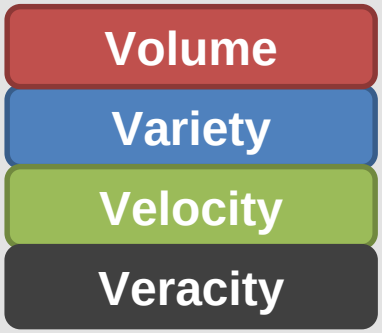
Key Challenges for Environmental Data and Information as viewed from NCEI

Thomas R. Karl, L.H.D
Director, NOAA's National Centers for Environmental Information
Chair, US Global Change Research Program

Department of Commerce Data Advisory Council
29-30 October 2015

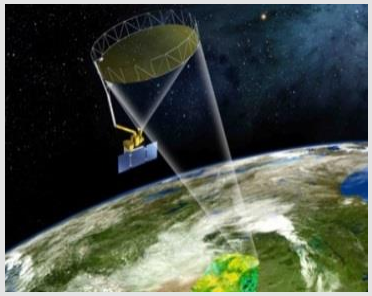
Key Challenges

- Supporting the Full Information Lifecycle
- Volume
  - Increasing Volume: Archival & Access
  - Rethinking compression
- Variety
  - Definition of a customer
  - Diversity of customers, needs, and products
  - Understanding customer requirements
- Velocity
  - Preventing bottlenecks
  - Cloud challenges
  - Future vision
- Veracity
  - Reference Environmental Data Records
  - Tiers of data stewardship
  - Integration, Harmonization & Environmental Intelligence

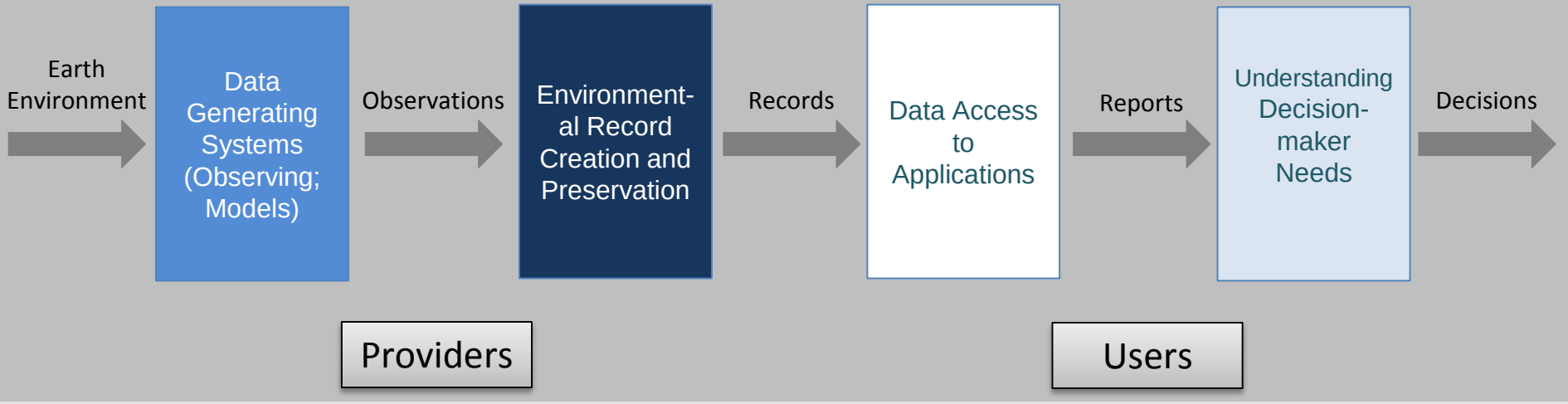


Supporting the Full Information Lifecycle

Volume
Variety
Velocity
Veracity



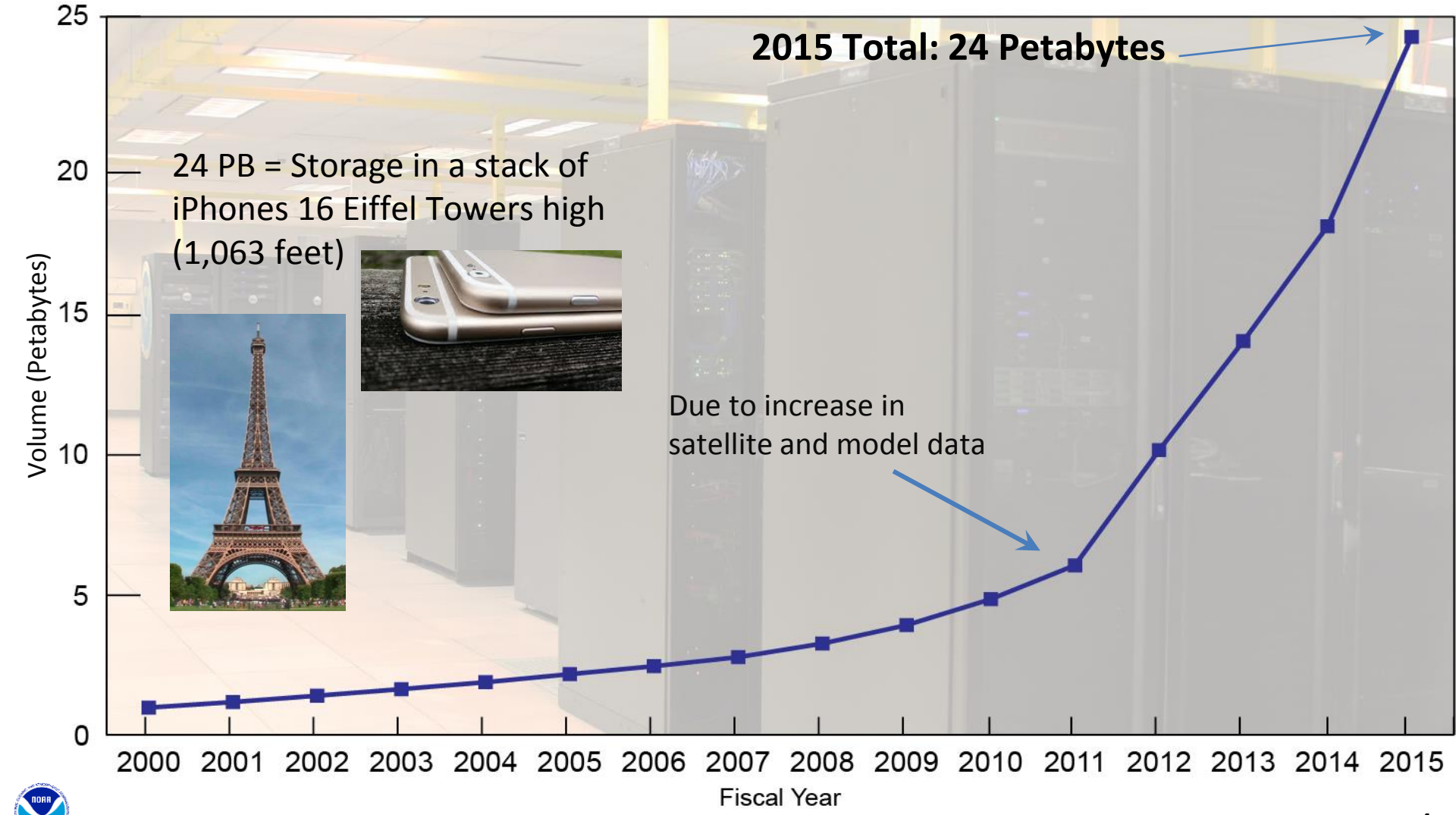
- NCEI makes foundational investments in environmental information production and preservation
- NCEI supports others' application development and policy/decision-making



Darker shading represents higher levels of effort

Increasing Archive Data Volumes from Station, Model, Radar, and Satellite Sources

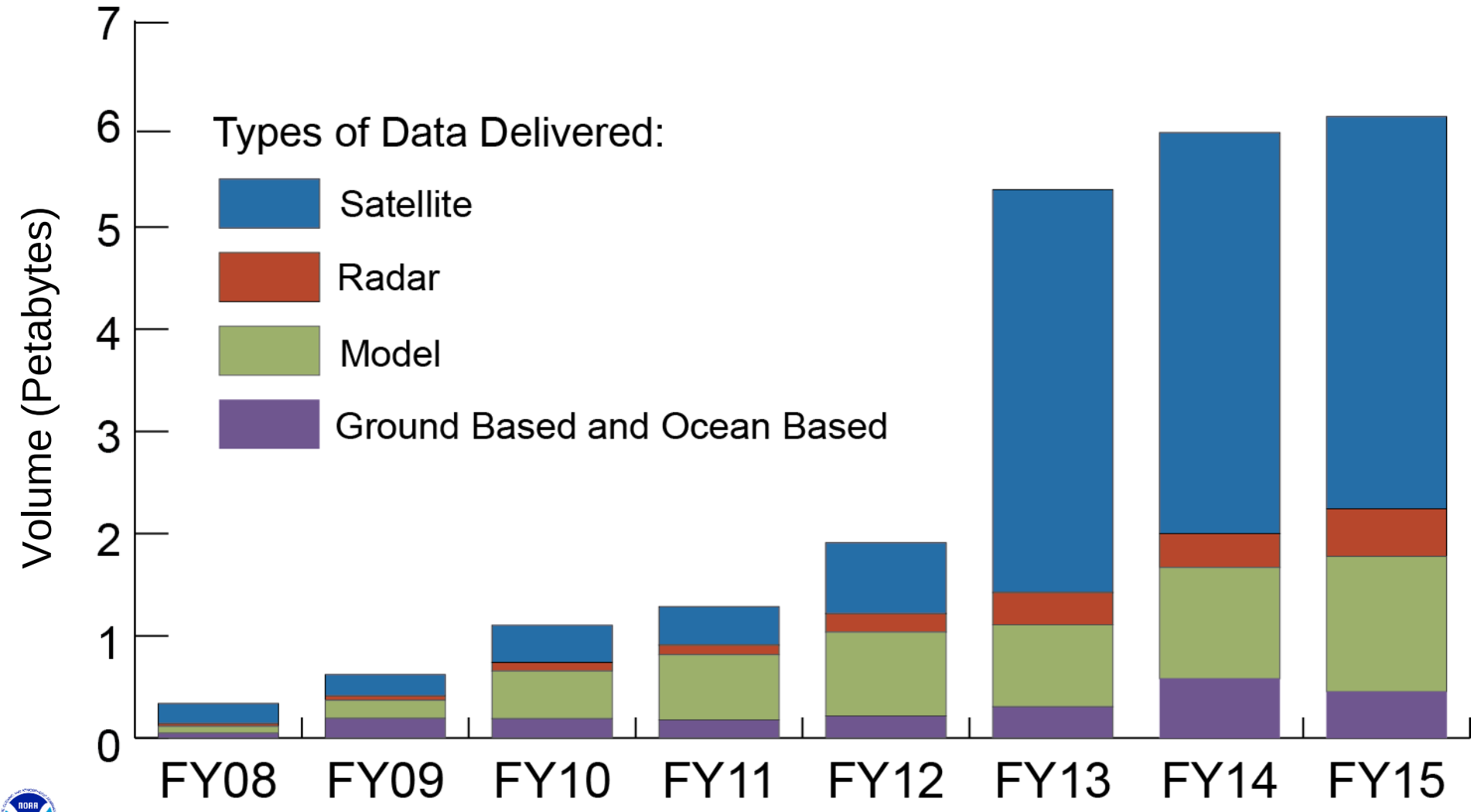
Volume
Variety
Velocity
Veracity



User-Requested Access

Increasing Data Requests for All Data Types

Volume
Variety
Velocity
Veracity



Increasing Volume

Volume
Variety
Velocity
Veracity

- Continued increases in customer demand for access and archive
  - U.S. Open Data Policy
  - NOAA plan for Public Access to Research Results (PARR)
  - Climate Data Initiative
  - Big Earth Data Initiative
- New data from around NOAA
  - Integrated data products
    - •Integrate across disciplines, e.g., coastal inundation
    - •Integrate across sensors, e.g., blended precipitation (radar, satellite, in-situ)
  - Increased volumes from new sensor technologies (e.g., JPSS and GOES-R), models, video, etc.
- Need to rethink compression
  - Lossless versus Noiseless
  - More than an IT challenge
    - •Radar data 16:1 → ?:1
    - •Satellite data 2:1 → ?:1
    - •Model data ?:1

Definition of "Customer"

Volume
Variety
Velocity
Veracity

- Customers: those to whom NCEI provides products and services
- Three types of customers
  - Providers
    - Provides archival and data stewardship services
  - Provider/User
    - Provides both archive and access services
  - Users
    - Provide data and information products
      - Range of space & time scales
      - Range of contextual expertise

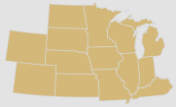
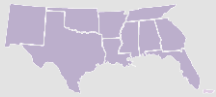
Understanding Diverse Customer Needs

Volume
Variety
Velocity
Veracity

Economic Sectors

Agriculture
Forestry
Fishing
Water
Energy
Utilities
NGO
Transportation
Academia
State, Local, Tribal Gov.
Recreation
Federal Government
Construction
Healthcare
Reinsurance

Regional

West

Central

South

Northeast

Alaska

Pacific Islands

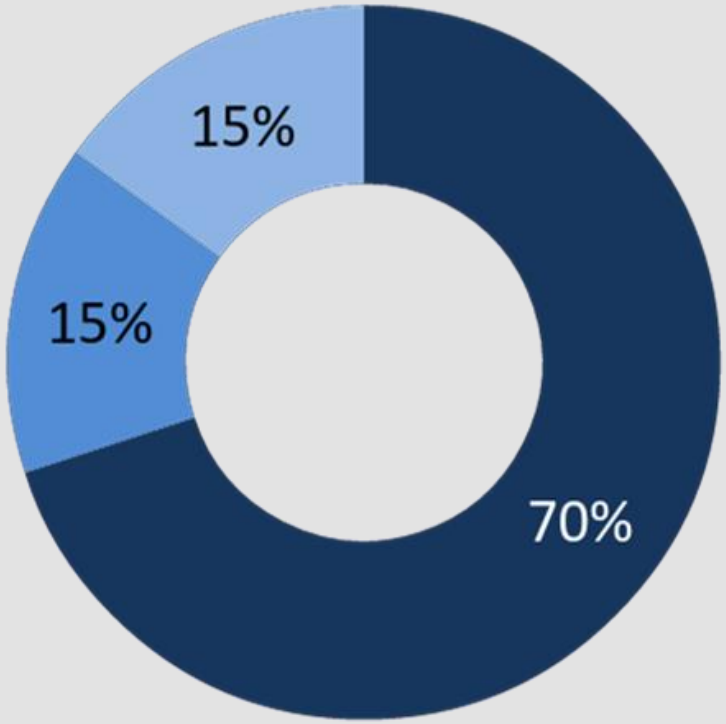
Societal Challenges

Marine
Coastal Resilience
Water Resources
Extremes
Health

Understanding Diverse Customer Expertise and Access Needs

Volume
Variety
Velocity
Veracity

Typical User	Data or Info Need	Preferred Format	Access Volume	Access Frequency
Business, media, public	Qualitative	Point-and-click, graphics, assessments	Low	High
Researchers, Climate consultancies	Quantitative	Digital downloads	High	Low
Value-added Providers (database scrapers)	Quantitative	Digital downloads	Low	High



Responding with a Diverse Set of Products

Volume
Variety
Velocity
Veracity

	Daily – Weekly	Monthly	Seasonal – Annual	Annual – Decadal
Local	Snowfall Impact Index – FEMA, disaster response	Heating & Cooling Degree Days – Energy Sector	Temperature & Precipitation Outlooks – Agriculture	Coastal Digital Elevation Models (DEM) – Hazard Mitigation
Regional	Hurricane Tracks – Emergency Planners	Solar Activity/Sun Spots – Power Distribution	Billion $ Disasters, Climate Extremes Index – Insurance	Climate Normals – Construction, Infrastructure, Agriculture
National & Global	Tsunami Warning – Emergency Managers	Global and U.S. Climate Summaries – Numerous Sectors	STATE OF THE CLIMATE IN 2013 Annual State of the Climate Reports – Scientists	IPCC & National Climate Assessments – Gov't Policymakers

Understanding Producer and User Requirements

Volume
Variety
Velocity
Veracity

- Direct Contact – customer support, questions, comments
  - Ad hoc, but often specific feedback received on important requirements
- Requests for Information – posted formally
  - Not well targeted to current and prospective users
  - Very low response, mostly vendors sending proposals
- Online – user registration & surveys
  - Only connects with existing users
  - Not a high response rate to surveys
  - Cumbersome regulations regarding surveying the public
- User Workshops – in-person events bringing together scientists, archivists, and users
  - Sometimes successful due to robust in-person interaction
  - Expensive to run, and hard to scale
- Executive Forums – invited business leaders
  - Can be helpful but considerable investment costs

Producer and User Requirements: Lessons Learned

Volume
Variety
Velocity
Veracity

- Private sector is very interested in our future plans
  - Need to know how to plan their products
  - Best to reach out as broadly and deeply as possible
- One size doesn't fit all
  - Regional differences in needs: coastal inundation, drought, heat…
  - Sectoral differences in needs: academic modelers, energy industry, water managers…
- Importance of standards and frameworks
  - Tiers of stewardship, cost model, Open Archival Information System reference model
- Desirability of a third party to facilitate connection
  - Partnership for Resilience is a research consortium currently being created to address this need
  - Challenges yet to be overcome include allocating intellectual property, collaborative projects, technology transfer, sharing best practices

Velocity: Preventing Archival Bottlenecks

Volume
Variety
Velocity
Veracity

- High I/O demand reprocessing and in place product processing
- Affording long-term archival storage while meeting velocity requirements
  - Robotic tape: Cheap, slow, secure
  - Spinning disk: Expensive, fast, secure
  - Cloud: TBD
  - Federated vs Centralized solutions

- Re-architecting to maximize cost, performance, and security
  - Meeting requirements of Federal Cloud Computing Strategy
  - Decoupling archive, access
- Transport
  - E.g., moving model data 2-5 TB/week, for over 860 TB of data

Preventing Access bottlenecks

Volume
Variety
Velocity
Veracity

PRIMARY DELIVERY SYSTEMS TODAY

- CLASS: heavy on satellite subscription-based services
  - Contains satellite data, and is incorporating non-satellite data
  - New users often require extensive support from NCEI to understand how to obtain the data they need
- Environmental Data Online: for non-satellite data
  - Delivery usually within 24 hours (typically 15 minutes or less)
  - Most effective for existing users
  - New users often require extensive support from NCEI to understand how to obtain the data they need
  - Aging system, and many subcomponents

Challenges of Cloud Access

Volume
Variety
Velocity
Veracity

- Enabling cloud dissemination while maintaining authoritative archive
  - Designing security into system
  - Supporting digital signatures to allow confirmation of data integrity prior to use
  - Routinely updating cloud copy to be consistent with the authoritative copy
- Cloud access expected to increase application of data, thereby increasing demand for good stewardship
  - Ensuring transparent data stewardship
  - Documenting the data for reuse beyond its original intended application
  - Ensuring data understandability
  - Improving data maturity through iterative scientific and preservation improvements
  - Generating reference environmental data records
  - Building trust with the data-using communities through ongoing engagement and responsiveness to their needs

Vision to Address Volume and Velocity Challenges: NOAA Environmental Data Stewardship System

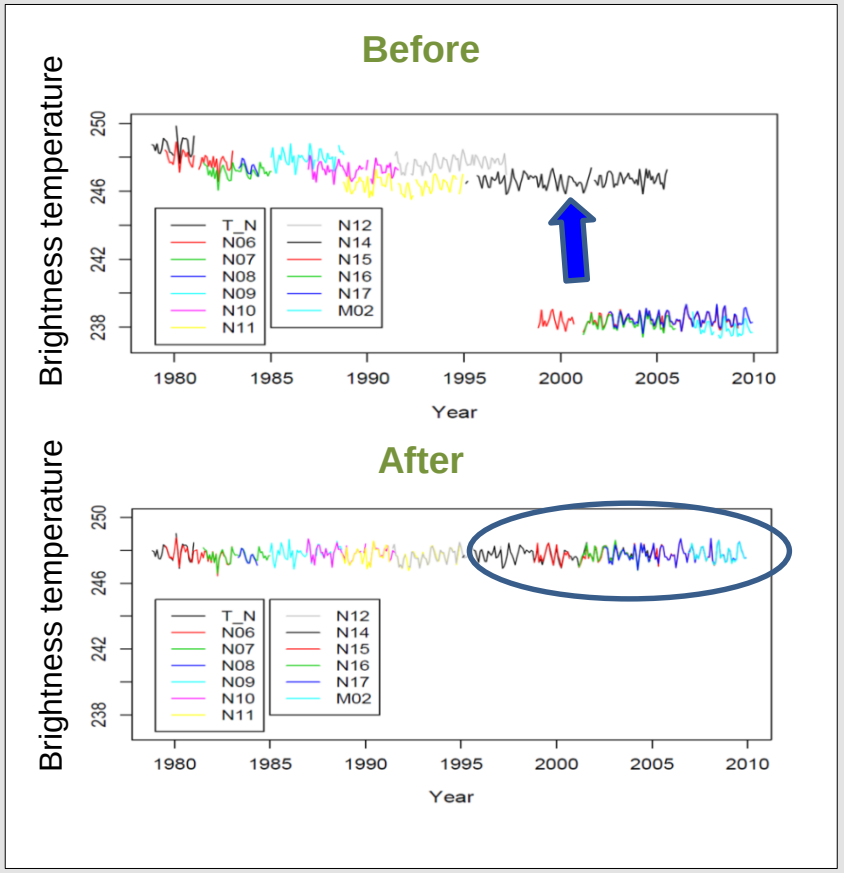
Volume
Variety
Velocity
Veracity

- Potential Solutions
  1. Integrate core archive systems (former Data Centers) and link to other large-volume storage capabilities around NOAA to form a federated archive;
  2. Leverage the Big Data Partnership to exploit scalable storage and processing in the public cloud for access and reprocessing; and
  3. Develop new data management tools and services to support the entire data lifecycle, from planning to archive and re-use
- Improve discovery, access, and usability of NOAA data through improvements in
  - User Interface
  - Catalog Services
  - Metadata
  - Search
  - Data management capabilities

Providing Reference Environmental Data Records

Volume
Variety
Velocity
Veracity

- Reference Environmental Data Records (formerly Climate Data Records – CDRs) help to minimize risk and maximize the opportunities for informed decision making
  - Quantifying and communicating uncertainties
  - Maintaining openness and transparency
  - Ensuring records are easy to understand and use

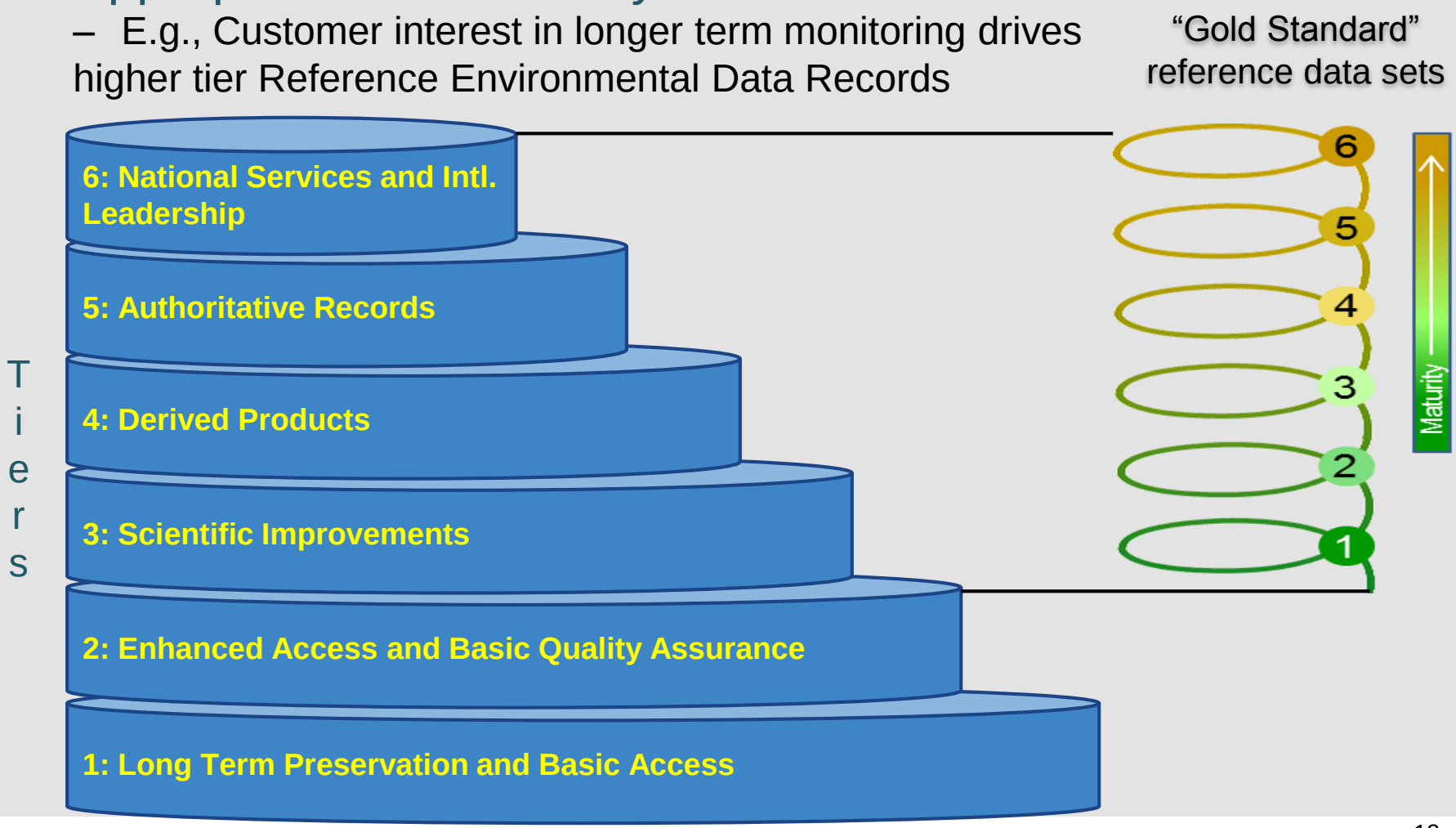


Tiers of Data Stewardship

A common language for dialog

Volume
Variety
Velocity
Veracity

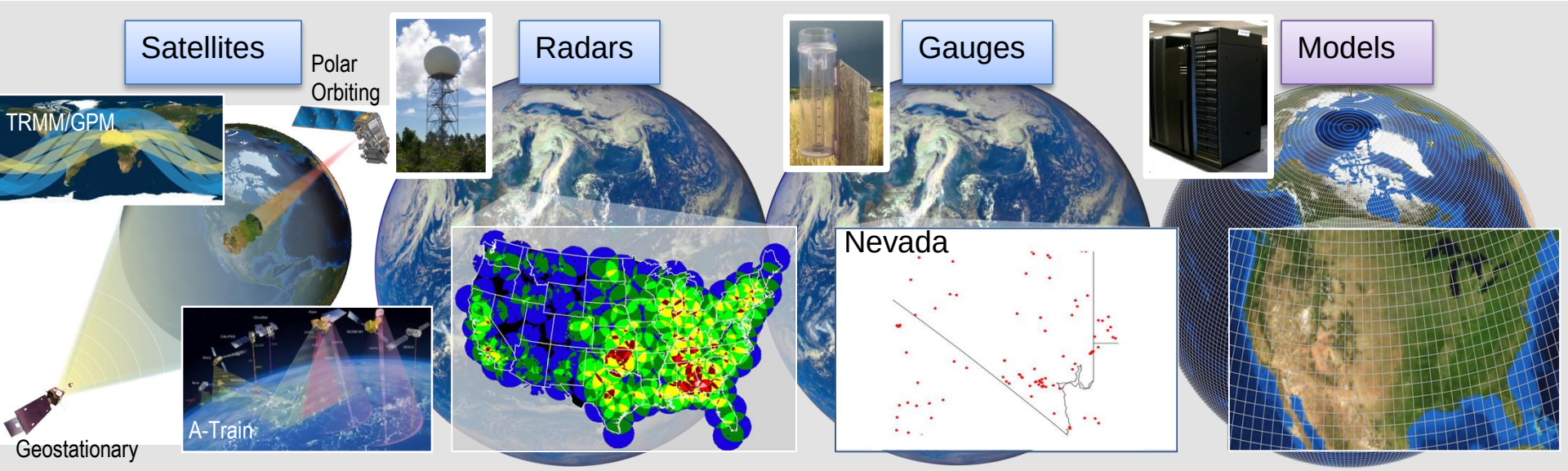
- Appropriate tier driven by customer needs
  - E.g., Customer interest in longer term monitoring drives higher tier Reference Environmental Data Records



Example: The Process of Integration and Harmonization

Data Harmonization: The blending of observations from different observing systems to improve accuracy, reduce bias, and increase resolution or coverage

An illustrative example: How does precipitation change over time at a given location?
Data from individual observing platforms and models all have unique strengths and weaknesses



- Global coverage
- Indirect measurement
- Biases: Viewing angles, orbital changes, hydrometeor type and size distribution

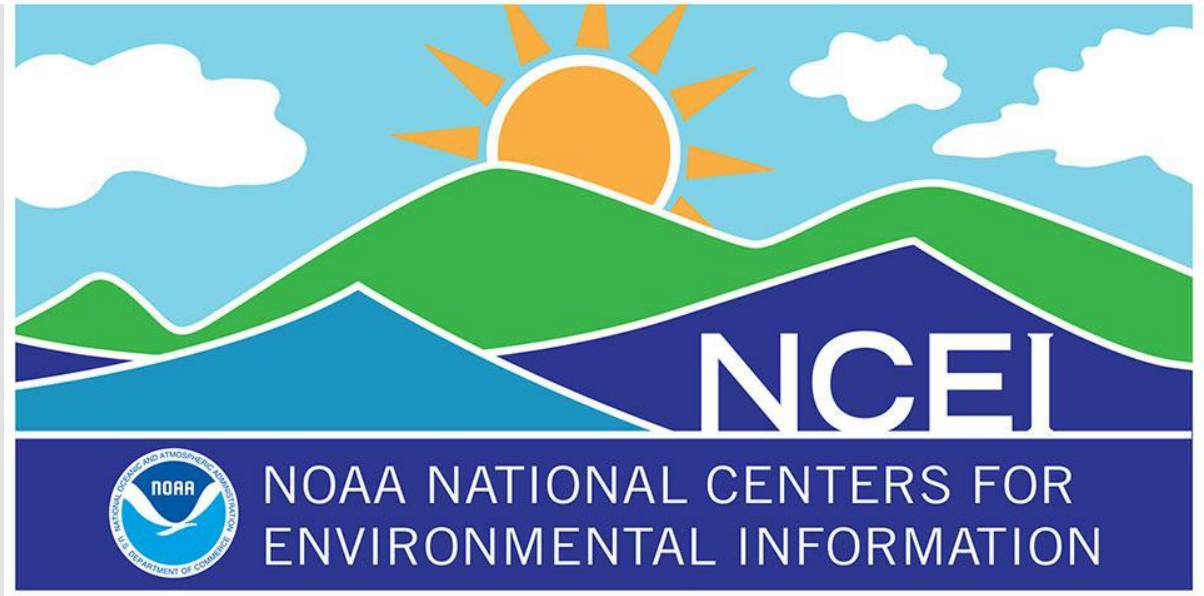
- High spatial resolution
- Indirect measurement
- Biases: Signal strength attenuation, operating methods

- Direct measurement of precipitation
- Long history
- Representativeness: a single point
- Biases: winds and other factors

- Global coverage
- Directly linked to physics/chemistry/ thermodynamics
- Interpretation affected by complex terrain and other factors

Collectively, harmonized data from multiple observing/modeling systems provide the most reliable results

Thank you!



www.ncei.noaa.gov

NCEI Climate Facebook: http://www.facebook.com/NOAANCEIclimate
NCEI Ocean & Geophysics Facebook: http://www.facebook.com/NOAANCEIoceangeo
NCEI Climate Twitter (@NOAANCEIclimate): http://www.twitter.com/NOAANCEIclimate
NCEI Ocean & Geophysics Twitter (@NOAANCEIocngeo): http://www.twitter.com/NOAANCEIocngeo