IOOS Advisory Committee Meeting

5 -6 April 2016

Big Data Brief

Thomas B. Curtin

Table 1: Summary of four big data characteristics

Characteristic	Description	Attributes	Drivers
Volume	The amount of data generated or data intensity that must be ingested, analyzed and managed to make decisions based on complete data analysis.	- Exabyte, zettabyte, yottabyte, etc.	- Increase in data sources - Higher resolution sensors - Scalable infrastructure
Velocity	How fast data is being produced and changed and the speed at which data is transformed into insight.	- Batch - Near real-time - Real-time - Streams - Rapid feedback loop	- Improved throughput connectivity - Competitive advantage - Precomputed information
Variety	The degree of diversity of data from sources both inside and outside an organization.	- Degree of structure - Complexity	- Mobile - Social media - Video - Genomics - M2M / IoT
Veracity	The quality and provenance of data.	- Consistency - Completeness - Integrity - Ambiguity	- Cost - Need of traceability and justification

Adapted from TechAmerica: Demystifying big data[23]

Big Data Value Chain

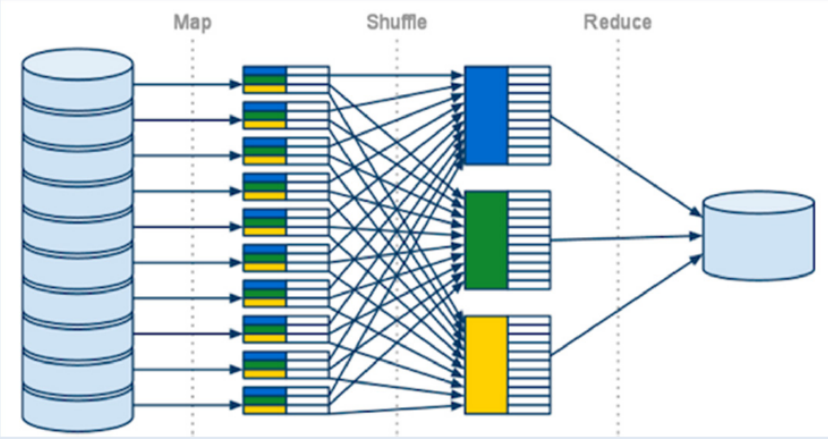
1) Data is collected where it originates from a variety of sources: sensors, human input, etc.

2) Raw data is combined with data from other sources, classified and stored in a data repository.

3) Algorithms and analysis are applied by an intelligence engine to interpret and provide utility to the aggregated data.

4) Outputs of the intelligence engine are converted to tangible values, insights or recommendations.

Box 4: Processing datasets in stages: Map/Reduce model in Google App Engine



- *Reading input:* An input reader reads data from storage (or some other source) and passes it to the next stage. You can choose a pre-defined reader from a list of choices. E.g., one input reader reads all datastore entities of a specified kind, passing each entity as input to the next stage.
- *Map:* You write a map function that runs once for each input value. It returns a collection of name-value pairs which are passed on to the next stage. If there are many, many input values, then this function runs many, many times. The framework divides up the input so that subsets are handled in parallel on multiple instances of your application. A typical map function could count things that occur in each input value that matches some filter.
- *Shuffle:* The Map/Reduce framework "shuffles" together the values returned by the map function. The shuffler groups together name-value pairs that have the same name and then passes those groups on to the next stage.
- *Reduce:* You write a reduce function that runs once for each "name" used in the name-value pairs. If there are many, many names then this function may run many, many times. It returns a collection of values that are passed along to the next stage.
- *Writing output:* An output writer concatenates together the outputs of the reduce functions in arbitrary order and writes them to persistent storage. You can choose a pre-defined writer from a list of choices.

Source: Google App Engine MapReduce Python Overview, https://developers.google.com/appengine/docs/python/dataprocessing/, licensed under the Creative Commons Attribution 3.0 License (http://creativecommons.org/licenses/by/3.0/)

Big Data Software
(free, open source)

Hadoop, MongoDB, Spark, Storm

Hadoop, a free and open-source implementation of the Map/Reduce programming model,

- **processes huge datasets (petabytes of data) in a scalable manner by commissioning parallel processing capabilities that move data subsets to distributed servers**

- **provides a distributed file system (HDFS) that can store data on thousands of computer nodes, providing very high aggregate bandwidth across the whole cluster**

Both Map/Reduce and the distributed file system are designed so that node failures are automatically handled by the framework to re-assign the tasks among other nodes.

Global Strategic Principles

The G8 has set out five strategic principles aimed at unlocking the economic potential of open data, supporting innovation, and providing greater accountability. These principles include an expectation that all government data will be published openly by default, alongside principles to increase the quality, quantity and re-use of the data released.

Table 2: Examples of open data portals

State / organization	Website
Belgium	http://data.gov.be/
Ghana	http://data.gov.gh/
India	http://data.gov.in/
Kenya	https://www.opendata.go.ke/
Morocco	http://data.gov.ma/
Russia	http://opengovdata.ru/
United Arab Emirates	http://www.government.ae/web/guest/uae-data
United Kingdom	http://data.gov.uk/
United States of America	http://www.data.gov/
European Union	http://open-data.europa.eu/
Organization for Economic Co-operation and Development (OECD)	http://stats.oecd.org/
United Nations (UN)	http://data.un.org/
United Nations High Commissioner for Refugees (UNHCR)	http://data.unhcr.org/
World Bank	http://data.worldbank.org/

Note: This list is by no means exhaustive.

Big Data Standards

The Cloud Security Alliance (GSA) established a big data working group in 2012 to identify scalable techniques for data-centric security and privacy problems. The group's investigation is expected to clarify best practices for security and privacy in big data, and also to guide industry and government in the adoption of those best practices.

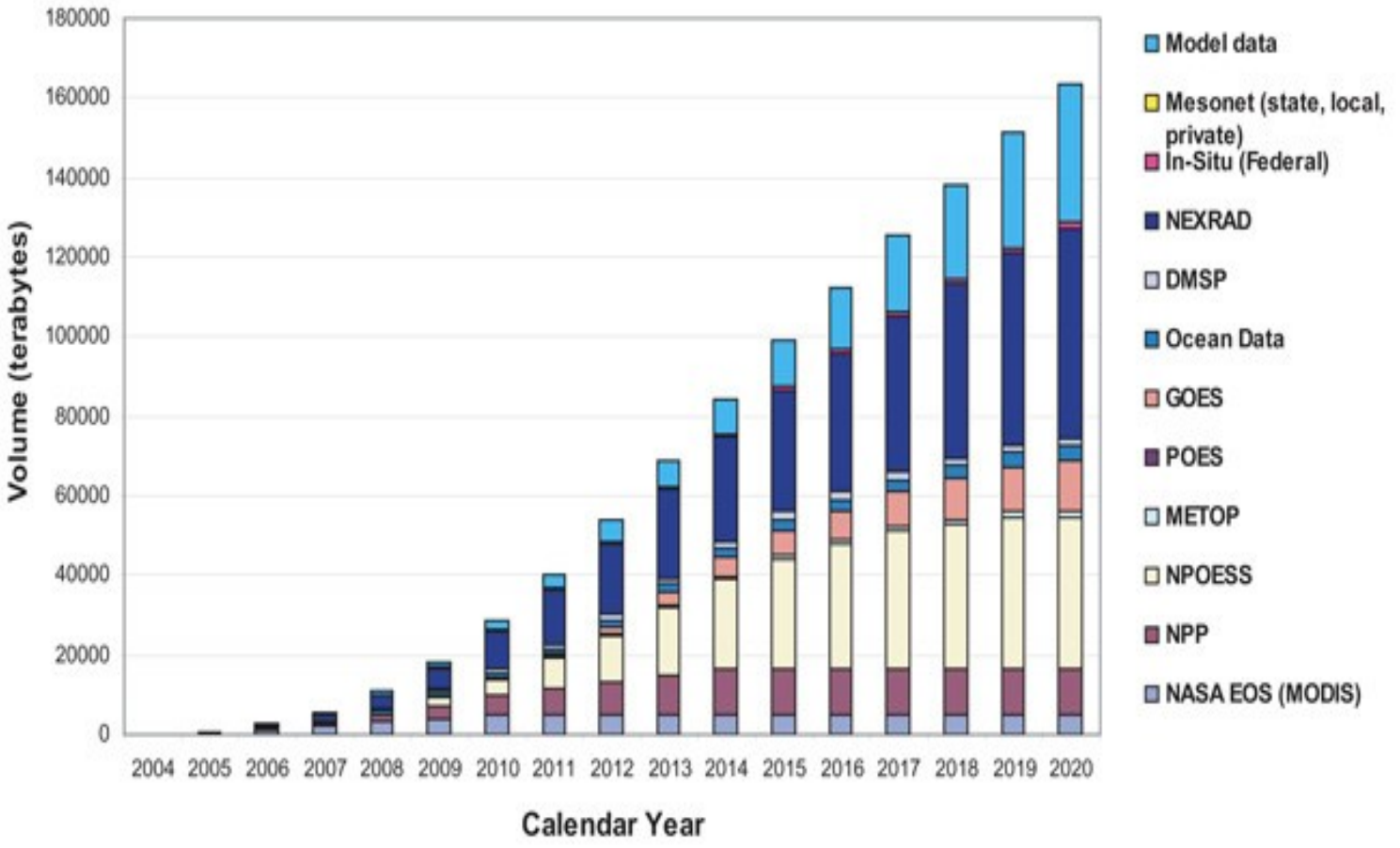
The U.S. National Institute of Standards and Technology (NIST) kicked-off its big data activities with a workshop in June 2012 and a year later launched a public working group. The NIST working group seeks consensus on definitions, taxonomies, secure reference architectures and a technology roadmap for big data analytic techniques and technology infrastructures.

ISO/IEC JTC1's data management and interchange standards committee (SC32) has initiated a study on next generation analytics and big data. The W3C has created several community groups on different aspects of big data.

ITU's standardization activities address individual infrastructure requirements, noting existing work in domains including optical transport and access networks, future network capabilities (e.g., software-defined networks), multimedia and security. A work item has been initiated to study the relationship between cloud computing and big data. The recently determined Recommendation ITU-T X.1600, "Security framework for cloud computing", matches security threats with mitigation techniques and the future standardization of the described threat-mitigation techniques is expected to incorporate big data use cases.

NOAA Data

- NOAA gathers over 20 terabytes of data every day – more than twice the data of the entire printed collection of the United States Library of Congress.
- This environmental intelligence comes from a wide variety of sources, including Doppler radar systems, weather satellites, buoy networks, water level stations, real-time weather stations, as well as ships and aircraft.
- Only a small percentage of this valuable data is easily accessible to the public.
- The demand for this data has increased, and it is imperative to find ways to effectively and efficiently distribute this data to decision makers and industries.



- NOAA's projected data growth is exponential
- By 2020, 160 petabytes of archived data
- Rate of data capture will continue to increase

Types of NOAA Data


Bathymetry & Global Relief


Climate Monitoring & Extremes


Coastal & Regional


Geomagnetism


Interactive Maps


Land-Based Stations


Marine Geophysics


Marine Surface


Models


Natural Hazards


Ocean Climatologies


Ocean Data


Paleoclimatology


Radar


Satellite


Space Weather

NOAA issued a Request for Information in 2014 to engage private industry to help make NOAA's data available in a rapid, scalable manner to the public.

American companies were asked to suggest ways for NOAA to more effectively distribute its data, allowing industry to take advantage of the untapped value of NOAA's public data resources by creating new and innovative products and services.

Secretary of Commerce Penny Pritzker announced on April 21, 2015 a Big Data Project to explore ways of bringing NOAA and the Department of Commerce closer to the goal of transforming Department data capabilities and supporting a data-driven economy.

NOAA's Big Data Project provides an innovative approach to publishing NOAA's growing data resources and positioning them near cost-efficient high performance computing, analytic, and storage services provided by the private sector. Hypotheses:

- There is additional value in NOAA data that has not yet been realized because of access and infrastructure difficulties;
- If NOAA data were accessible in the Cloud, alongside computing capability, private enterprise might generate new value-added products, services, and lines of business; and
- Private business might be willing to support the cost of transferring and storing large datasets because of these new lines of business.

The BDP Dissemination Model

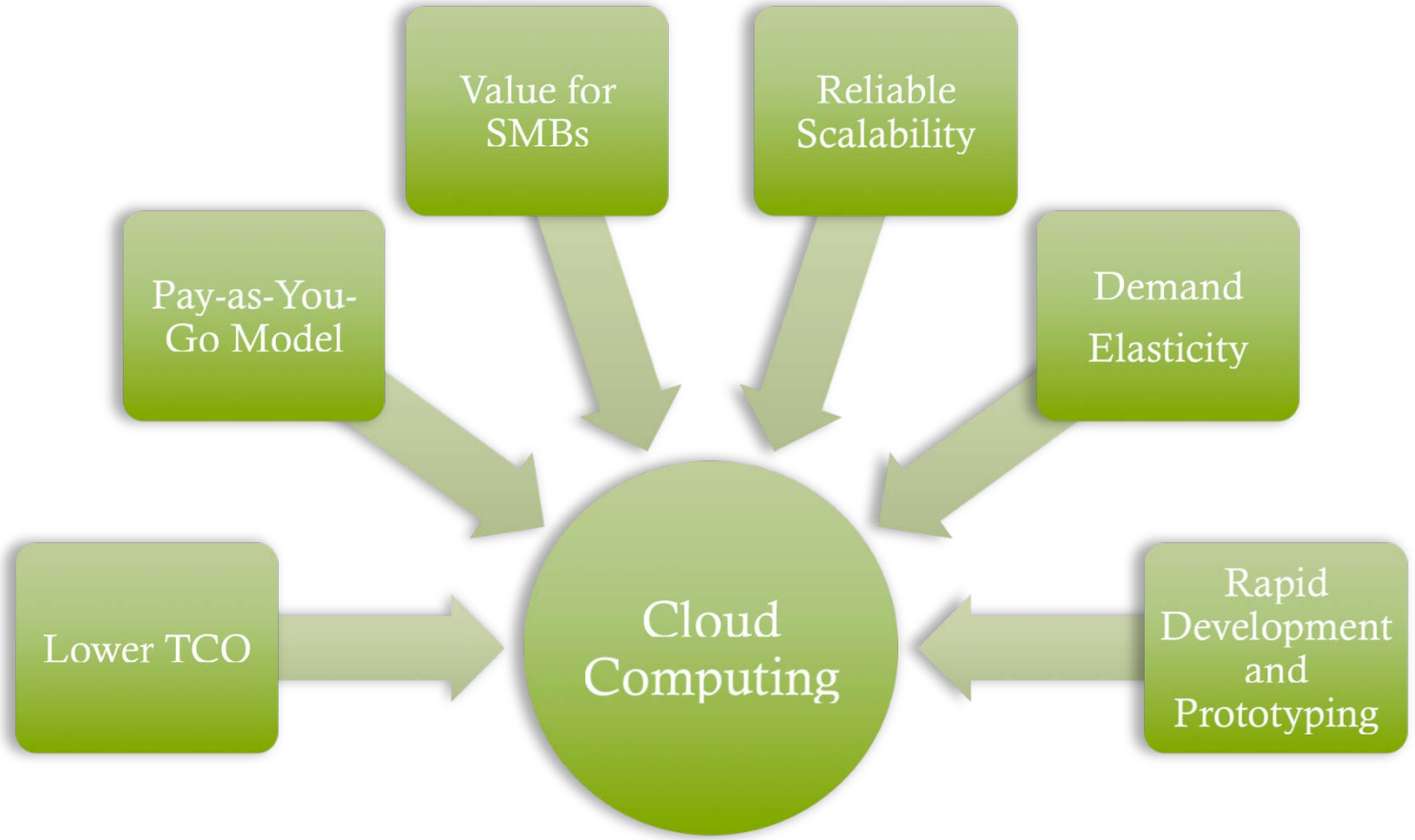
Traditional Data Request

Big Data Project

CRADA's:
Amazon Web Services, Google Cloud Platform, IBM, Microsoft, and the Open Cloud Consortium

- Existing NOAA infrastructure and funding cannot support growing data or demand
- CRADA a low-risk mechanism for scaled testing of the full market ecosystem
- Provides an iterative approach to dissemination without disrupting NOAA operations
- Allows for lessons learned and real-time modifications to the dissemination process
- Valid for three years with annual renewal option
- Collaborators and/or NOAA may choose to terminate with 30 days' notice
- Collaborators have non-exclusive access to NOAA data
- All NOAA data is up for discussion (excluding ITAR-restricted and NS sensitive data)
- Collaborators must provide users with equal access to NOAA data equal terms

Cloud Services



Cloud Collaborators

NOAA's cloud collaborators will provide the framework for a set of data alliances led by each of the anchor companies. Data alliances, which consist of participating organizations across the private and public sectors, will work to research and test solutions for bringing NOAA's vast information to the cloud, where both the public and industry can easily and equally access, explore, and create new products from it, fostering new ideas and spurring economic growth.

NOAA and its partners are starting small in order to minimize the collaborators' investment costs and they will use initial datasets to establish baselines and demonstrate the proof of concept. This will be a market driven process, whereby the collaborates reach back into NOAA for the datasets being most requested by their users. Except for those restricted due to national security concerns, all NOAA datasets are open for discussion on availability.

The collaborators will have equal access to NOAA's historical archive of large datasets and each must provide users with equal access to NOAA data on equal terms.

NOAA

- Ensure free and open access to all data, regardless of market interest;
- Provide authoratative data, metadata, information, forecasts, warnings, and analysis;
- Perform research to improve sensors, numerical models, and algorithms;
- Ensure long-term preservation of the data master copy;
- Perform scientific data stewardship as an unbiased, objective partner; and
- Provide expertise and skills to support proper use and application of data.

Pilot Data Set

NEXRAD is a network of 160 high- resolution Doppler weather radars that detect atmospheric precipitation and winds, which allow scientists to track and anticipate weather events, such as rain, ice pellets, snow, hail, and tornadoes.

Real-time feed and full historical archive of original resolution (LevelII) Next Generation Weather Radar (NEXRAD) data, from June 1991 to present has been released to Amazon, Microsoft, OCC, and Google.

Planned Data Sets

- Geostationary Operational Environmental Series satellite (GOES) – provides satellite information every 30 minutes using either the Visible, Infrared, Shortwave Infrared (4um), or Water Vapor images into one larger composite image using GOES East and West Imagers.

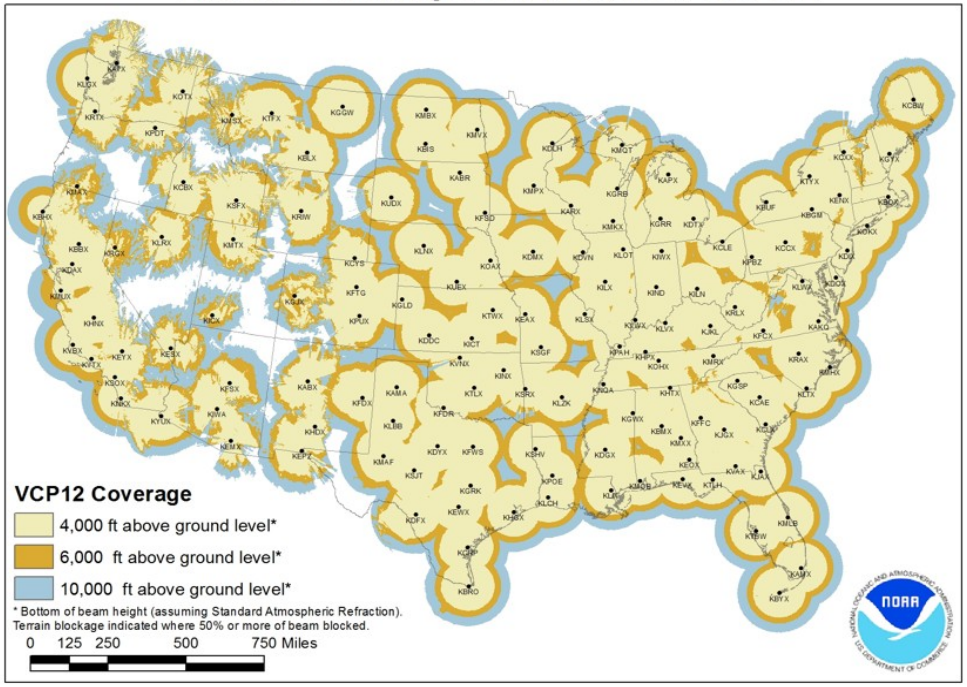
- Geostationary Operational Environmental Series satellite- R (GOES-R) – scheduled for launch in October 2016, the GOES-R will collect three times more data and provide four times better resolution and more than five times faster coverage than current satellites. This means the satellite will scan Earth's Western Hemisphere every five minutes and as often as every 30 seconds in areas where severe weather forms.

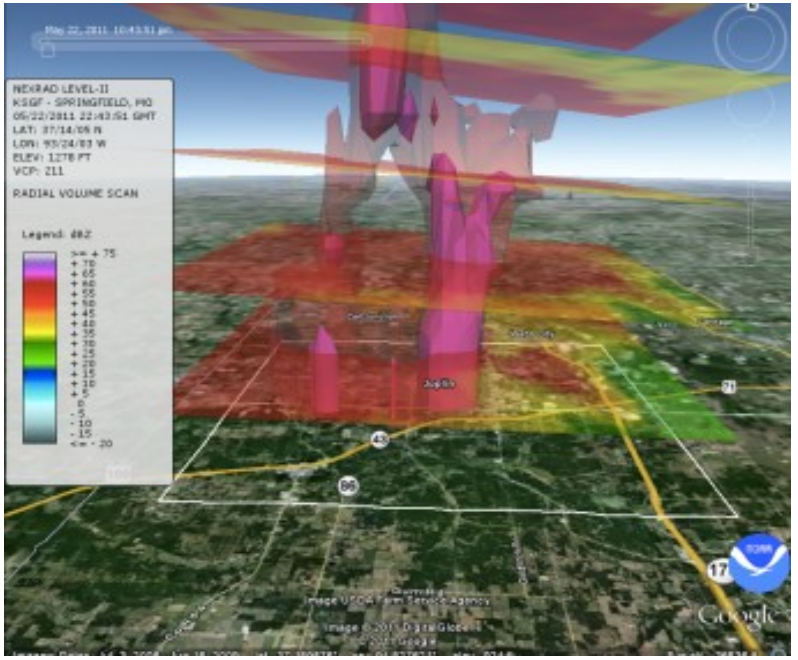
- Multi-Radar/Multi-Sensor System (MRMS) – quickly and intelligently integrates data streams from multiple radars, surface and upper air observations, lightning detection systems, and satellite and forecast models. Numerous two-dimensional multiple-sensor products offer assistance for hail, wind, tornado, quantitative precipitation estimation forecasts, convection, icing, and turbulence diagnosis.

NEXRAD Doppler Radar

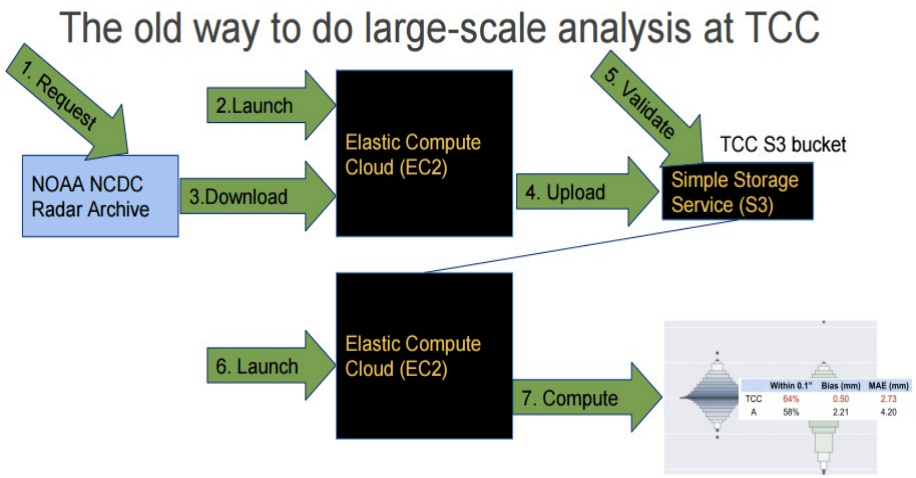
- Original resolution (Level II) available now
- Precipitation and atmospheric movement (wind)
- Tornado visualization
- Cold fronts
- Thunderstorm gusts
- Rainfall rate/hydrological forecasting

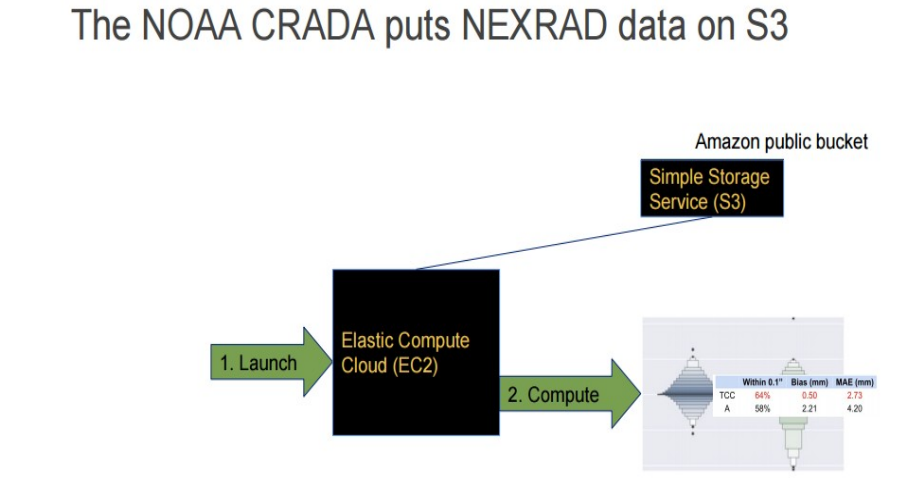
NEXRAD Coverage Below 10,000 Feet AGL



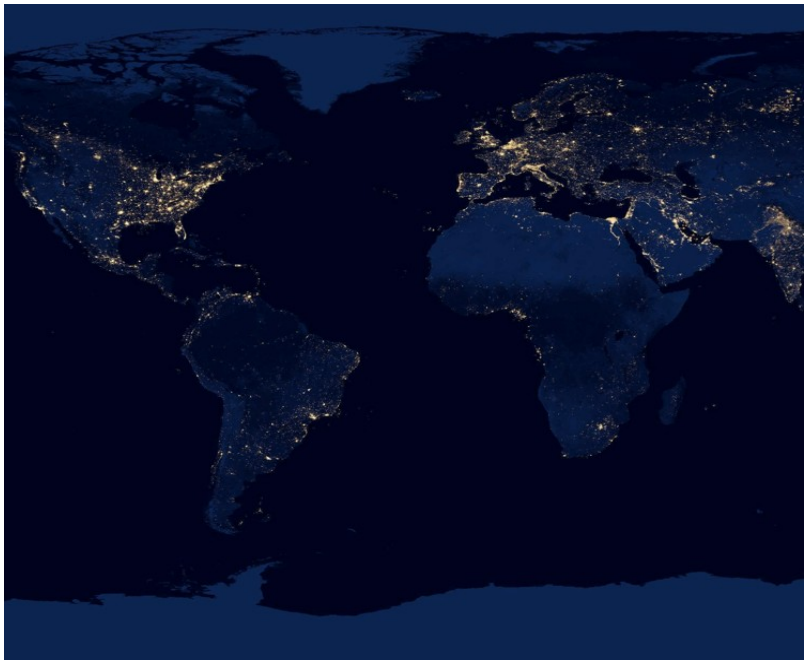


Reducing Operational Time at The Climate Corporation





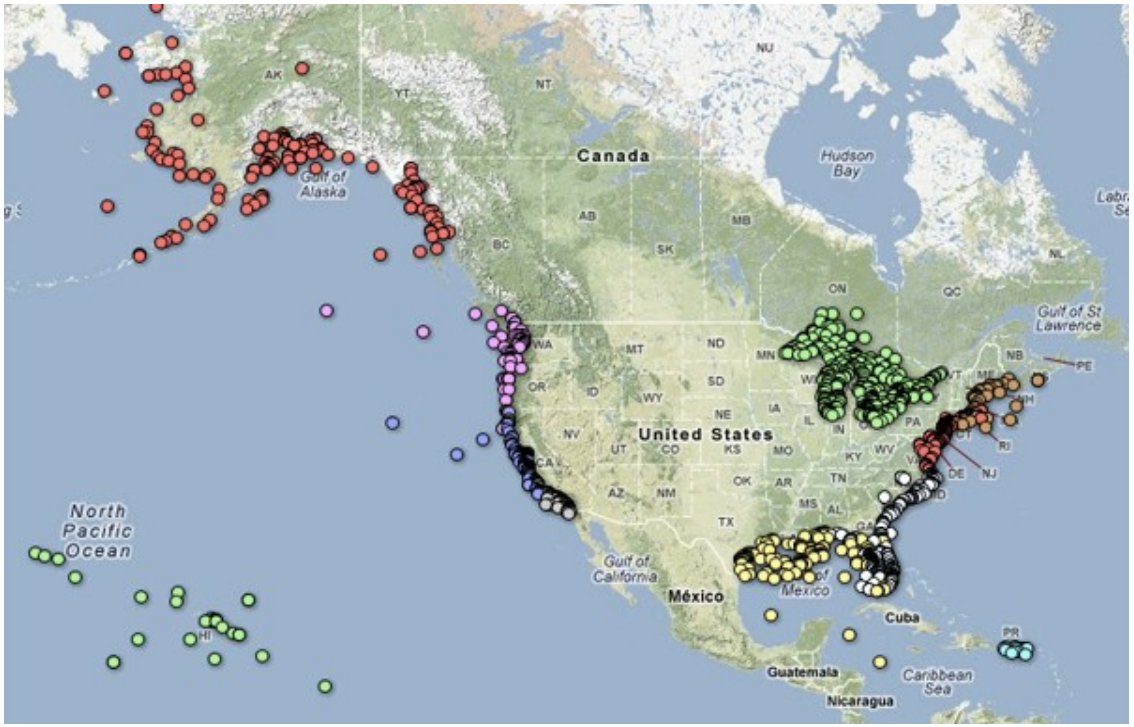
- **Project timelines at TCC are now several weeks shorter**
- **TCC can evaluate new forecasting models on larger datasets**
- **TCC reduces project cost; only pay for temporary cloud storage and compute cycle**
- **NOAA improves archive through resolution of data issues**



VIIRS Infrared Satellite

- Assists in long-term assessment of:
  - Emerging global storm patterns
  - Global economic activity
  - Arctic and ocean ice

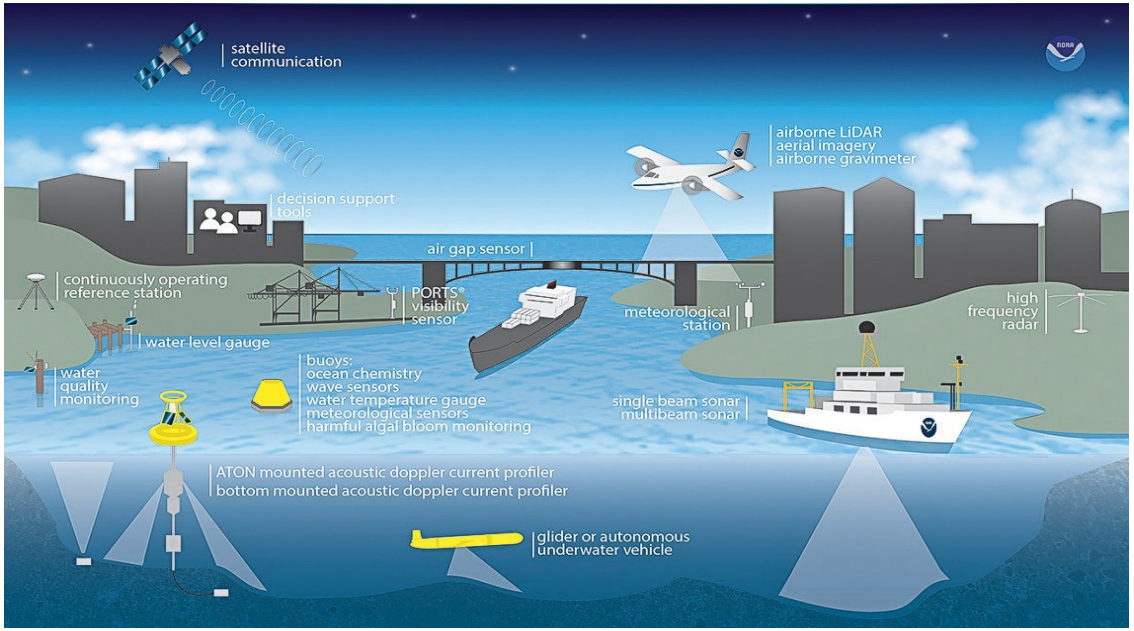
- Covers the entire globe every 14 hours
- Enables short-term monitoring of:
  - Wildfires
  - Drought
  - Vegetation
  - Ocean temperature



IOOS Ocean Observations

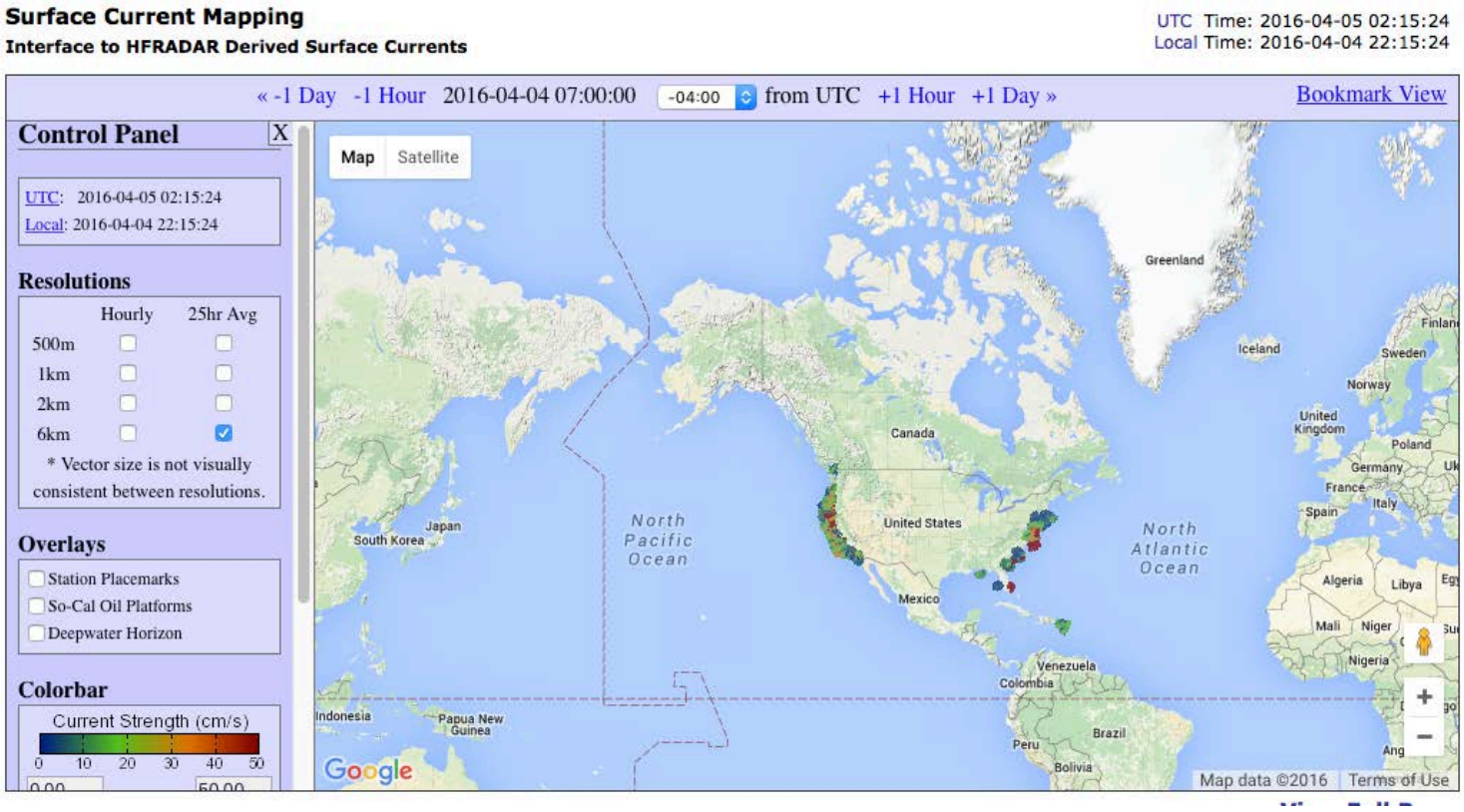
10,000+ oceanographic data sets

- Buoys
- High-frequency radar
- Water level gauges
- Gliders
- Animal telemetry
- Coastal and estuary stations



Recommendations

Explore interfacing HFRNet with CRADA partners as an ocean Pilot Data Set (similar to NEXRAD)



Data access is available in a number of formats and protocols:

Google Earth KML (7 day)
Mapping API (For off-site maps)
Data Access via CORDC THREDDS Server
Data Access via NDBC THREDDS Server
Data Access via FTP (3 day rolling archive)
ArcGIS Toolbox
Learn more about data access.

Amy Gaskins, Big Data Project Director

- **Email: Amy.Gaskins@noaa.gov**
- **Twitter: @AmyVGaskins**
- **LinkedIn: www.linkedin.com/in/amygaskins**