OceanXtremes: Oceanographic Data-Intensive Anomaly Detection and Analysis Portal
AIST-14

PI: Thomas Huang
Co-Is: Ed Armstrong, George Chang, Toshio Chin, and Brian Wilson
Collaborators: Tong (Tony) Lee, and Victor Zlotnicki

Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive
Pasadena, CA 91109-8099
United States of America

Overview

- OceanXtremes is a computational platform powered by an intelligent, Cloud-based analytic service backend that enables execution of domain-specific, multi-scale anomaly and feature detection algorithms across the entire archive of ocean science datasets.

- On-Premise Cloud Computing environment in JPL, where it is closed to the oceanography data center

- Using this platform scientists can efficiently search for anomalies or ocean phenomena, compute data metrics for events or over time-series of ocean variables, and efficiently find and access all of the data relevant to their study (and then download only that data).

- The OceanXtremes' analytic backend will demonstrate three new technology ideas to provide rapid turn around on climatology computation and anomaly detection:
  1. An adaption of the MapReduce framework for **parallel data-mining** of science datasets, typically large 3 or 4-dimensional arrays packaged in NetCDF and HDF.
  2. An algorithm profiling service to efficiently and cost-effectively scale up **hybrid Cloud computing resources** based on the needs of scheduled jobs (CPU, memory, network, and bursting from a private Cloud computing cluster to public cloud provider like Amazon Cloud services)
  3. An extension to industry-standard search solutions (OpenSearch and Faceted search) to provide support for **shared discovery and exploration of ocean phenomena and anomalies**, along with unexpected correlations between key measured variables.

Anomaly Detection

- Anomaly detection is a process of identifying items, events or observations, which do not conform to an expected pattern in a dataset or time series.

- Typically this is a two-stage procedure
  1. Determine a long-term/periodic mean ("climatology")
  2. Deviations from the mean are searched. Step 1 could be omitted in cases where a climatology data set already exists.
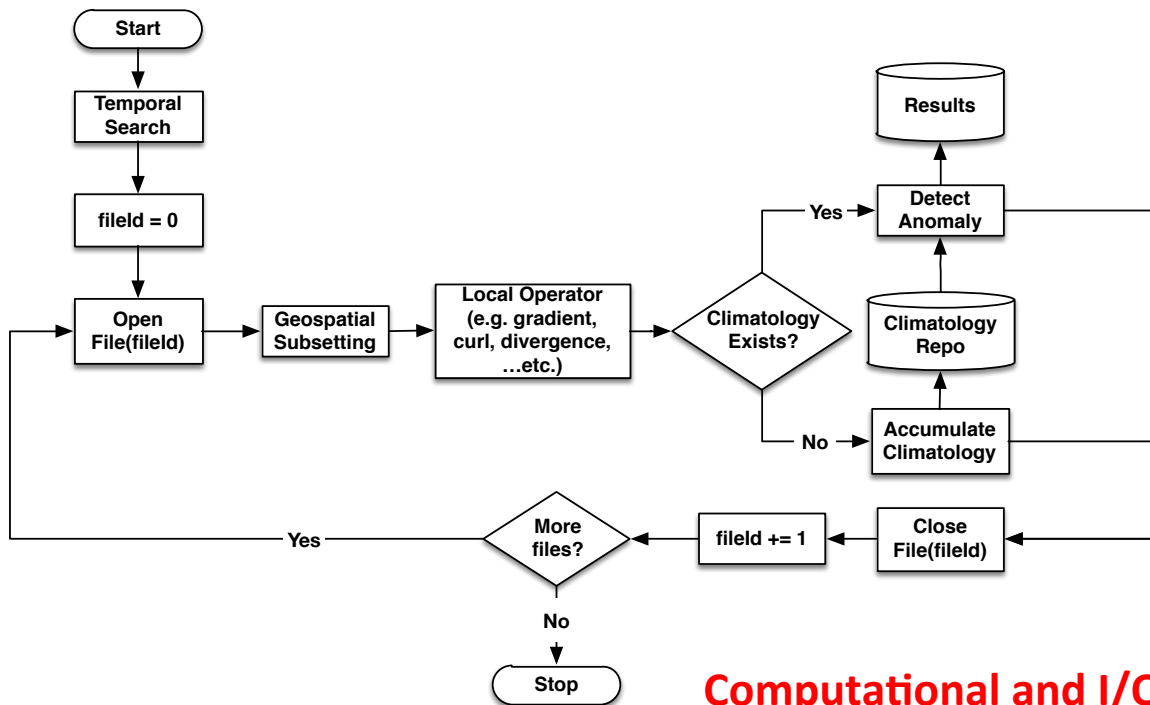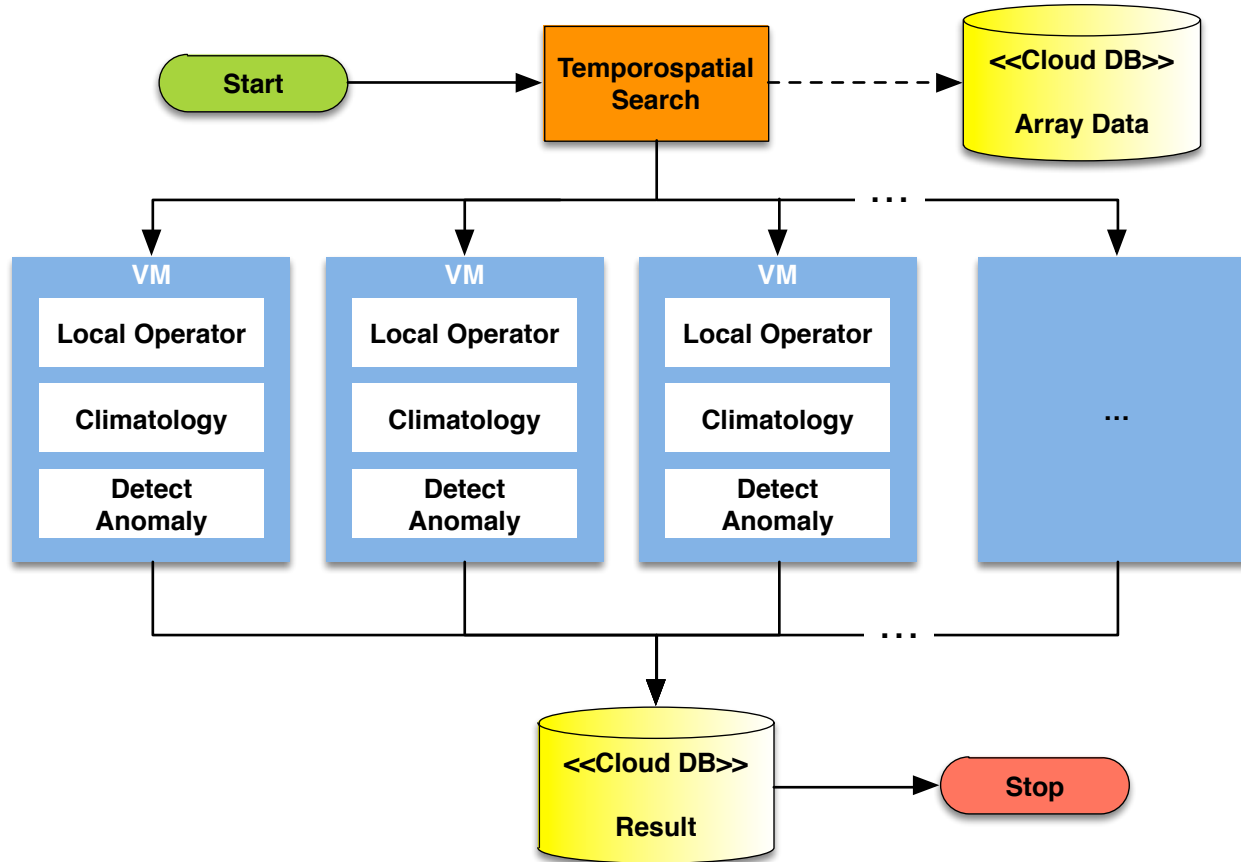

Use Cases

- El Niño/La Nina anomaly characterization
- El Niño /La Nina teleconnections
- Ocean features: Gradients, frontal detection, upwelling
- Rogue waves in high resolution altimeter data

Computational and I/O Intensive

# OceanXtremes: High-Level Workflow

Proposed high-level workflow

- ✓ Leverage Virtual Machine technology

- ✓ Leverage the elasticity of Cloud Computing

- ✓ Leverage Cloud data store for high-performance search and read

- ✓ Leverage and extend technologies developed at the NASA Physical Oceanography Distributed Active Archive Center (PO.DAAC)

- ✓ Leverage and extend technologies developed through several other funded projects in relation to PO.DAAC

- ✓ Leverage industry standard, open-source data processing/analysis solutions

Our Experience with Cloud Computing

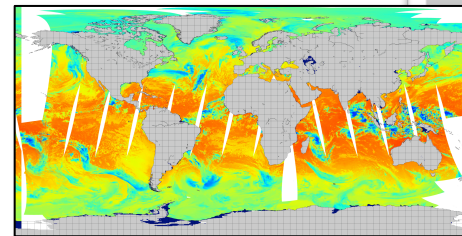- **2011 Technology Infusion – Cloud Computing Study**
  - Amazon
  - NASA Nebula
  - Apache Hadoop and HBase
  - Built Climatology Service
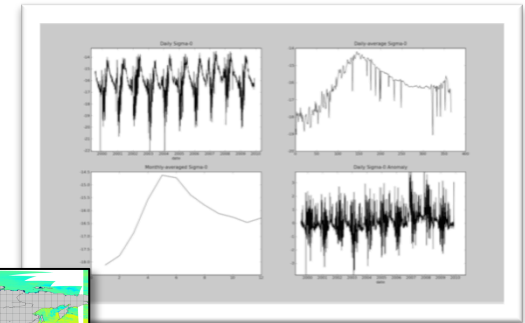
- **2014 NASA ITLabs Cloudbursting**

- **2014 ESDSWG Cloud Tasks**
  - Platform as a Service
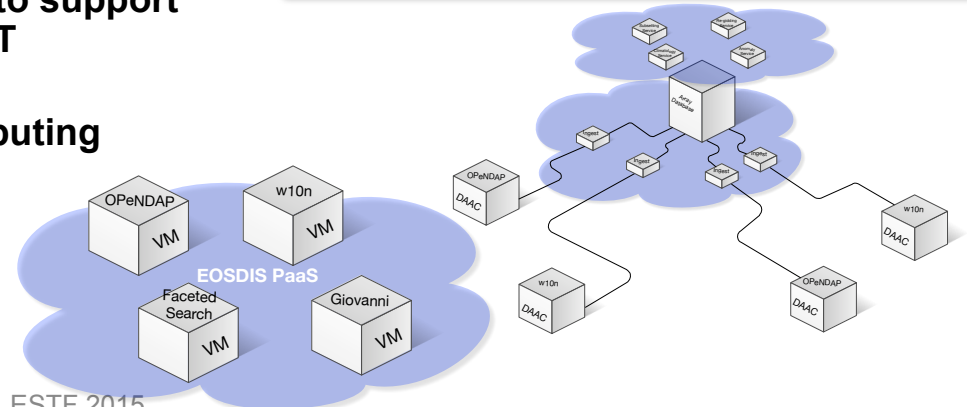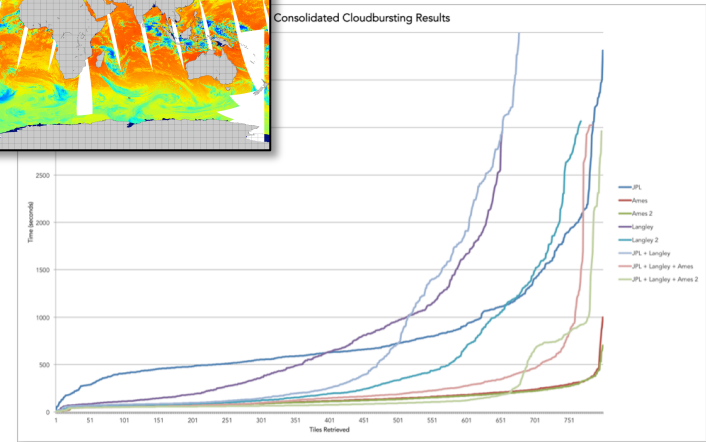 - Science Data Service Platform

- **2015 SMAP Big Data Analysis Service**

- **2015 SMAP Imagery System on Amazon GovCloud**

- **2015 PO.DAAC Private Cloud Platform to support future Big Data missions such as SWOT**
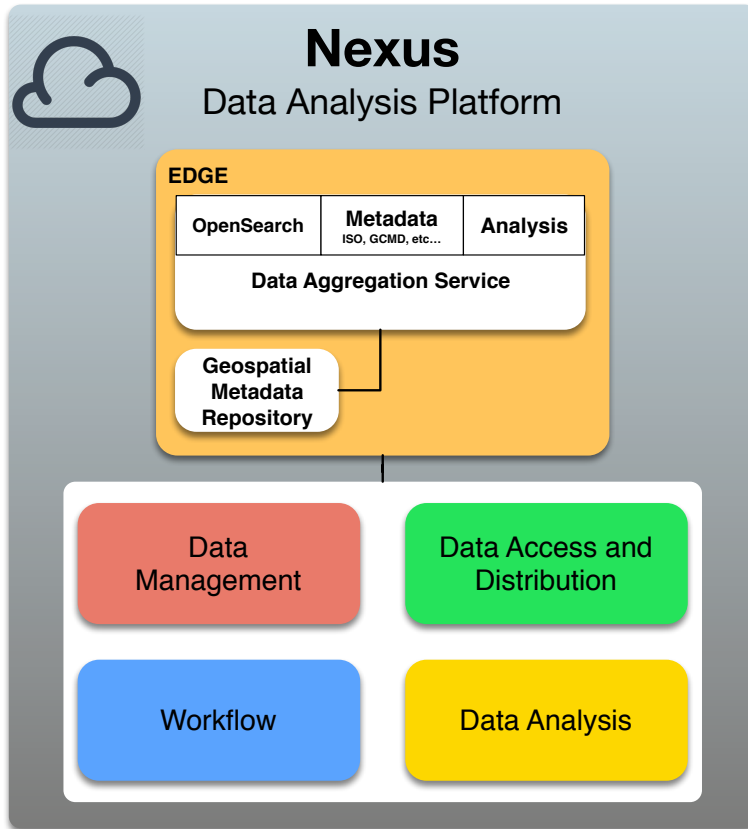
- **Working with Big Data and Cloud Computing Communities**
  - ESIP Federation
  - NASA ESDSWG
 - NIST Big Data Working Group
  - EarthCube

Nexus: Data Analysis Platform

Nexus
Data Analysis Platform

EDGE

OpenSearch	Metadata ISO, GCMD, etc…	Analysis

Data Aggregation Service

Geospatial Metadata Repository

Data Management

Data Access and Distribution

Workflow

Data Analysis

- Data analysis platform on the Cloud
- Data management and transformation
- Multi-disciplinary data coordination
- On-the-fly analysis services
  - Time series
  - Correlation
  - Re-gridding
  - Data subsetting
  - Data visualization service
- RESTful access to geospatial array data

- Applications
  - NASA Sea Level Change Portal
  - AIST-14: DOMS
  - ACCESS-13: Virtual Quality Screening Service
  - PO.DAAC's next generation of subsetting and data analysis suites

Integration with SMAP Interactive Map Client
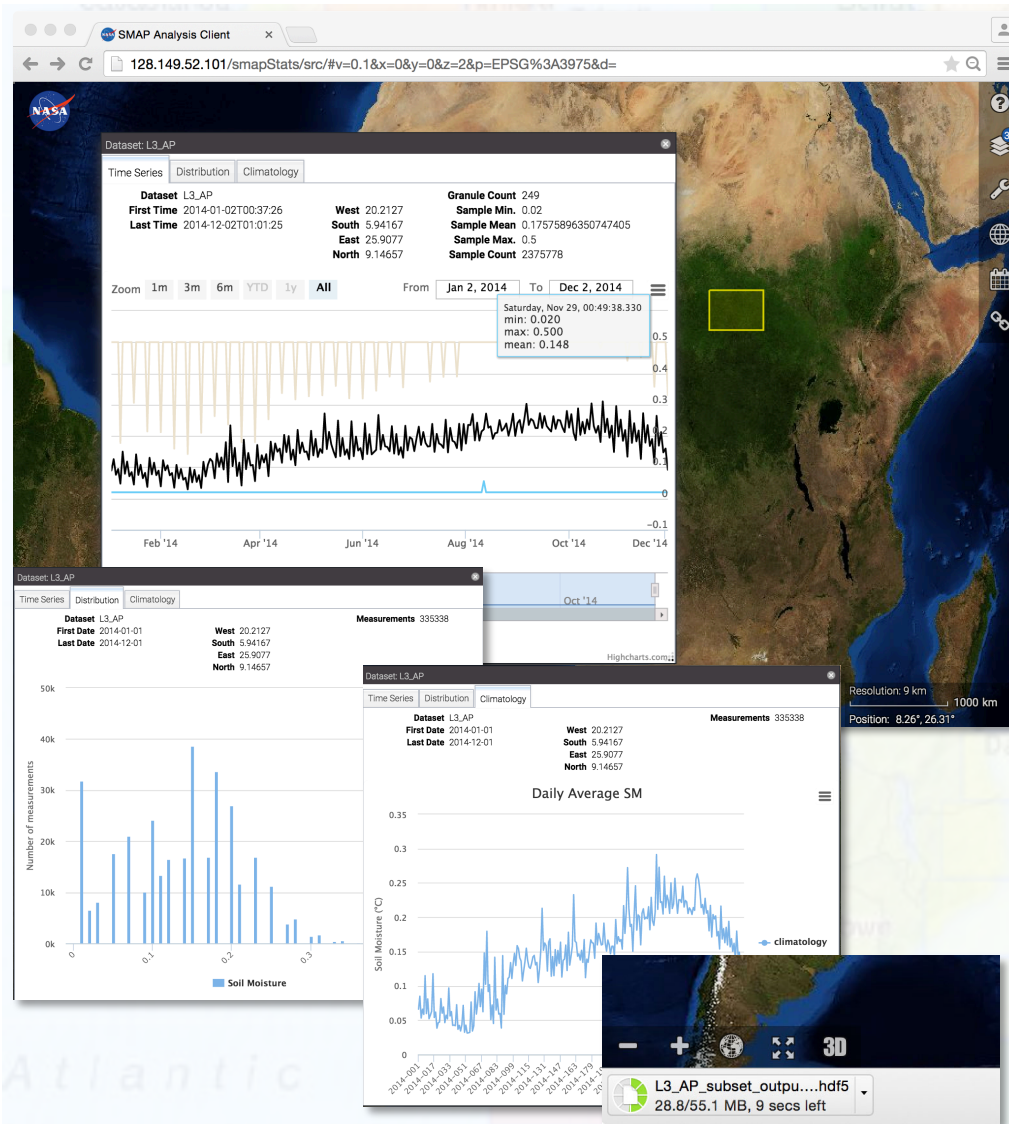
Provide on-the-fly data analysis
- L2/L3 Soil Moisture Active
- L2/L3 Soil Moisture Passive
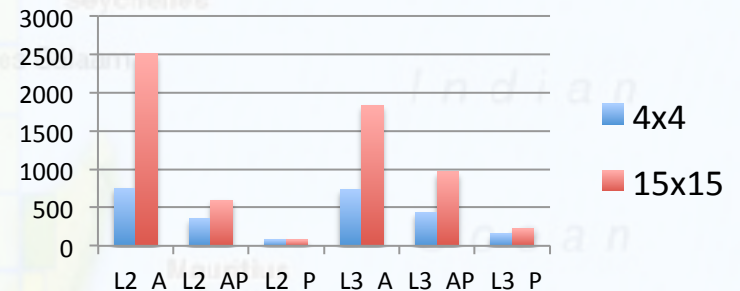- L2/L3 Soil Moisture Active/Passive

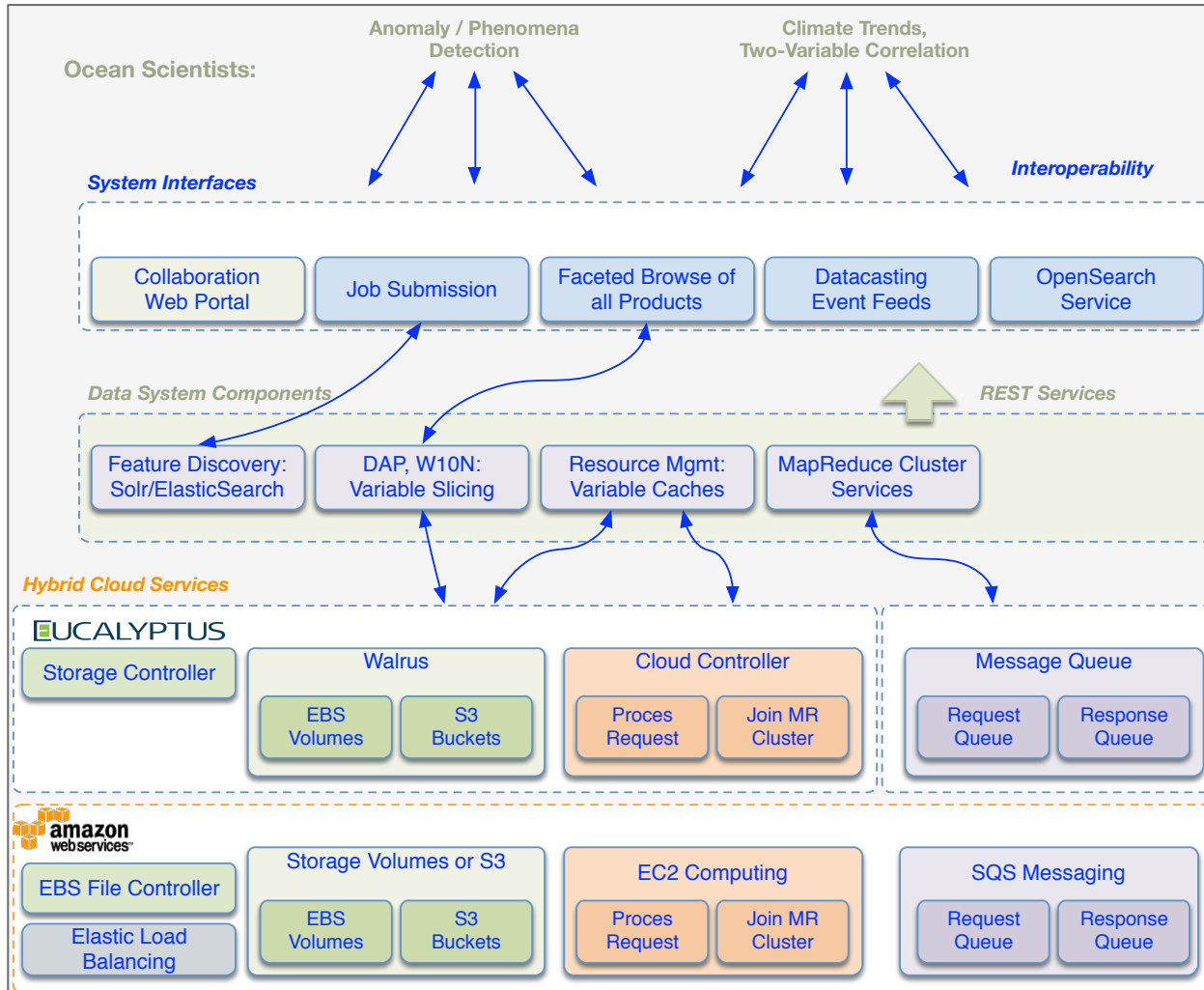On-the-fly generation of
- Daily, weekly, monthly averages
- Variance
- Histogram

Monthly Distribution/Climatology Generation (ms)

OceanXtremes High-Level Architecture

Key Milestones

Project scheduled start May, 2015

OceanXtremes	May-Jul'15	Aug-Oct'15	Nov-Jan'16	Feb-Apr'16	May-Jul'16	Aug-Oct'16	Nov-Jan'17	Feb-Apr'17
Procure and install OceanXtremes hardware	◇							
Design OceanXtremes backend system		◇						
Select data(s) and algorithm(s)		◇						
Develop and test OceanXtremes backend				◇				
Design web portal					◇			
Develop and test web portal							◇	
Expand OceanXtremes datasets and algorithm support							◇	
Integrate Datacasting capability							◇	
Evaluate and integrate data visualization solution							◇	
Perform end-to-end demonstration and benchmarking								◇

THANKS

Questions, and more information

thomas.huang@jpl.nasa.gov