Improving Oceanographic Anomaly Detection Using High Performance Computing

Thomas Huang, **Ed Armstrong**, George Chang, Toshio Chin, Brian Wilson, Tong (Tony) Lee, Victor Zlotnicki. Jorge Vazquez and Michelle Gierach

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

2015 Pacific Anomalies Symposium

5 May 2015
CL#15-1921

Introduction

- Anomaly detection is a process of identifying items, events or observations outside the “norm” or expected patterns
- Current and future oceanographic missions for SSH, SST, OC, Ocean Wind will present us with challenges to identify features and anomalies in increasingly complex and voluminous observations
- **OceanXtremes**, a NASA technology effort, is powered by an intelligent, elastic 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.
  - User defines own anomaly or feature types with continuous backend executing the chosen data mining algorithm (e.g. differences from climatology or gradients above a specified threshold)
  - Feature types: Anomalies, gradients, eddies
  - Algorithms and data mining: Thresholds, curl and divergence, correlations, EOFs
- A key idea is that the parallel data-mining operations will be run “near” the ocean data archives (a local “network” hop)
- Funded by NASA Advanced Information Systems Technology (AIST) program in 2015
- Stakeholder inputs
  - Provide use cases
  - Feedback and testing

OceanXtremes Overview

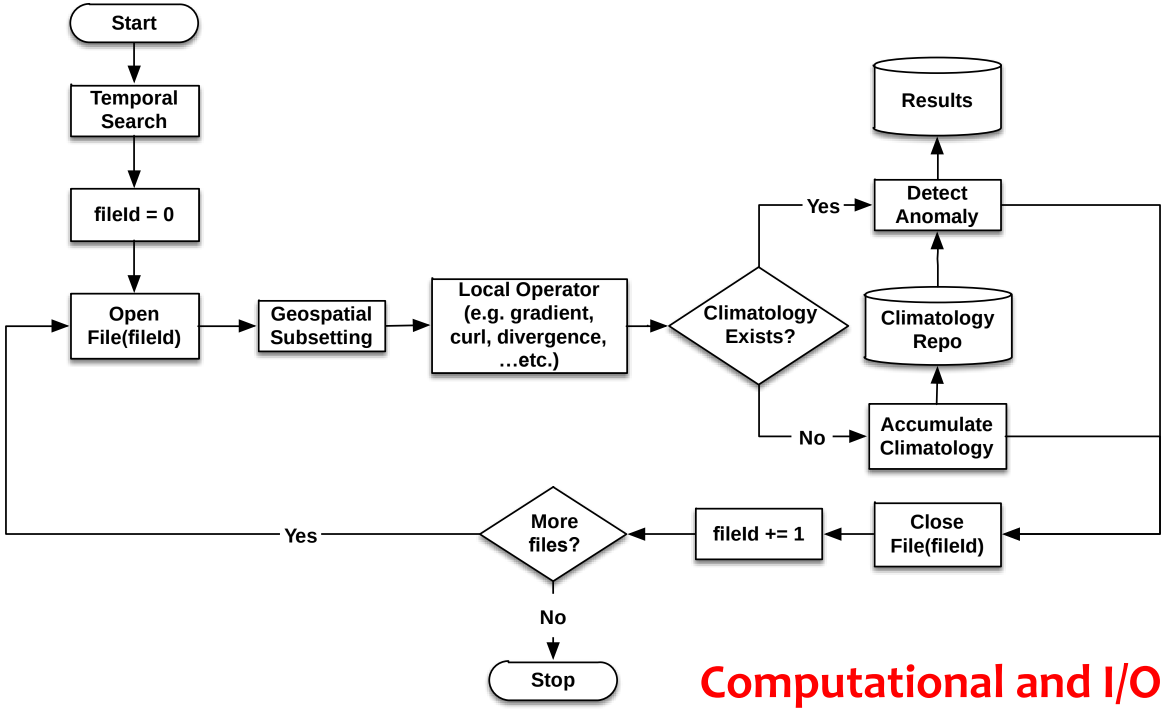
- OceanXtremes is a computational platform powered by an intelligent, Cloud-based analytic service backend that enables execution of domain-specific, multi-scale anomaly and feature detection algorithms across the entire archive of ocean science datasets.
- On-Premise Cloud Computing environment in JPL,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

Datasets

Dataset	Key Variables	Time Range	Data Mining Operators Needed	Phenomenon
CCMP L4, Pathfinder L3, Integrated Altimeter L4	Wind SST SSH	1987-2011 1982-2013 1992-2013	Anomaly calculation from fixed or on-the-climatology, Threshold detection. Variance characterization	El Niño genesis, anomaly detection and characterization in different regions (3.4 vs 4). Coastal upwelling
MODIS L3, Pathfinder L3, CCMP L4, Integrated Altimeter L4, MODIS L3, Aquarius L3	SST Wind SSH Color Salinity	2000-present 1982-2013 1987-2011 1992-2013 2000-present 2011-present	Cross correlations. Covariabilty and EOFs.	El Niño and other teleconnections. Regional correlations
MUR L4, MODIS L3, CCMP L4	SST Wind	2002-present 2000-present 1987-2011	Divergence and curl.	Upwelling. Hurricane genesis
MODIS L3, MODIS L3	SST Color	2000-present	Matched filter (e.g., Sobel operator). First derivatives.	Gradients, edges, and eddy detection
Pathfinder L3, CCMP L4, Integrated Altimeter L4	SST Wind SSH	1982-2013 1987-2011 1992-2013	Regression, Polynomial fits. Variance.	Trends. Basin scale variability

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.



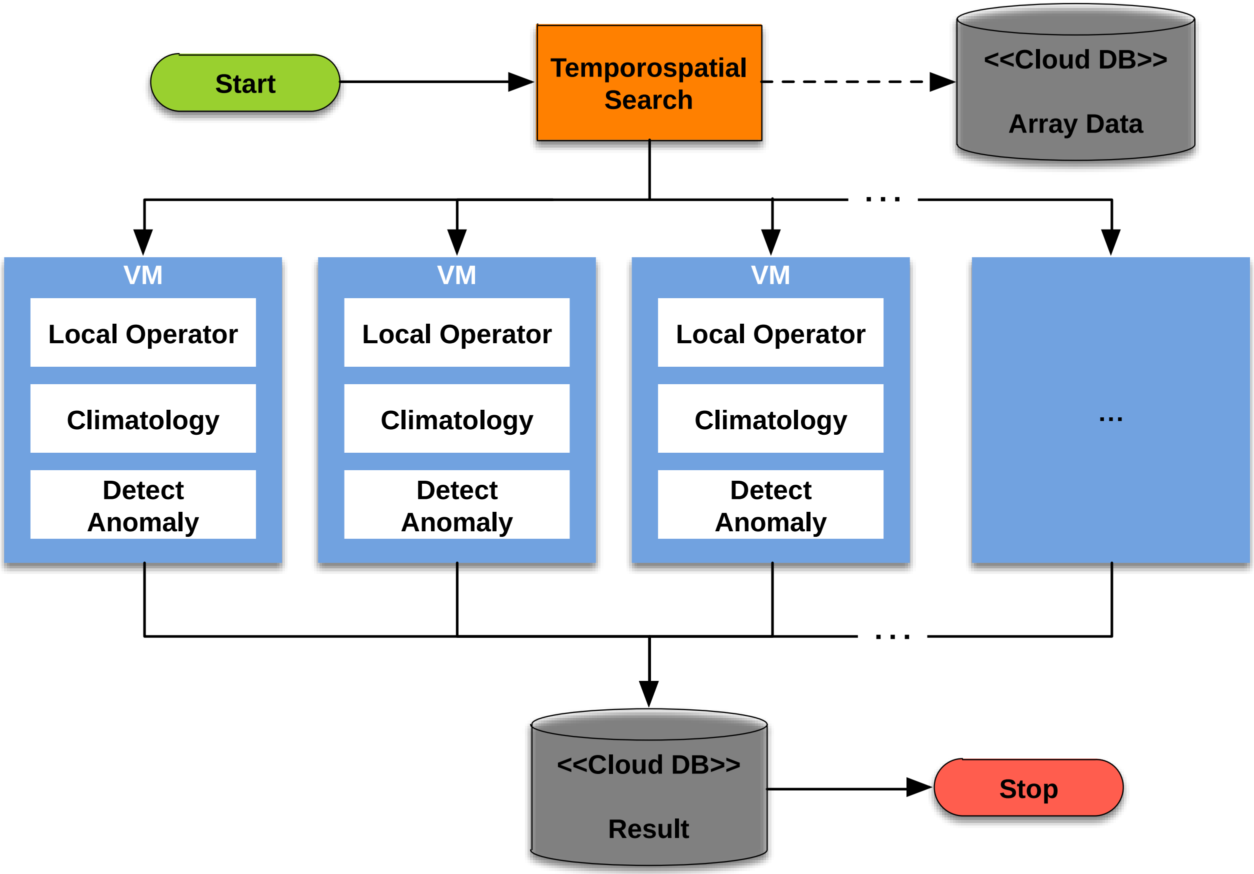
User Cases

- El Niño/La Nina anomaly detection and 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

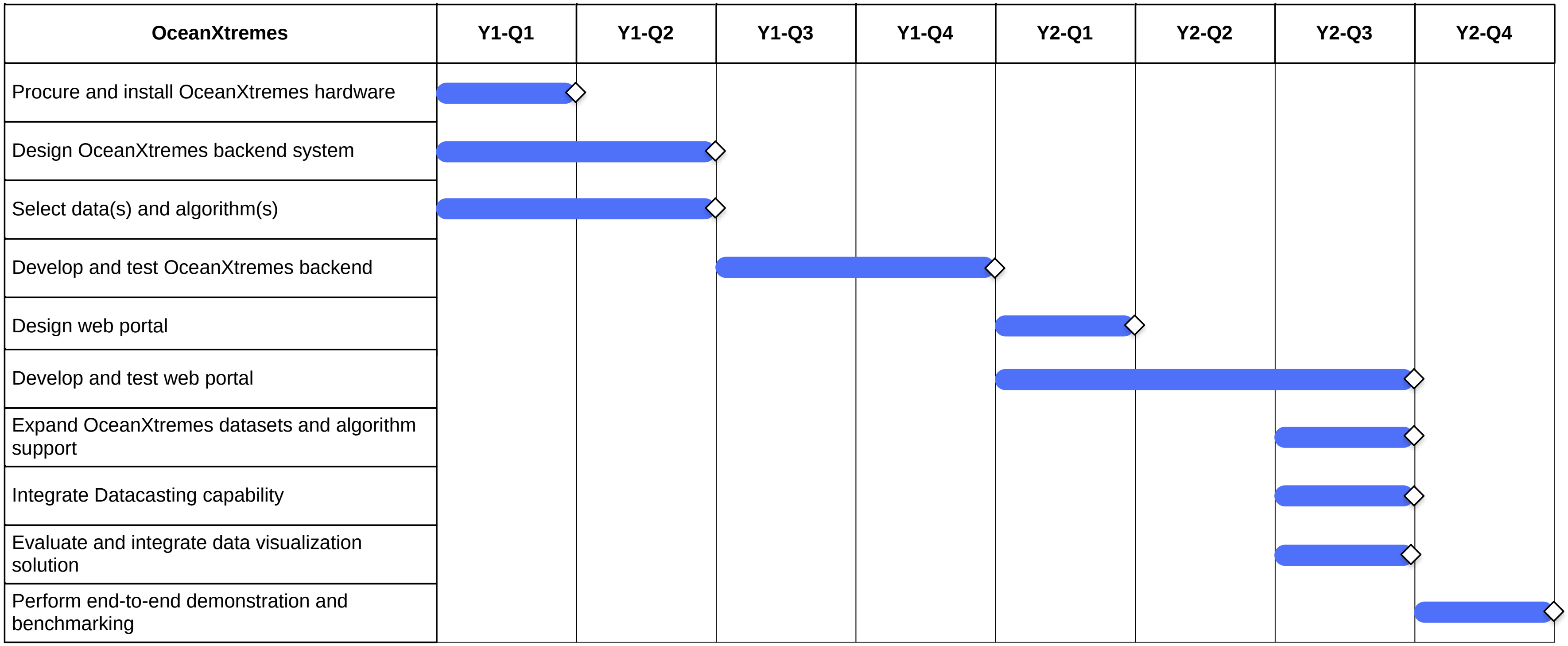
Proposed high-level workflow



- ✓ 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

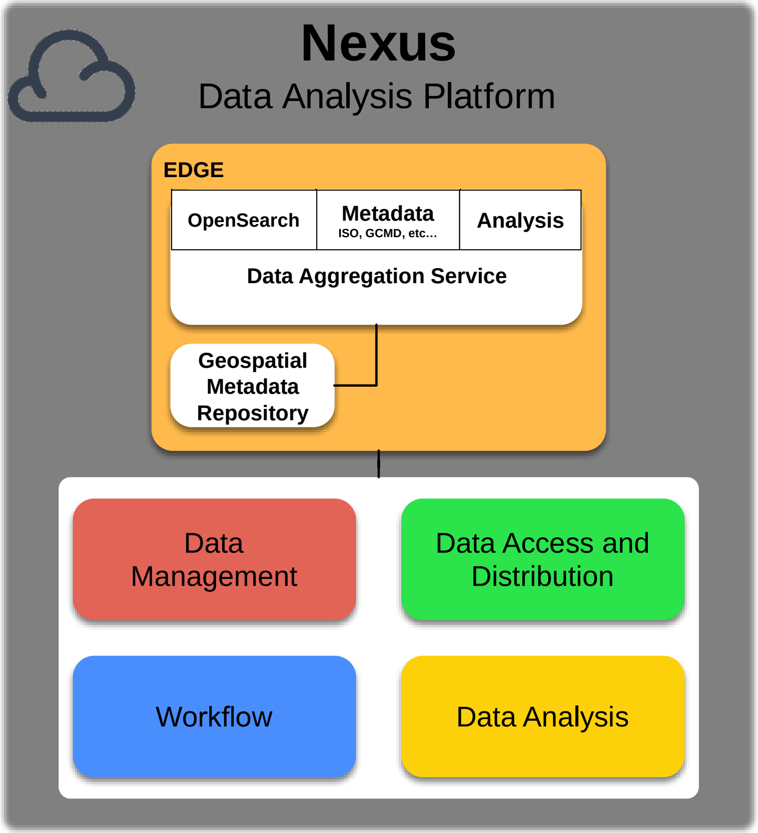
Key Milestones

Project scheduled to start on June 1, 2015



OceanXtremes	Y1-Q1	Y1-Q2	Y1-Q3	Y1-Q4	Y2-Q1	Y2-Q2	Y2-Q3	Y2-Q4
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								━◇

Nexus: Data Analysis Platform

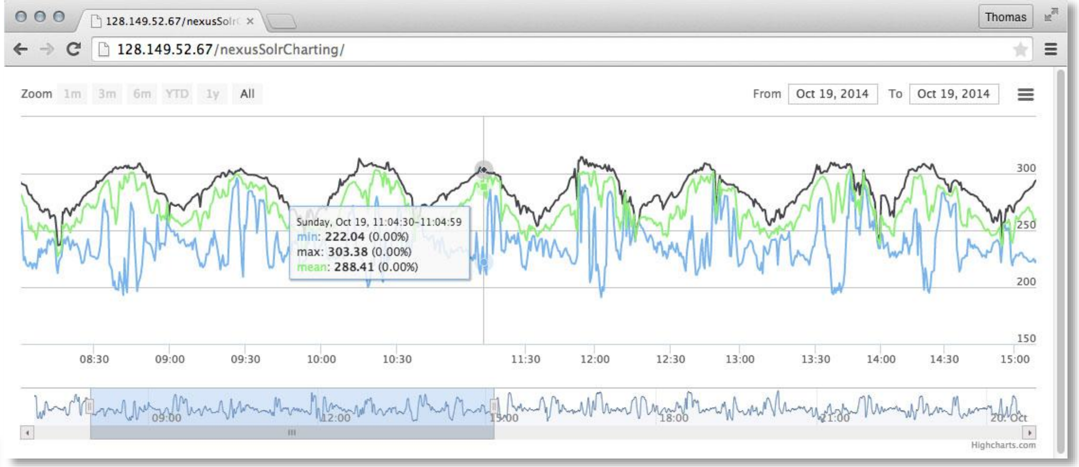


- Data analysis platform on the Cloud
- Data management and transformation
- Multi-disciplinary data coordination
- On-the-fly analysis services
  - Time series
  - Correlation
  - Re-gridding
  - Data subsetting
  - Data visualization service
- RESTful access to geospatial array data

- 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

Nexus in Action

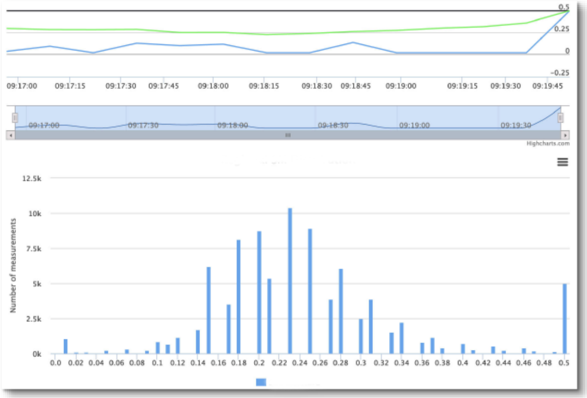
- Integrated to support PO.DAAC data and others
  - Supports L2+ data
  - On-the-fly Time-series generation
  - On-the-fly Histogram generation
  - On-the-fly data subsetting of oceanographic data



128.149.52.67/nexusSolrCharting/

Sunday, Oct 19, 11:04:30-11:04:59
min: 222.04 (0.00%)
max: 303.38 (0.00%)
mean: 288.41 (0.00%)

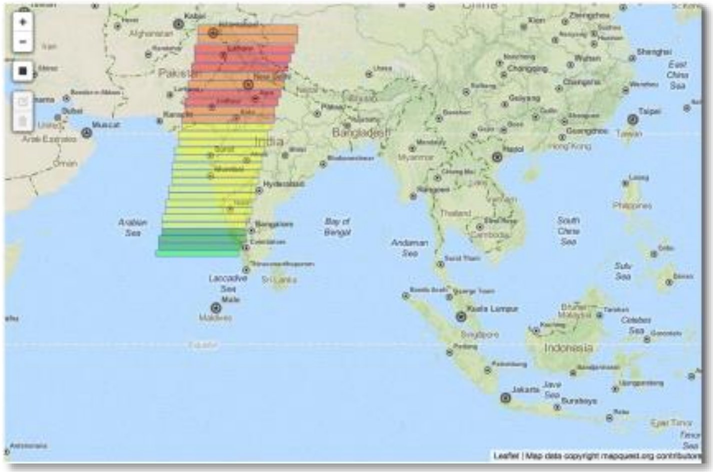
Time-series



L3_AP_subset_outpu....hdf5
28.8/55.1 MB, 9 secs left

Data subsetting

Histogram



New approach to mange data for analysis

THANKS

Questions, and more information

thomas.huang@jpl.nasa.gov
Edward.m.armstrong@jpl.nasa.gov