed (subsetting); etc.
3. Results - The resulting product of the analytics. e.g., New product; New finding being sort; A decision based on some prediction
f. Current technical issues/requirements to take into account that may impact needed data analytics. These can include:
- Data Source (distributed/centralized)
- Volume (size)
- Velocity (e.g. real time)
- Variety - Bringing distributed heterogeneous data together (data formats, data types)
- Veracity (Robustness Issues) / Data Quality
- Visualization
- Compute(System), storage, networking
- Specialized Software
  - Current technologies required to perform research

- Is there a specific tool, implemented or commercial, that is being used
- Software needs should be identified to understand compatibility with available Data Analytics Tools and Techniques

- Current Data Analytics tools applied

g. Data Analytics Challenges (Gaps) - Identifying known data analytics challenges, roadblocks, areas needing attention to accomplish goals

h. Type of User - Taken from the Use Analysis Study, types of user performing use case. Earth science data user classes, who might be performing data analytics, include:

- Public - interested user of no or limited scientific skill
- Graduate student - person of moderate to high scientific skill at a university or college working towards an advanced degree
- Production Center - large organization that handles/processes vast quantities of data
- Science Team - group of scientists focused on a specific area of study or on a specific instrument type, can include calibration and validation scientists
- QA/Testing - developers or scientists using data to test software operation or to determine quality of a product, can include cal/val scientists
- Data Analyst - person using NASA data to perform a specific analysis
- Domain Scientist - person using data to do research and publish within a discipline, comes in with some expertise in using the data
- Interdisciplinary Scientist - person using high-level data products from multiple sources
- Operational User - Data analyst or technician using data for operational support (applications) and emergency response
- Assimilation Modeler - person or group that routinely obtains vast quantities of data for incorporation into models, can have operational needs
- Government Agency or Private Think Tank - a researcher or group of researchers using and analyzing data for decision or/and policy making
- Decision Support Systems - computer systems that routinely incorporate data into a system that a less knowledgeable user can use for deriving information

i. Dominant Data Analytics Skills Needed - Skills needed to perform use case analytics

j. Science Research Areas - NASA Earth science research areas (http://science.nasa.gov/earth-science/focus-areas/)

k. Societal Benefit Areas - GEOS (http://www.earthobservations.org/geoss.php) or NASA Applications (http://appliedsciences.nasa.gov)

l. Potential for and/or issues for generalizing this use case (e.g. for ref. architecture)

m. More Information and relevant URLs (e.g. who to contact or where to go for more information)

5.1 **Use Case: MERRA Analytics Services: Climate Analytics-as-a-Service**

a. **Title -** MERRA Analytics Services: Climate Analytics-as-a-Service
b. **Author/Company/Email** - Ethan McMahon, US EPA, mcmahon.ethan@epa.gov
c. **Actors/Stakeholders/Project URL and their roles and responsibilities -** Commercial parties that could use climate model data, such as air conditioning manufacturers, water vendors, farmers and investors. These parties could project the demand for their products and services. NASA Goddard is offering the data to whomever may want to use it.
d. **Use Case Goal** - The goal of Modern-Era Retrospective Analysis for Research and Applications (MERRA) is to enable a climate-quality analysis that places NASA's Earth Observing System (EOS) observations into a climate context. And provide a library of commonly used spatiotemporal operations (canonical ops) that can be composed to enable higher-order analyses
(#10 - Derive New Analytics Tools)
e. **Use Case Description** - Modern-Era Retrospective Analysis for Research and Applications (MERRA) is a MERRA is a historical reprocessing of climate data. It enables Climate Analytics-as-a-Service by combining iRODS data management, Cloudera MapReduce, and the Climate Data Services API to serve MERRA reanalysis products.
f. **Current technical issues/requirements to take into account that may impact needed data analytics:**
    - **Data Source (distributed/centralized)** - Distributed - Bringing data close to the places where it is processed will speed analyses and reduce input/output concerns.
    - Volume (size)
    - **Velocity (e.g. real time**) - No, because the data inputs are from the past.
    - Variety - Bringing distributed heterogeneous data together (data formats, data types)
    - Veracity (Robustness Issues) / Data Quality
    - **Visualization** - Outputs are displayed in map form on large (approximately 20 foot by 8 foot) screens.
    - **Compute(System), storage, networking** - Store the MERRA reanalysis data collection in HDFS to enable parallel, high-performance, storage-side data reductions; Manage storage-side <driver, mapper, reducer> code sets and realized objects for users; Deliver end-user and application capabilities through the NASA Climate Data Services API.
    - **Specialized Software** - Cloudera MapReduce
    - **Current Data Analytics tools applied** - Cloudera MapReduce
g. **Data Analytics Challenges (Gaps)** - None Identified
h. **Type of User** -
i. **Dominant Data Analytics Skills Needed** -
j. **Science Research Areas** -
k. **Societal Benefit Areas** -

l. Potential for and/or issues for generalizing this use case (e.g. for ref. architecture) Unclear. This is a massive analysis that very for other parties are likely to duplicate. However, this use case illustrates ways to keep the data close to the processing location in order to avoid I/O issues.

m. More Information and relevant URLs - Project managers: John Schnase: Senior Computer Scientist in NASA Goddard Space Flight Center's Computational and Information Sciences and Technology Office (CISTO); Dan Duffy: NASA Center for Climate Simulation (NCCS). http://gmao.gsfc.nasa.gov/merra/

5.2 **Use Case: MUSTANG QA: Ability to detect seismic instrumentation problems**

a. **Title -** MUSTANG QA: Ability to detect seismic instrumentation problems
b. **Author/Company/Email** - Robert Casey, Incorporated Research Institutions for Seismology (IRIS) rob@iris.washington.edu
c. **Actors/Stakeholders/Project URL and their roles and responsibilities -** IRIS Instrumentation Services; IRIS Global Seismic Network (GSN); IRIS Quality Assurance Working Group. Features are currently deployed at http://service.iris.edu/mustang in the form of web services.
d. **Use Case Goal** - To automatically gather various metrics on incoming and archival seismic data to assess and improve the quality of the stored seismic data as well as detect adverse conditions with the seismic instrumentation in the field.
(#3 - Assess data quality; #4 To perform coarse data preparation; #8 - Forecast/Predict)
e. **Use Case Description** - Examine one or more QA measurements to form high-level inferences as to data quality and instrument state of health. Certain metrics may highlight a number of different conditions so making use of multiple metrics tends to narrow the field as to diagnoses. Time trending and cumulative metric representations can point to other issues that may be otherwise hard to detect.
f. **Current technical issues/requirements to take into account that may impact needed data analytics:**
    - **Data Source (distributed/centralized)** - Data for analysis is centralized: an archival system with web service interface access
    - **Volume (size)** - 310TB in total, growing 50TB per year and accelerating
    - **Velocity (e.g. real time**) - Metrics are gathered at rate of the data access and metrics pipeline. About 800 GB of data scanned per day. Real time data accumulates about 100 GB per day, contributing to steady source data growth. Some data arrives as an entire package, many GB to TB at a time, for any point in time, resulting in bursty growth of the archive.
    - **Variety** - Seismic data being analyzed, along with mass position and pressure readings are currently returned in uniform fashion. Variety comes in the nature of the instrumentation, their response curves, whether triggered or not, and their colocation to other seismic stations. Native data format is SEED, an FDSN seismic standard for more than 25 years. Newly introduced are markups and representations of this data amenable to web services, among these a new XML metadata standard. Text tabular representations of data are also offered via web services.
    - **Veracity (Robustness Issues) / Data Quality** - Veracity is one of the key issues being addressed by MUSTANG, where the issues with robustness of the dataset are many. Bad data encoding, gappy data, missing data, incorrect metadata entries, bad time clock values, and overlapping datasets are all common problems encountered and must be accounted for and corrected. Real time data feeds generally have small flaws that are later corrected by a followup, post-processed

data set. The Quality Assurance Working Group for IRIS is charged with answering the question of what constitutes quality seismic data. MUSTANG is meant to generate figures that can help to detect a number of anomalies. If anomalies are few or not detected, then this could be one condition of flagging quality data. The data must cleanly and correctly provide a high definition measurement of motion and vibrations in the ground.

- **Visualization** - Visualization is a key feature that is desired but not mature in MUSTANG. This is partly because the main goal of MUSTANG is to produce the numbers and make them easily accessible and visualization is a key feature of seismic data quality analysis. As a result, many traditional approaches are currently in use such as gathering metrics in Excel for making plots. Visualization of waveforms and noise images is also very common, e.g. measuring diurnal patterns to noise. None of these are seamlessly integrated to MUSTANG, however, they are easily within reach.
- **Compute(System), storage, networking** - 10 VMs (ea. 8 core, 24 GB RAM) locally, (15 VMs, ea. 8 core, 24 GB RAM) remotely. NetApp filer with ~3 TB allocation, projecting need to ~5TB. NetApp filer replicated at remote site. Servers at each site mount common file system for code and configuration deployment, each has separate home FS. 10 Gbit load balancers and outbound network pipe between local and remote site. VPN connectivity (at much reduced bandwidth) and external connectivity for public HTTP access.
- **Specialized Software** - Job/Resource coordination using a master scheduler with listeners on remote machines. Analytics carried out using a statistical system like R, central data storage using an SQL database, job state persistence centralized on SQL database. All metrics transactions occur through HTTP web services (read/write). Also implementing a persistent queue (RabbitMQ) to control metrics capture flow from internal and external sources.
- **Current Data Analytics tools applied** - All original statistics in MUSTANG originate from processing in R. There are almost 50 different metrics being generated for each sensor per day. Analytical processes that occur after the fact take place in R, Matlab, Excel, and seismic viz tools like PQLX. Automated characterization of data quality or anomalies is only now being explored.

g. Data Analytics Challenges (Gaps) - Data sets are very large, so some complex data accesses can be slow when conducted via web services. Unsure of the veracity of metrics algorithms, so expert analysis of hits vs. misses must be carried out with each and every metric to check for proper characterization, correct numbers, proper tuning, and appropriate data selection. Visualization and exploration tools are lacking in MUSTANG (at this stage) to be really effective at analytics. Complete acquisition of data metrics can be difficult due to numerous errors and failures that can happen in the distributed compute pipeline. Detecting these gaps and visualizing data coverage are both immediate concerns being worked on.

h. Type of User - Network operator, scientist, graduate student, data analyst

i. Dominant Data Analytics Skills Needed -

j. Science Research Areas -

k. Societal Benefit Areas -

l. Potential for and/or issues for generalizing this use case (e.g. for ref. architecture) - Generating metrics to ensure data quality and instrumentation health can probably be viewed as applicable to many forms of instrumentation data gathering efforts.

m. More Information (URLs) - (1) http://ds.iris.edu/ds/nodes/dmc/quality-assurance/ , (2) http://www.adv-geosci.net/40/31/2015/ , (3) http://service.iris.edu

5.3 **Use Case:** Inter-calibrations among datasets

a. **Title -** Inter-calibrations among datasets
b. **Author/Company/Email** - Tiffany Mathews, SSAI (@ NASA ASDC) tiffany.j.mathews@nasa.gov
c. **Actors/Stakeholders/Project URL and their roles and responsibilities -** Committee on Earth Observation Satellites (CEOS) Working Group on Calibration and Validation (WGCV); World Meteorological Organization (WMO); Global Space based Intercalibration System (GSICS) community; Instrument and Science Teams
d. **Use Case Goal** - To improve target instrument calibration and remove temporal calibration trend by being able to quickly compare co-located measurements with matched viewing geometries from different sensors on separate spacecraft. All of which re-use existing science algorithms and information technology software to increase the interconnectedness of existing distributed information systems and the quality of datasets.
(#1 - Calibration; #2 - Validation; #5 - Inter-comparison)
e. **Use Case Description** - Instrument Calibration, offering a solution to instrument teams responsible for calibration and validation of target instrument data. Specifically, best inter-calibration practices for sensors on Geosynchronous and Low Earth Orbiting satellites.
f. **Current technical issues/requirements to take into account that may impact needed data analytics:**
    - **Data Source (distributed/centralized)** -
    - **Volume (size)** -
    - **Velocity (e.g. real time)** -
    - **Variety** -
    - **Veracity (Robustness Issues) / Data Quality** -
    - **Visualization** -
    - **Compute(System), storage, networking** - Remote Data Servers executing algorithms such as convolution over sensor point spread functions and spectral response functions.
    - **Specialized Software** -
    - **Current Data Analytics tools applied** - MIIC II Framework. Uses: Extensible Markup Language (XML) for inter-calibration plan (the event specifications are inserted into this); Open-source Project for Network Access Protocol (OPeNDAP) (event inter-calibration algorithms are executed on remote servers using OPeNDAP server-side functions prior to delivery of the data to the instrument teams for further analysis)
g. **Data Analytics Challenges (Gaps)** -
h. **Type of User** -
i. **Dominant Data Analytics Skills Needed** -
j. **Science Research Areas** -
k. **Societal Benefit Areas** -

l. Potential for and/or issues for generalizing this use case (e.g. for ref. architecture) -

m. More Information (URLs)-

https://earthdata.nasa.gov/our-community/community-data-system-programs/access-projects/multi-instrument-inter-calibration-miic-framework

http://clarreo.larc.nasa.gov/2014-01STM/Wednesday/Currey_MIIC_Framework_CLARREO_STM_123013.pdf

5.4 **Use Case:** Inter-comparisons between multiple model or data products

a. **Title -** Inter-comparisons between multiple model or data products
b. **Author/Company/Email** - Tiffany Mathews, SSAI (@ NASA ASDC) tiffany.j.mathews@nasa.gov
c. **Actors/Stakeholders/Project URL and their roles and responsibilities -** Science and Instrument Teams as well as Researchers (Academia)
d. **Use Case Goal** - To be able to obtain high-resolution inter-comparisons between two or more model or data products in order to get more accurate measurements than what can be received from observing the global monthly means such as enabling one to make a point by point comparisons of specific locations.
(#5 - Intercomparison)
e. **Use Case Description** - Instead of only being provided the global monthly means (averages), if users had access to two more broken down data of two or more model or data products and high-resolution intercalibration tools, users would be better able to make a point by point comparisons of specific locations and better identify phenomena that contribute to certain patterns.
f. **Current technical issues/requirements to take into account that may impact needed data analytics:**
    - **Data Source (distributed/centralized)** - Data is within the same product or model
    - **Volume (size)** - Depending on the size of the volume (to break down a global monthly means to a more specific point in time) the volume could become overwhelming if the appropriate tools are unavailable to identify specific times and places.
    - **Velocity (e.g. real time)** -
    - **Variety** - Heterogeneous (as it comes from the same product but is more broken down) or from similar products measuring the same place and point in time.
    - **Veracity (Robustness Issues) / Data Quality** -
    - **Visualization** - Visualization used to identify phenomena (e.g. the global monthly mean might not reflect two equally different and extreme phenomena).
    - **Compute(System), storage, networking** -
    - **Specialized Software** -
    - **Current Data Analytics tools applied** -
g. **Data Analytics Challenges (Gaps)** -
h. **Type of User** - Public user with significant scientific skill, Graduate students, Science Teams, Data Analyst, Domain Scientist - person using data to do research and publish within a discipline, comes in with some expertise in using the data, Interdisciplinary Scientist, Operational User, Assimilation Modelers, Decision Support Systems
i. **Dominant Data Analytics Skills Needed** -
j. **Science Research Areas** -
k. **Societal Benefit Areas** -
l. **Potential for and/or issues for generalizing this use case (e.g. for ref. architecture)** -
m. **More Information (URLs)-** http://nacp.ornl.gov/MsTMIP_products.shtml

http://power.larc.nasa.gov/solar/publications/solar2006_A218.pdf

5.5 **Use Case:** Sampling Total Precipitable Water Vapor using AIRS and MERRA

a. **Title -** Sampling Total Precipitable Water Vapor using Atmospheric Infrared Sounder (AIRS) and MERRA
b. **Author/Company/Email** - Steve Kempler, NASA/GSFC, Steven.J.Kempler@nasa.gov
c. **Actors/Stakeholders/Project URL and their roles and responsibilities -** GMAO; Reseacrh scientists
d. **Use Case Goal** - AIRS and MERRA data inter-comparison **(#2 - Validation; #5 - Intercomparison)**
e. **Use Case Description** - MERRA data is being used to cross-validate with AIRS Total water vapor. Being of different characteristics, datasets need to be manipulated to enable intercomparison. The result of the analytics will be datasets that can be more homogeneously compared. Differences can be further analyzed to assess deficiencies in both the observations and the reanalyses.
f. **Current technical issues/requirements to take into account that may impact needed data analytics:**
    - **Data Source (distributed/centralized)** - Data is co-located in the same facility
    - **Volume (size)** -
    - **Velocity (e.g. real time)** -
    - **Variety** - Heterogeneous datasets are made to be more homogeneous so they can be inter-compared: Grid, resolution, etc.
    - **Veracity (Robustness Issues) / Data Quality** -
    - **Visualization** - Visualization heavily used to spot measurement differences and signatures
    - **Compute(System), storage, networking** -
    - **Specialized Software** - Creating common sampling (time series), gridding. Accounting for temporal and instrument effects. Applying MERRA Sampled like AIRS with Quality control
    - **Current Data Analytics tools applied** -
g. **Data Analytics Challenges (Gaps)** -
h. **Type of User** - Science Teams
i. **Dominant Data Analytics Skills Needed** -
j. **Science Research Areas** -
k. **Societal Benefit Areas** -
l. **Potential for and/or issues for generalizing this use case (e.g. for ref. architecture)** - Same/similar issues arise when inter-comparing any heterogeneous datasets
m. **More Information (URLs) - NASA GES DISC, Thomas Hearty (thomas.j.hearty@nasa.gov), et al;**

5.6 **Use Case:** Using Earth Observations to Understand and Predict Infectious Diseases

a. **Title -** Using Earth Observations to Understand and Predict Infectious Diseases
b. **Author/Company/Email** - Steve Kempler, NASA/GSFC, Steven.J.Kempler@nasa.gov
c. **Actors/Stakeholders/Project URL and their roles and responsibilities -** Public Health Decision Makers
d. **Use Case Goal** - Identify meteorological parameters associated with disease outbreaks; Disease forecast using meteorological parameters.
(#8 - Forecast/Predict; #9 - Derive conclusions)
e. **Use Case Description** - Meteorological data, TRMM precipitation and GLDAS near surface temperature and near surface specific humidity is modeled using a variety of methods to ultimately predict influenza outbreak. Uses the Disease Surveillance database for inter-comparisons.
f. **Current technical issues/requirements to take into account that may impact needed data analytics:**
    - **Data Source (distributed/centralized)** - Distributed, but brought together
    - **Volume (size)** - Storage capacity problem.
    - **Velocity (e.g. real time)** -
    - **Variety** - Heterogeneous datasets are ingested into models that can accommodate the dataset
    - **Veracity (Robustness Issues) / Data Quality** -
    - **Visualization** - Visualizations are used to verify seasonal patterns of the aggregated meteorological parameters, for data exploration (i.e. scatter plot between disease outbreak and meteorological parameters to assess initial or existence of relationship), to illustrate disease prediction and visually compare with observed data.
    - **Compute(System), storage, networking** -
    - **Specialized Software** - In-house database of remote sensing products where a user can retrieve a time series data for any administrative region (country, province and district) or any user-specified rectangular area. Options to average data into weekly and monthly are also available. Mathematical and statistical modeling (i.e. regression, neural network) are performed using Matlab and R software.
    - **Current Data Analytics tools applied** - Regression Modeling, Machine Training and Prediction, Neural Network
g. **Data Analytics Challenges (Gaps)** - Meteorological Data and Processing
    - Changes in or heterogeneity of: location, formats, algorithm, availability (data continuity)
    - Data products validation
    - Uncovering patterns & modeling
    - Choice of mathematical and statistical models
    - Each model has assumptions such that results and prediction may need to be appropriately interpreted
    - Parameter constraints and prediction validation

h. Type of User - Assimilation Modelers

i. Dominant Data Analytics Skills Needed -

j. Science Research Areas -

k. Societal Benefit Areas -

l. Potential for and/or issues for generalizing this use case (e.g. for ref. architecture) - Tools used can be applied elsewhere. Need to determine where

m. More Information (URLs) - NASA Global Change Data Center (GCDC), Radina P. Soebiyanto (radina.p.soebiyanto@nasa.gov),Richard Kiang (richard.kiang@nasa.gov)

5.7 **Use Case: CREATE-IP -** Collaborative REAnalysis Technical Environment - Intercomparison Project

a. **Title - CREATE-IP -** Collaborative REAnalysis Technical Environment - Intercomparison Project
b. **Author/Company/Email** - Gerald L. Potter, Laura Carriere laura.carriere@nasa.gov
c. **Actors/Stakeholders/Project URL and their roles and responsibilities -** Five major reanalysis projects - NASA's MERRA; European Centre for Medium-Range Weather Forecasts (ECMWF) ERA-Interim; NOAA/NCEP's Climate Forecast System Reanalysis (CFSR); NOAA/ESRL's 20CR; Japan Meteorolocical Agency (JMA) JRA-25 and JRA-55.
d. **Use Case Goal** - Reanalysis scientists are interested in reproducing the success of Coupled Model Intercomparison Project Phase 5 (CMIP5) by studying reanalysis differences and similarities to improve reanalysis techniques. Reanalysis data also allows interdisciplinary scientist to compare their datasets with 30 or more years of gridded climate data.
(#5 - Intercomparison)
e. **Use Case Description** - Climate reanalysis. Data sets - NASA's MERRA, ECMWF's ERA-Interim, NOAA/NCEP's CFSR, NOAA/ESRL's 20CR, and JMA's JRA-25 and JRA-55. New datasets to be added as they become available. Resulting Product - Improved climate reanalysis code.
f. **Current technical issues/requirements to take into account that may impact needed data analytics:**
    - **Data Source (distributed/centralized)** - Distributed - 2D and 3D global gridded model output, multiple variables located at their respective data centers.
    - **Volume (size)** - Currently 0.5 TB. Growth to 500 TB to 1 PB expected.
    - **Velocity (e.g. real time**) - Data is generated on supercomputers. Most of the data is currently available or will be made available as soon as the model runs have completed.
    - **Variety** - Native formats from centers differ (netcdf, grib, variable names, etc). CREATE-IP will prepare the data in ESGF standard format to facilitate comparison. Netcdf (some data is originally grib and hdf).
    - **Veracity (Robustness Issues) / Data Quality** - Errors can be found in the model output resulting in a reprocessing. Data quality dependent on input observations (which can improve with new satellite instrumentation) and climate model.
    - **Visualization** - CREATE-IP will provide multiple visualization services, including a Web Map Service (WMS) tool, UV-CDAT, and ArcGIS.
    - **Compute(System), storage, networking** - Each reanalysis project has access to their own computer. Each reanalysis project has local access to their own data but would need to download and format the other reanalysis datasets. Issues associated with moving data between the US, Japan, and Europe.

- **Specialized Software** - Based on climate model code, e.g. GEOS-5. Each reanalysis center has their own code but some are based on the same internal physics code.
- **Current Data Analytics tools applied** - Comparison of anomalies, i.e. departures from the climatology (one year's average of the model's 30+ years of data). Analysis of "innovations", the correction applied to the model data to bring the model back to the observations after a set number of model timesteps.

g. Data Analytics Challenges (Gaps) - Data volumes, particularly for hourly data. Download times. Format differences. Differences in the definition of the variables by each center.

h. Type of User - Domain Scientists, Interdisciplinary Scientists (Note, above description is related to Domain Scientist usage.)

i. Dominant Data Analytics Skills Needed -

j. Science Research Areas -

k. Societal Benefit Areas -

l. Potential for and/or issues for generalizing this use case (e.g. for ref. architecture) - Building an infrastructure to allow scientists to run similar workflows on the multiple variables from multiple analyses. This could be generalized to other ESGF projects as the data formatting is the basis of CREATE-IP

m. More Information (URLs) - http://esgf.nccs.nasa.gov Precursor to CREATE-IP: **https://earthsystemcog.org/projects/ana4MIPs**

5.8 **Use Case:** The GSSTF Project (MEaSUREs-2006)

a. **Title -** The GSSTF Project (MEaSUREs-2006)
b. **Author/Company/Email** - Chung-Lin Shie, UMBC/JCET, NASA/GSFC Chung-Lin.Shie-1@nasa.gov
c. **Actors/Stakeholders/Project URL and their roles and responsibilities -** Data producers/providers; Research scientists; Community users
d. **Use Case Goal** - Produce long-term (21.5 yrs) global air-sea surface turbulent fluxes gridded datasets ("hybrid" and L4) using the SSM/I multple-sensor (L2 or L3) retrievals and the NCEP reananlysis gridded products (L3 or L4).
(#6 - To tease out information from data)
e. **Use Case Description** - Gobal Water and Energy Cycle.
Input (applied) datasets: a) SSM/I multi-sensor (F08, F10, F11, F13, F14 and F15) brightness temperature (TB) and total precipitable water (W), surface air humidity at 10 m (Q), surface wind speed (U); b) NCEP reanalysis of sea surface/skin temperature (SST/SKT), air temperature at 2m (Tair), and sea level pressure (SLP); c) Cross-Calibrated Multi-Platform (CCMP) ocean surface wind vectors (mainly based on SSM/I multi-sensors). Output (resultant) datasets: Combined (multi sensors) and "hybrid" (satellite and model reanalysis) products consisting of new physical parameters, i.e., the air-sea turbulent fluxes: latent heat flux (LHF), sensible heat flux (SHF) and wind stress (WST).
f. **Current technical issues/requirements to take into account that may impact needed data analytics:**
    - **Data Source (distributed/centralized)** - Involve (input) multiple datasets of multiple variables from multiple data source (e.g., multi satellites/sensors and model renanlysis)
    - **Volume (size)** - Currently ~1.0 TB. May grow if more funding granted. The current data has covered from July 1987-Dec 2008, i.e., 21.5 years. Assume that the PI (data provider Chung-Lin Shie) has successful won a 3-yr grant (say 2016-2018), and thus would be able to resume and extend the data producton for another 10-yr (Jan 2009-Dec 2018). Then, a 0.5 TB data volume would be grown by Dec 2018
    - **Velocity (e.g. real time)** -
    - **Variety** - Using heterogeneous data (mainly in binary format) and creating a new data product (physical quantities) of HDF-EOS5 format.
    - **Veracity (Robustness Issues) / Data Quality** - The (resultant) data quality depends not only on the model (satellite data retrieval algorithms and flux model) quality but also on the (input) data quality.
    - **Visualization** - Visualization may mainly serve research purpose so far.
    - **Compute(System), storage, networking** - Datasets produced by scientists using local workstations, then transferred to GES DISC and distributed by GES DISC.
    - **Specialized Software** - Algorithms or/and models for retrieving surface air humidity and computing surface turbulent fluxes, respectively, are required.

- **Current Data Analytics tools applied** - Data format convention -- from binary to HDF-EOS5.

g. Data Analytics Challenges (Gaps) - Producing new sets of massive data products containing valuable physical quantities based on/using several sets of existing massive data (satellite observations or model reanalysis)

h. Type of User - Research Scientists

i. Dominant Data Analytics Skills Needed -

j. Science Research Areas - Air-sea Interactions. Energy and Water Cycle. P-E (Fresh Water). Monsoons. etc.

k. Societal Benefit Areas - Climate Changes.

l. Potential for and/or issues for generalizing this use case (e.g. for ref. architecture) - May not be practical to build an infrastructure that may not become popular to be followed by other scientists, but some potential issues may deserve our attention and further discussions such as 1) The Rice Cooker Theory, 2) Producing new (more) "big" datasets based on existing "big" datasets, how do we leverage and optimize such kind of "post-processing" productions, especially when they are inevitable for the Earth science research purpose?

m. More Information (URLs) -
http://disc.sci.gsfc.nasa.gov/measures/documentation/Science_of_the_data.GSSTF3.pdf

5.9 **Use Case:** Science- and Event-based Advanced Data Service Framework at GES DISC

a. **Title -** Science- and Event-based Advanced Data Service Framework at GES DISC
b. **Author/Company/Email** - Chung-Lin Shie, UMBC/JCET, NASA/GSFC Chung-Lin.Shie-1@nasa.gov
c. **Actors/Stakeholders/Project URL and their roles and responsibilities -** Data producers/providers; Research scientists; Community users
d. **Use Case Goal** - Efficiently provide the users with a sophisticated integrated (i.e., knowledge-based) data package ("bundle") via user-friendly selecting a system-predefined science- or event-based topic (e.g., hurricane, volcano, etc.) among the currently in-developing knowledge database of the framework.
(#5 - Intercomparison; #10 - Derive New Analytics Tools)
e. **Use Case Description** - A prototype portal related to a specific topic of Hurricane Sandy (Oct 22-31 2012) that has been developed (currently as a "recipe") at GES DISC is used here to describe the “bundle” data service. As Hurricane Sandy being selected as a user targeted case, a system-prearranged table consisting of related data variables (i.e., precipitation, winds, sea surface temperature, sea level pressure, air temperature, relative humidity, aerosols, soil moisture and surface runoff, trace gases) linked to the respective data products with fine temporal and spatial resolution of various in-house sources is provided. All the listed data variables that should be readily applied by users for studies such as hurricane track, hurricane intensity, disaster analysis, and its impacts etc. can then be readily downloaded through the data search engine, Mirador. The powerful visualization tools Giovanni (online; built in-house) is accessible for users to acquire quick and informative views of their interested data variables/products of Level 3 (gridded) in the Giovanni database. For Level 2 data (swath) and certain Giovanni-unavailable Level 3 data, the system provides a link to data recipes that give a step-by-step how-to guide to read or visualize the data using offline tools, such as Panoply, GrADS, or IDL, etc.
f. **Current technical issues/requirements to take into account that may impact needed data analytics:**
    - **Data Source (distributed/centralized)** - Involve (input) multiple datasets of multiple variables from multiple data source (e.g., multi satellites/sensors and model renanlysis)
    - **Volume (size)** - The "bundle" data service here is not about to focus mainly on the data volume growth (e.g., a constantly increased data volume), but rather to provide users a knowledge-based service to efficiently (quicker and more accurately) acquire (extract) the targeted data parameters/datasets that the users are interested with a considerably reduced volume size out of the various and originally massive data products (of big volume size).
    - **Velocity (e.g. real time)** -
    - **Variety** - "Bundle" together the diverse (yet knowledge-based) data parameters of interests from various data products.

- **Veracity (Robustness Issues) / Data Quality** - This "knowelege-based" (can also be called as "value-added") service has ensured users to acquire the "right" (adequate and needed) data products.
- **Visualization** - The visualization tool, e.g., the powerful Giovanni, is accessible for users to acquire quick and informative views of their interested data variables/products.
- **Compute(System), storage, networking** - The "Bundled" and the original data variables/datasets are distributed at GES DISC.
- **Specialized Software** - Currently, this data service framework is developed on top of existing data services at GES DISC, such as Mirador (search engine), Giovanni (visualization), OPeNDAP, and data recipes. It also involves other popular data tools, such as Panoply, GrADS, IDL, etc. A "Virtual Collection" concept for "Data Bundles" would involve more computer-driven, machine-learning and software-development
- **Current Data Analytics tools applied** - Data format conversion -- from binary to HDF-EOS5.

g. Data Analytics Challenges (Gaps) - To develop a Virtual Collection of Data Bundles.

h. Type of User - Research Scientists, Applications Scientists, General Publics, and Students

i. Dominant Data Analytics Skills Needed -

j. Science Research Areas - Hurricanes. Volcanos. etc

j. Societal Benefit Areas - Disasters

k. Potential for and/or issues for generalizing this use case (e.g. for ref. architecture)

- This value-added service of "Data Bundles" could serve as a good model for other NASA DAACs or non-NASA data centers to consider if they have not developed a similar one of their own. Or, this prototype service may develop into a cross-DAAC collaboration or implementation.

m. More Information (URLs) -

http://disc.sci.gsfc.nasa.gov/recipes/?q=recipes/How-to-Obtain-Data-for-Conducting-Hurricane-Case-Study

5.10 **Use Case:** Risk analysis for environmental issues

a. **Title -** Risk analysis for environmental issues
b. **Author/Company/Email** - Joan Aron, joanaron@ymail.com
c. **Actors/Stakeholders/Project URL and their roles and responsibilities -** Climate and Air Quality studies/decision makers
d. **Use Case Goal** - Trends (comparisons) across time and space; Other end user needs may be different; Example: Emergency response systems need the latest data but not historical data.
   (#8 - Prediction)
e. **Use Case Description** -
f. **Current technical issues/requirements to take into account that may impact needed data analytics:**
   - **Data Source (distributed/centralized)** - Ability to link datasets from different agencies, sectors and/or disciplines
   - **Volume (size)** -
   - **Velocity (e.g. real time)** -
   - **Variety** -
   - **Veracity (Robustness Issues) / Data Quality** -
   - **Visualization** -
   - **Compute(System), storage, networking** -
   - **Specialized Software** -
   - **Current Data Analytics tools applied** -
g. **Data Analytics Challenges (Gaps)** - Robust, machine-readable metadata to determine which model outputs are suitable as inputs for companion models, and tools for converting units, regridding, etc.
h. **Type of User** - Research Scientists
i. **Dominant Data Analytics Skills Needed** -
j. **Science Research Areas** -
k. **Societal Benefit Areas** -
l. **Potential for and/or issues for generalizing this use case (e.g. for ref. architecture)** -
m. **More Information (URLs) -**

5.11 **Use Case:** Aerosol Characterization

a. **Title -** Aerosol Characterization
b. **Author/Company/Email** - Steve Kempler, NASA/GSFC, Steven.J.Kempler@nasa.gov
c. **Actors/Stakeholders/Project URL and their roles and responsibilities -** Atmospheric Science Researchers
d. **Use Case Goal** - Discover/uncover global and regional aerosol attributes utilizing data from multiple Earth observations, satellite and suborbital, combined with climate and air quality models.
(#5 – Inter-comparison; #9 - Derive conclusions)
e. **Use Case Description** - To understand the anthropogenic and natural aerosol contributions to global climate forcing and regional air quality; Validate global and regional aerosol transport models; Applying analysis techniques that are often unique to the research.
f. **Current technical issues/requirements to take into account that may impact needed data analytics:**
• **Data Source (distributed/centralized)** - Ability to link datasets from different instruments, agencies, sectors and/or disciplines
• **Volume (size)** - working with huge volumes of data (~700 GB/day for 15+ years)
• **Velocity (e.g. real time)** -
• **Variety** - MODIS, MISR, CALIPSO, OMI, AERONET, aircraft field campaign and surface-station data, and model simulations
• **Veracity (Robustness Issues) / Data Quality** - Part of analysis
• **Visualization** – Required, often non-standard
• **Compute(System), storage, networking** -
• **Specialized Software** – developed as needed
• **Current Data Analytics tools applied** -
g. **Data Analytics Challenges (Gaps)** - Pattern recognition to identify wildfire smoke, volcanic ash, and dust plumes with less that 10% false positives and false negatives, etc.
h. **Type of User** - Research Scientists
i. **Dominant Data Analytics Skills Needed** -
j. **Science Research Areas** -
k. **Societal Benefit Areas** -
l. **Potential for and/or issues for generalizing this use case (e.g. for ref. architecture)** - Applicable to dust, smoke, volcanic ash, and pollution studies
m. **More Information (URLs) -** Ralph Kahn, ralph.a.kahn@nasa.gov

5.12 **Use Case:** Creating One Great Precipitation Data Set From Many Good Ones

a. **Title -** Creating One Great Precipitation Data Set From Many Good Ones
b. **Author/Company/Email** - Steve Kempler, NASA/GSFC, Steven.J.Kempler@nasa.gov
c. **Actors/Stakeholders/Project URL and their roles and responsibilities -** Precipitation researchers, applications, near real-time users
d. **Use Case Goal** - Create a merged global precipitation dataset utilizing several precipitation measuring sources
(#6 - To tease out information from data)
e. **Use Case Description** - Various precipitation measurements are combined to create a single Integrated Multi-satellitE Retrievals for GPM (IMERG) product (0.5 hour, 0.1° lat/lon, currently 60°N – 60°S)
f. **Current technical issues/requirements to take into account that may impact needed data analytics:**
- **Data Source (distributed/centralized)** - Distributed; highly heterogeneous
- **Volume (size)** -
- **Velocity (e.g. real time**) - Near real time resulting product
- **Variety** - Several satellite (Low-Earth passive and active microwave; Geostationary Infrared (frequent images but low quality) and precipitation gauges
- **Veracity (Robustness Issues) / Data Quality** - Microwave: Infrequent overpasses but good quality; Infrared: frequent images but low quality
- **Visualization** - Very helpful
- **Compute(System), storage, networking** -
- **Specialized Software** - Intercalibration process software; Morphing
- **Current Data Analytics tools applied -** Kalman filter-based technique to weight forward and backward propagated estimates

g. **Data Analytics Challenges (Gaps)** - Produce a combined product using the strengths and mitigating the weaknesses of individual data sets; Different sensors provide different precipitation estimates so we must intercalibrate all estimates to the best standard (GPM combined GMI/DPR); Fill data gaps
h. **Type of User** - Public; Research Scientists; Applications; Decision Makers
i. **Dominant Data Analytics Skills Needed** -
j. **Science Research Areas** -
k. **Societal Benefit Areas** -
l. **Potential for and/or issues for generalizing this use case (e.g. for ref. architecture)** -
m. **More Information (URLs) -** Dave Bolvin, david.t.bolvin@nasa.gov, George Huffman, george.j.huffman@nasa.gov

5.13 **Use Case:** Reconstructing Sea Ice Extent from Early Nimbus Satellites

a. **Title -** Reconstructing Sea Ice Extent from Early Nimbus Satellites
b. **Author/Company/Email** - Steve Kempler, NASA/GSFC, Steven.J.Kempler@nasa.gov
c. **Actors/Stakeholders/Project URL and their roles and responsibilities -** Snow/ice researchers, climate change researchers
d. **Use Case Goal** - To recover 1960's vintage Nimbus data into a usable form; To get this data to earth science researchers in a modern format (NetCDF); To potentially extend the satellite sea ice 1979-2014 records by as much as 16 years.
(#1 - Calibrate data; #4 To perform coarse data preparation)
e. **Use Case Description** - Digitize images; Create metadata for each image; Flag bad data; Merge with ephemeris data; Merge into time composite; Calibrate
f. **f. Current technical issues/requirements to take into account that may impact needed data analytics:**
   - **Data Source (distributed/centralized)** - Tapes
   - **Volume (size)** - 250,000 images; over 1,000,000 frames
   - **Velocity (e.g. real time)** -
   - **Variety** -
   - **Veracity (Robustness Issues) / Data Quality** - Some of the information is not readable
   - **Visualization** -
   - **Compute(System), storage, networking** -
   - **Specialized Software** -
   - **Current Data Analytics tools applied -**

g. **Data Analytics Challenges (Gaps)** - Not possible to automate because some of the information is not readable
h. **Type of User** - Research Scientists; Earth science modelers
i. **Dominant Data Analytics Skills Needed** -
j. **Science Research Areas** -
k. **Societal Benefit Areas** -
l. **Potential for and/or issues for generalizing this use case (e.g. for ref. architecture)** -
m. **More Information (URLs) -** David Gallaher, david.gallaher@nsidc.org

5.14 **Use Case:** DOE-BER AmeriFlux and FLUXNET Networks (borrowed, with permission, from NIST Big Data Use Case Submissions [http://bigdatawg.nist.gov/usecases.php])

a. **Title -** DOE-BER AmeriFlux and FLUXNET Networks
b. **Author/Company/Email** - Steve Kempler, NASA/GSFC, Steven.J.Kempler@nasa.gov, adapted from Deb Agarwal, Lawrence Berkeley Lab. daagarwal@lbl.gov
c. **Actors/Stakeholders/Project URL and their roles and responsibilities -** AmeriFlux scientists, Data Management Team, ICOS, DOE TES, USDA, NSF, and Climate modelers
d. **Use Case Goal** - AmeriFlux Network and FLUXNET measurements provide the crucial linkage between organisms, ecosystems, and process-scale studies at climate-relevant scales of landscapes, regions, and continents, which can be incorporated into biogeochemical and climate models. Results from individual flux sites provide the foundation for a growing body of synthesis and modeling analyses.
(#6 - Tease out information; #9 - Derive conclusions)
e. **Use Case Description** - AmeriFlux network observations enable scaling of trace gas fluxes (CO2, water vapor) across a broad spectrum of times (hours, days, seasons, years, and decades) and space. Moreover, AmeriFlux and FLUXNET datasets provide the crucial linkages among organisms, ecosystems, and process-scale studies—at climate-relevant scales of landscapes, regions, and continents—for incorporation into biogeochemical and climate models
f. **Current technical issues/requirements to take into account that may impact needed data analytics:**
    - **Data Source (distributed/centralized)** - ~150 towers in AmeriFlux and over 500 towers distributed globally collecting flux measurements
    - **Volume (size)** -
    - **Velocity (e.g. real time)** -
    - **Variety** - The flux data is relatively uniform, however, the biological, disturbance, and other ancillary data needed to process and to interpret the data is extensive and varies widely. Merging this data with the flux data is challenging in today's systems
    - **Veracity (Robustness Issues) / Data Quality** - Each site has unique measurement and data processing techniques. The network brings this data together and performs a common processing, gap-filling, and quality assessment. Thousands of users
    - **Visualization** - Graphs and 3D surfaces are used to visualize the data.
    - **Compute(System), storage, networking** - NSERC, ESNet
    - **Specialized Software** - EddyPro, Custom analysis software, R, python, neural networks, Matlab
    - **Current Data Analytics tools applied -** Data mining, data quality assessment, cross-correlation across datasets, data assimilation, data interpolation, statistics, quality assessment, data fusion, etc
g. **Data Analytics Challenges (Gaps)** - Translation across diverse datasets that cross domains and scales.

h. Type of User -
i. Dominant Data Analytics Skills Needed -
j. Science Research Areas -
k. Societal Benefit Areas -
l. Potential for and/or issues for generalizing this use case (e.g. for ref. architecture) -
m. More Information (URLs) - Ameriflux.lbl.gov, www.fluxdata.org

5.15 **Use Case:** DOE-BER Subsurface Biogeochemistry Scientific Focus Area (borrowed, with permission, from NIST Big Data Use Case Submissions [http://bigdatawg.nist.gov/usecases.php])

a. **Title -** DOE-BER Subsurface Biogeochemistry Scientific Focus Area
b. **Author/Company/Email** - Steve Kempler, NASA/GSFC, Steven.J.Kempler@nasa.gov**,** adapted from Deb Agarwal, Lawrence Berkeley Lab. daagarwal@lbl.gov
c. **Actors/Stakeholders/Project URL and their roles and responsibilities -** LBNL Sustainable Systems SFA 2.0, Subsurface Scientists, Hydrologists, Geophysicists, Genomics Experts, JGI, Climate scientists, and DOE SBR
d. **Use Case Goal** - The Sustainable Systems Scientific Focus Area 2.0 Science Plan ("SFA 2.0") has been developed to advance predictive understanding of complex and multiscale terrestrial environments relevant to the DOE mission through specifically considering the scientific gaps defined above.
(#8 - Forecast/Predict)
e. **Use Case Description** - Development of a **G**enome-**E**nabled **Wa**tershed **S**imulation **C**apability (**GEWaSC)** that will provide a predictive framework for understanding how genomic information stored in a subsurface microbiome affects biogeochemical watershed functioning, how watershed-scale processes affect microbial functioning, and how these interactions co-evolve. While modeling capabilities developed by our team and others in the community have represented processes occurring over an impressive range of scales (ranging from a single bacterial cell to that of a contaminant plume), to date little effort has been devoted to developing a framework for systematically connecting scales, as is needed to identify key controls and to simulate important feedbacks. A simulation framework that formally scales from genomes to watersheds is the primary focus of this GEWaSC deliverable.
f. **Current technical issues/requirements to take into account that may impact needed data analytics:**
    - **Data Source (distributed/centralized)** - Terabase-scale sequencing data from JGI, subsurface and surface hydrological and biogeochemical data from a variety of sensors (including dense geophysical datasets) experimental data from field and lab analysis
    - **Volume (size)** -
    - **Velocity (e.g. real time)** -
    - **Variety** - Data crosses all scales from genomics of the microbes in the soil to watershed hydro-biogeochemistry. The SFA requires the synthesis of diverse and disparate field, laboratory, and simulation datasets across different semantic, spatial, and temporal scales through GEWaSC. Such datasets will be generated by the different research areas and include simulation data, field data (hydrological, geochemical, geophysical), 'omics data, and data from laboratory experiments.
    - **Veracity (Robustness Issues) / Data Quality** - Each of the sources samples different properties with different footprints – extremely heterogeneous. Each of the sources has different levels of uncertainty and precision associated with it. In

addition, the translation across scales and domains introduces uncertainty as does the data mining. Data quality is critical.

- **Visualization** - Visualization is crucial to understanding the data.
- **Compute(System), storage, networking** - NSERC, ESNet
- **Specialized Software** - PFLOWTran, postgres, HDF5, Akuna, NEWT, etc
- **Current Data Analytics tools applied -** Data mining, data quality assessment, cross-correlation across datasets, reduced model development, statistics, quality assessment, data fusion, etc

g. Data Analytics Challenges (Gaps) - Translation across diverse and large datasets that cross domains and scales.

h. Type of User -

i. Dominant Data Analytics Skills Needed -

j. Science Research Areas -

k. Societal Benefit Areas -

l. Potential for and/or issues for generalizing this use case (e.g. for ref. architecture) - A wide array of programs in the earth sciences are working on challenges that cross the same domains as this project.

m. More Information (URLs) - Under development

5.16 **Use Case:** Climate Studies using the Community Earth System Model at DOE's NERSC center (borrowed, with permission, from NIST Big Data Use Case Submissions [http://bigdatawg.nist.gov/usecases.php])

a. **Title -** Climate Studies using the Community Earth System Model at DOE's NERSC center
b. **Author/Company/Email** - Steve Kempler, NASA/GSFC, Steven.J.Kempler@nasa.gov, adapted from Warren Washington, NCAR
c. **Actors/Stakeholders/Project URL and their roles and responsibilities -** Climate scientists, U.S. policy makers
d. **Use Case Goal** - The goals of the Climate Change Prediction (CCP) group at NCAR are to understand and quantify contributions of natural and anthropogenic-induced patterns of climate variability and change in the 20th and 21st centuries by means of simulations with the Community Earth System Model (CESM).
   (#8 - Forecast/Predict; #9 - Derive conclusions;#10 - Derive analytics tools)
e. **Use Case Description** - With these model simulations, researchers are able to investigate mechanisms of climate variability and change, as well as to detect and attribute past climate changes, and to project and predict future changes. The simulations are motivated by broad community interest and are widely used by the national and international research communities.
f. **Current technical issues/requirements to take into account that may impact needed data analytics:**
   - **Data Source (distributed/centralized)** - Data is produced at computing centers. The Earth Systems Grid is an open source effort providing a robust, distributed data and computation platform, enabling world wide access to Peta/Exa-scale scientific data. ESGF manages the first-ever decentralized database for handling climate science data, with multiple petabytes of data at dozens of federated sites worldwide. It is recognized as the leading infrastructure for the management and access of large distributed data volumes for climate change research. It supports the Coupled Model Intercomparison Project (CMIP), whose protocols enable the periodic assessments carried out by the Intergovernmental Panel on Climate Change (IPCC).
   - **Volume (size)** - 30 PB at NERSC (assuming 15 end-to-end climate change experiments) in 2017; many times more worldwide
   - **Velocity (e.g. real time**) - 42 GByles/sec are produced by the simulations
   - **Variety** - Data must be compared among those from observations, historical reanalysis, and a number of independently produced simulations. The Program for Climate Model Diagnosis and Intercomparison develops methods and tools for the diagnosis and intercomparison of general circulation models (GCMs) that simulate the global climate. The need for innovative analysis of GCM climate simulations is apparent, as increasingly more complex models are developed, while the disagreements among these simulations and relative to climate observations remain significant and poorly understood. The nature and causes of these disagreements

must be accounted for in a systematic fashion in order to confidently use GCMs for simulation of putative global climate change.

- **Veracity (Robustness Issues) / Data Quality** - EData produced by climate simulations is plays a large role in informing discussion of climate change simulations. Therefore it must be robust, both from the standpoint of providing a scientifically valid representation of processes that influence climate, but also as that data is stored long term and transferred world-wide to collaborators and other scientists**.**
- **Visualization** - Visualization is crucial to understanding a system as complex as the Earth ecosystem.
- **Compute(System), storage, networking** - NERSC (24M Hours), DOE LCF (41M), NCAR CSL (17M)
- **Specialized Software** - NCAR PIO library and utilities NCL and NCO, parallel NetCDF
- **Current Data Analytics tools applied -** There is a need to provide data reduction and analysis web services through the Earth System Grid (ESG). A pressing need is emerging for data analysis capabilities closely linked to data archives

g. Data Analytics Challenges (Gaps) - The rapidly growing size of datasets makes scientific analysis a challenge. The need to write data from simulations is outpacing supercomputers' ability to accommodate this need; Data from simulations and observations must be shared among a large widely distributed community

h. Type of User -

i. Dominant Data Analytics Skills Needed -

j. Science Research Areas -

k. Societal Benefit Areas -

l. Potential for and/or issues for generalizing this use case (e.g. for ref. architecture) - ESGF is in the early stages of being adapted for use in two additional domains: biology (to accelerate drug design and development) and energy (infrastructure for California Energy Systems for the 21st Century (CES21)).

m. More Information (URLs) -

http://esgf.org/
http://www-pcmdi.llnl.gov/
http://www.nersc.gov/
http://science.energy.gov/ber/research/cesd/
http://www2.cisl.ucar.edu/

5.17 **Use Case:** Radar Data Analysis for CReSIS (borrowed, with permission, from NIST Big Data Use Case Submissions [http://bigdatawg.nist.gov/usecases.php])

a. **Title -** Radar Data Analysis for CReSIS
b. **Author/Company/Email** - Steve Kempler, NASA/GSFC, Steven.J.Kempler@nasa.gov, adapted from Geoffrey Fox, Indiana University gcf@indiana.edu
c. **Actors/Stakeholders/Project URL and their roles and responsibilities -** Research funded by NSF and NASA with relevance to near and long term climate change. Engineers designing novel radar with "field expeditions" for 1-2 months to remote sites. Results used by scientists building models and theories involving Ice Sheets
d. **Use Case Goal** - Determine the depths of glaciers and snow layers to be fed into higher level scientific analyses
   (#6 - Tease out information)
e. **Use Case Description** - Build radar; build UAV or use piloted aircraft; overfly remote sites (Arctic, Antarctic, Himalayas). Check in field that experiments configured correctly with detailed analysis later. Transport data by air-shipping disk as poor Internet connection. Use image processing to find ice/snow sheet depths. Use depths in scientific discovery of melting ice caps etc.
f. **Current technical issues/requirements to take into account that may impact needed data analytics:**
   - **Data Source (distributed/centralized)** - Aircraft flying over ice sheets in carefully planned paths with data downloaded to disks
   - **Volume (size)** - ~0.5 Petabytes per year raw data
   - **Velocity (e.g. real time)** -
   - **Variety** - All data gathered in real time but analyzed incrementally and stored with a GIS interface.
   - **Veracity (Robustness Issues) / Data Quality** - Essential to monitor field data and correct instrumental problems. Implies must analyze fully portion of data in field.
   - **Visualization** - Rich user interface for layers and glacier simulations; Main engineering issue is to ensure instrument gives quality data.
   - **Compute(System), storage, networking** - Field is a low power cluster of rugged laptops plus classic 2-4 CPU servers with ~40 TB removable disk array. Off line is about 2500 cores; Removable disk in field. (Disks suffer in field so 2 copies made) Lustre or equivalent for offline; Terrible Internet linking field sites to continental USA
   - **Specialized Software** - Radar signal processing in Matlab. Image analysis is MapReduce or MPI plus C/Java. User Interface is a Geographical Information System
   - **Current Data Analytics tools applied -** Sophisticated signal processing; novel new image processing to find layers (can be 100's one per year).
g. **Data Analytics Challenges (Gaps)** - Data volumes increasing. Shipping disks clumsy but no other obvious solution. Image processing algorithms still very active research.
h. **Type of User** -
i. **Dominant Data Analytics Skills Needed** -

j. **Science Research Areas** -

k. **Societal Benefit Areas** -

l. Potential for and/or issues for generalizing this use case (e.g. for ref. architecture) - Loosely coupled clusters for signal processing. Must support Matlab.

m. More Information (URLs) - http://polargrid.org/polargrid; https://www.cresis.ku.edu/; See movie at http://polargrid.org/polargrid/gallery

5.18 **Use Case:** UAVSAR Data Processing, Data Product Delivery, and Data Service (borrowed, with permission, from NIST Big Data Use Case Submissions [http://bigdatawg.nist.gov/usecases.php])

a. **Title -** UAVSAR Data Processing, Data Product Delivery, and Data Service
b. **Author/Company/Email** - Steve Kempler, NASA/GSFC, Steven.J.Kempler@nasa.gov, adapted from Andrea Donnellan, NASA JPL, andrea.donnellan@jpl.nasa.gov; Jay Parker, NASA JPL, jay.w.parker@jpl.nasa.gov
c. **Actors/Stakeholders/Project URL and their roles and responsibilities -** NASA UAVSAR team, NASA QuakeSim team, ASF (NASA SAR DAAC), USGS, CA Geological Survey
d. **Use Case Goal** - Use of Synthetic Aperture Radar (SAR) to identify landscape changes caused by seismic activity, landslides, deforestation, vegetation changes, flooding, etc; increase its usability and accessibility by scientists
   (#6 - Tease out information)
e. **Use Case Description** - A scientist who wants to study the after effects of an earthquake examines multiple standard SAR products made available by NASA. The scientist may find it useful to interact with services provided by intermediate projects that add value to the official data product archive..
f. **Current technical issues/requirements to take into account that may impact needed data analytics:**
   - **Data Source (distributed/centralized)** - Data initially acquired by unmanned aircraft. Initially processed at NASA JPL. Archive is centralized at ASF (NASA DAAC). QuakeSim team maintains separate downstream products (GeoTIFF conversions).
   - **Volume (size)** - RPI Data: 1-2 TB/year. Polarimetric data is faster.
   - **Velocity (e.g. real time)** -
   - **Variety** - Two main types: Polarimetric and RPI. Each RPI product is a collection of files (annotation file, unwrapped, etc). Polarimetric products also consist of several files each.
   - **Veracity (Robustness Issues) / Data Quality** - Provenance issues need to be considered. This provenance has not been transparent to downstream consumers in the past. Versioning used now; versions described in the UAVSAR web page in notes.
   - **Visualization** - Uses Geospatial Information System tools, services, standards.
   - **Compute(System), storage, networking** - Raw data processing at NASA AMES Pleiades, Endeavour. Commercial clouds for storage and service front ends have been explored; File based; Data require one time transfers between instrument and JPL, JPL and other NASA computing centers (AMES), and JPL and ASF. Individual data files are not too large for individual users to download, but entire data set is unwieldy to transfer. This is a problem to downstream groups like QuakeSim
   - **Specialized Software** - ROI_PAC, GeoServer, GDAL, GeoTIFF-suporting tools

- **Current Data Analytics tools applied -** Done by downstream consumers (such as edge detections): research issues.

g. Data Analytics Challenges (Gaps) - Data processing pipeline requires human inspection and intervention. Limited downstream data pipelines for custom users. Cloud architectures for distributing entire data product collections to downstream consumers should be investigated, adopted.

h. Type of User -

i. Dominant Data Analytics Skills Needed -

j. Science Research Areas -

k. Societal Benefit Areas -

l. Potential for and/or issues for generalizing this use case (e.g. for ref. architecture) - Data is geolocated, and may be angularly specified. Categories: GIS; standard instrument data processing pipeline to produce standard data products.

m. More Information (URLs) - http://uavsar.jpl.nasa.gov/; http://www.asf.alaska.edu/program/sdc; http://quakesim.org

2. State of Available Earth Science Data Analytics Tools and Techniques

3. Gap Analysis: Where Information Technology Potential Needs to be Applied

4. Introduction

5. Goal: Utilize Information Technologies to Glean Knowledge from Information through Advanced Information Analysis Tools and Techniques: Data Analytics

6. Scope and State of Earth Science Data Analytics
   Data Preparation – Making heterogeneous data so that they can ‘play’ together
   Data Reduction – Smartly removing data that do not fit research criteria
   Data Analysis – Applying techniques/methods to derive results

7. Community Activities

USE CASE TEMPLATE and USE CASES next page