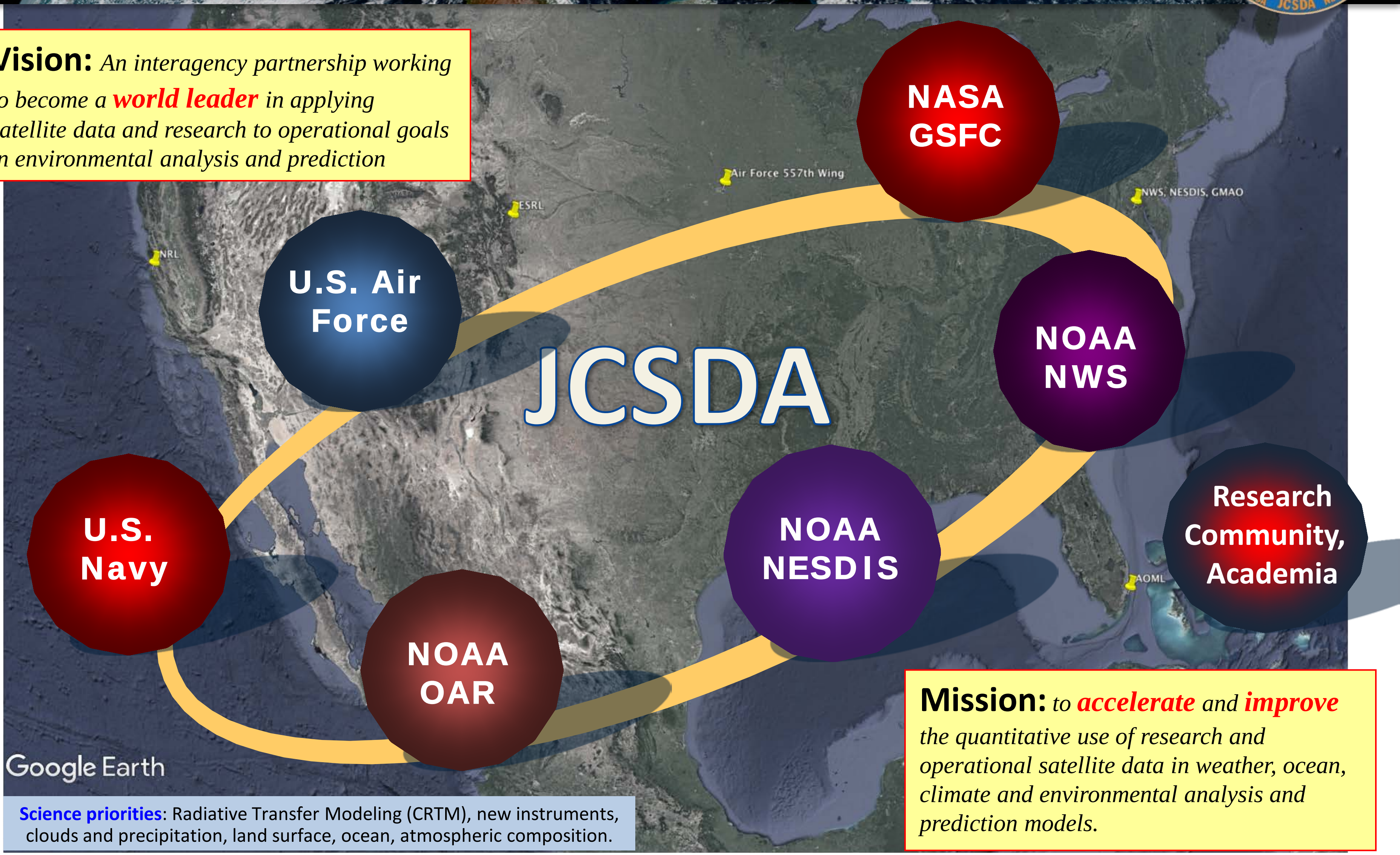
Cloud- Efforts at the Joint Center for Satellite Data Assimilation

Tom Auligné - Director, Joint Center for Satellite Data Assimilation (JCSDA)

Joint Center for Satellite Data Assimilation

Vision: *An interagency partnership working to become a* ***world leader*** *in applying satellite data and research to operational goals in environmental analysis and prediction*



Science priorities: Radiative Transfer Modeling (CRTM), new instruments, clouds and precipitation, land surface, ocean, atmospheric composition.

Mission: *to* ***accelerate*** *and* ***improve*** *the quantitative use of research and operational satellite data in weather, ocean, climate and environmental analysis and prediction models.*

Introduction

Disclaimer

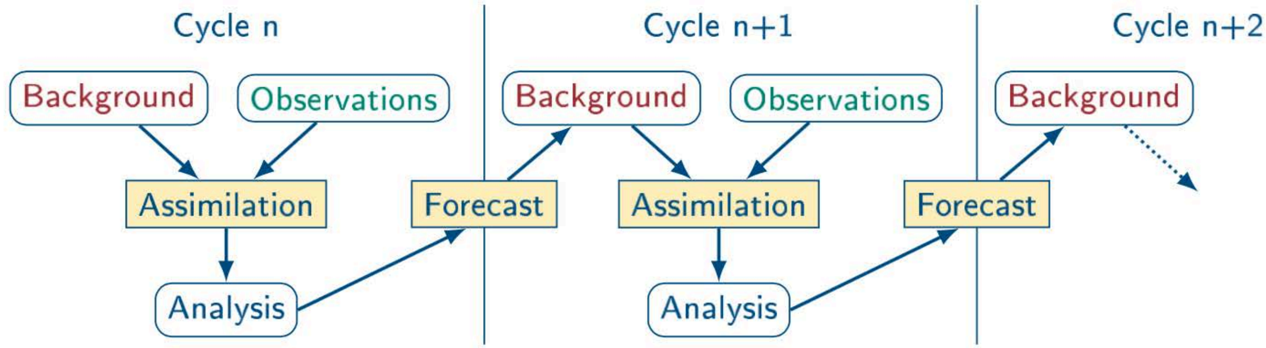
- Present exploratory effort and plans @JCSDA. No lessons learned yet

Outline

- Quick introduction to Data Assimilation
- Ongoing work for Joint Effort for Data assimilation Integration (JEDI)
- Challenges and opportunities relevant to cloud computing

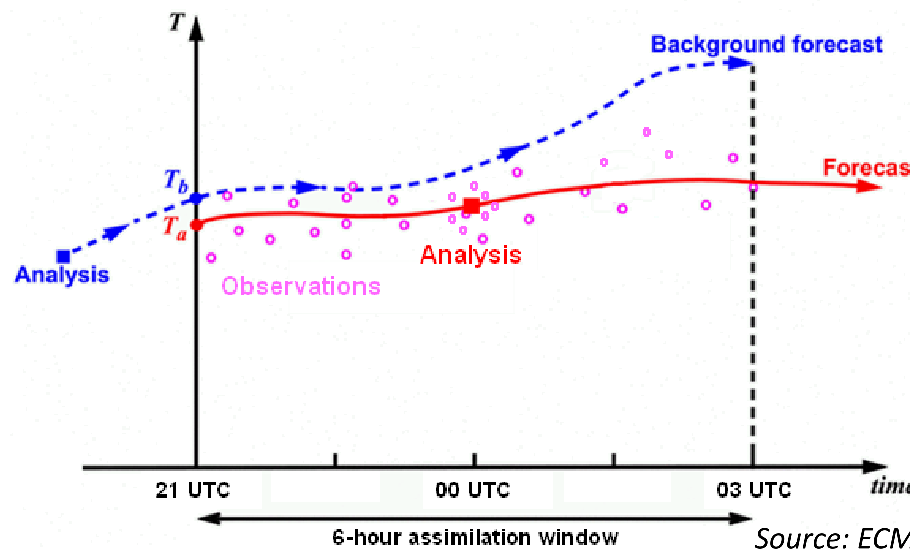
Introduction to Data Assimilation

Data assimilation systems usually combine together information from a set of observations, a short term forecast, and possibly other information to estimate the most probable state of a physical system.



- **Observations** provide information about “reality” but are disparate and irregular in space and time

- **Models** provide regular, physically consistent information about the system, but are prone to systematic errors



Source: ECMWF

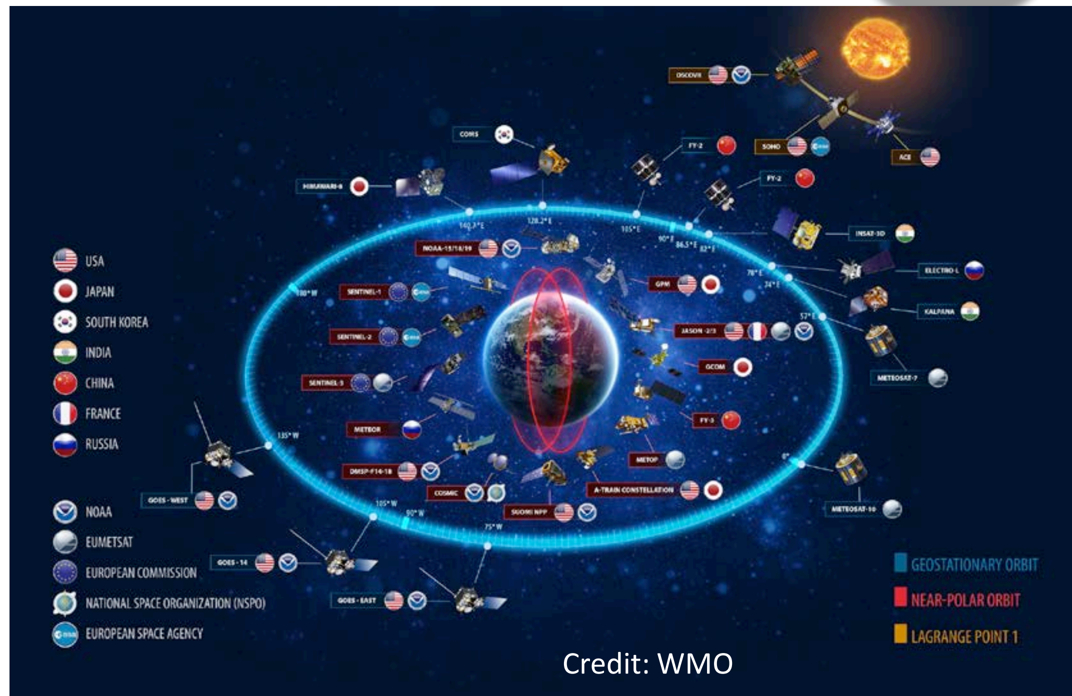
Eye Candy worth 1000 equations

Source: Will McCarty (NASA/GMAO)

Challenges in Earth System DA

Observations

- Big Data paradigm (volume, variety, velocity): most of forecast error reduction comes from large number of observations with **small or moderate individual impact**



Credit: WMO

Models

- Better value for society: forecast model for more components of Earth system
- Models are getting coupled to better account for interactions

Data Assimilation

- Data Assimilation (DA) systems becoming increasingly complex as science progresses: comparing all algorithms almost impossible

Joint Effort for Data assimilation Integration (JEDI)

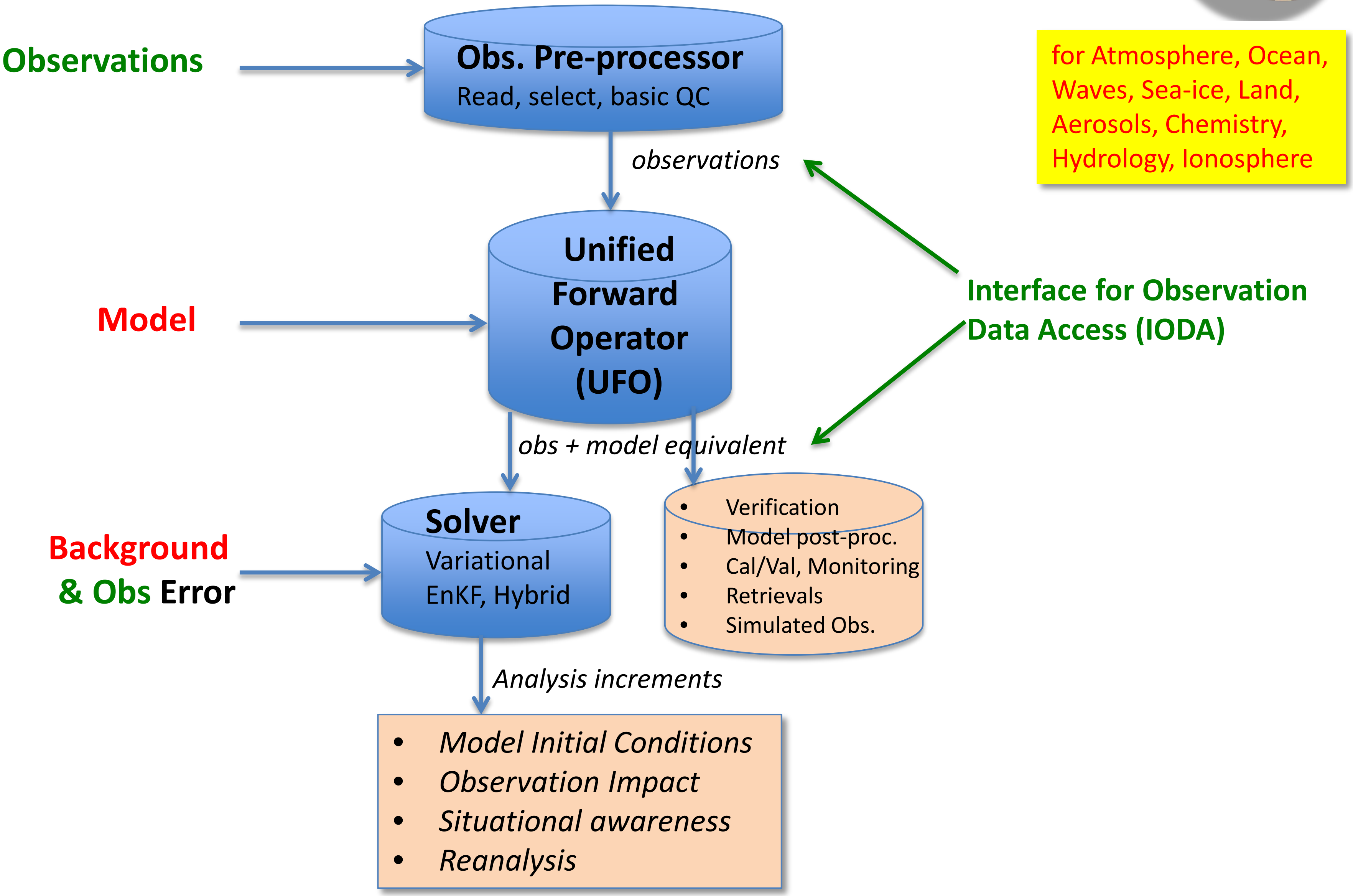
OBJECTIVES

1. Facilitate **innovation** to address next scientific grand challenges
2. Increase **R2O** transition rate
3. Increase **science productivity** and code **performance**

STRATEGY

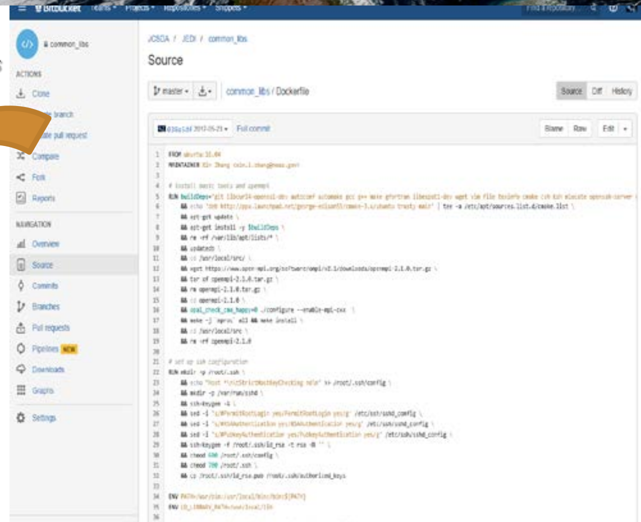
1. Collective path toward Nation Unified Next-Generation Data Assimilation
2. Modular, Object-Oriented code for flexibility, robustness and optimization
3. Mutualize **model-agnostic** components across
   - Applications (atmosphere, ocean, land, aerosols, etc.)
   - Models & Grids (regional/global, FV3, MPAS)
   - Observations (past, current and future)

Data Assimilation Components

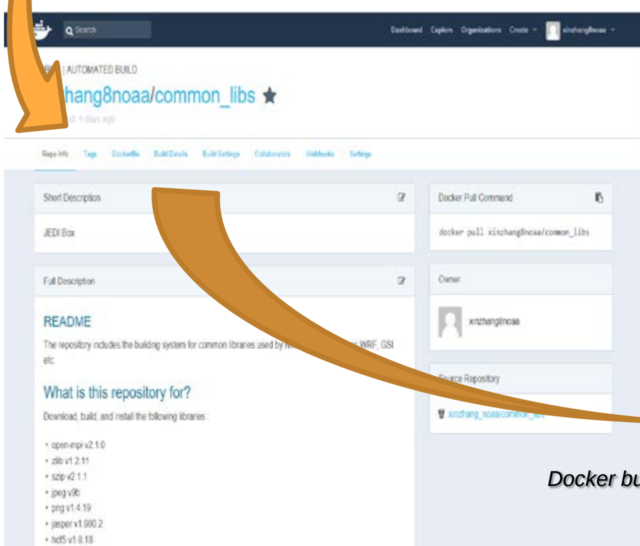


Bitbucket + Docker for JEDI

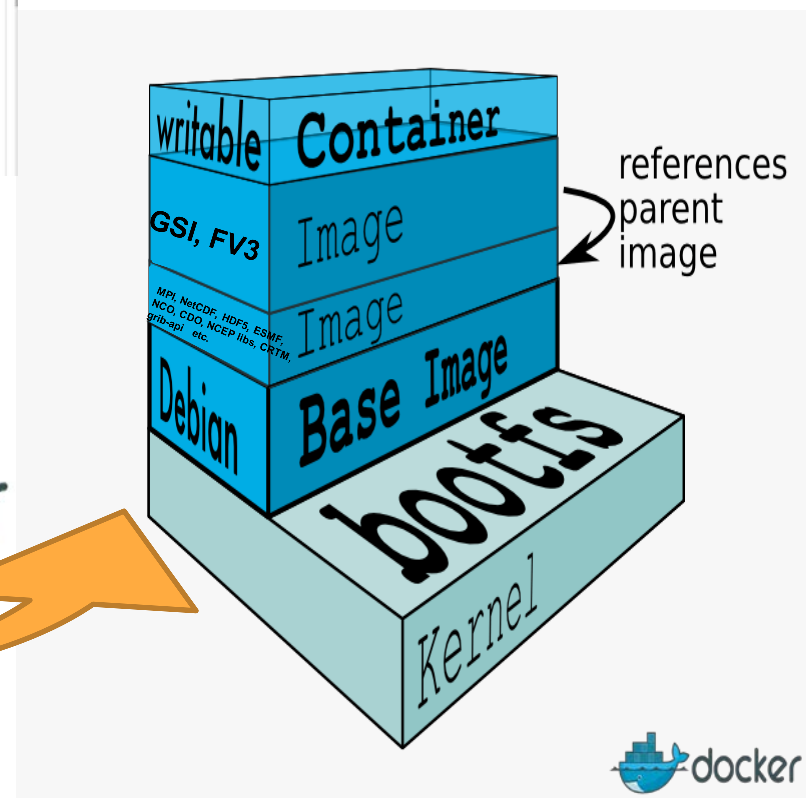
Bitbucket version-controls the codes, scripts



Docker pulls the codes and scripts from Bitbucket

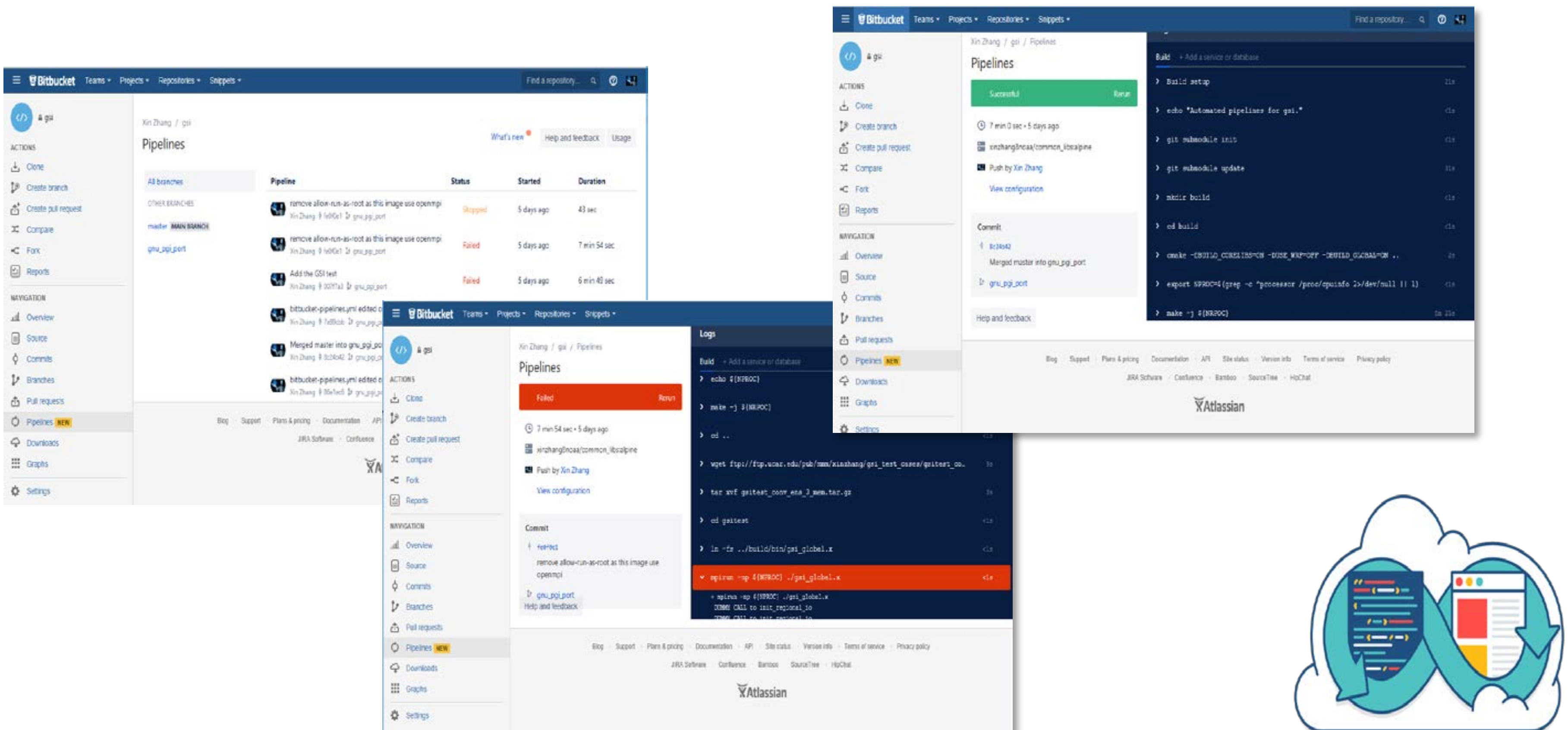


Docker builds and stores the JEDI image

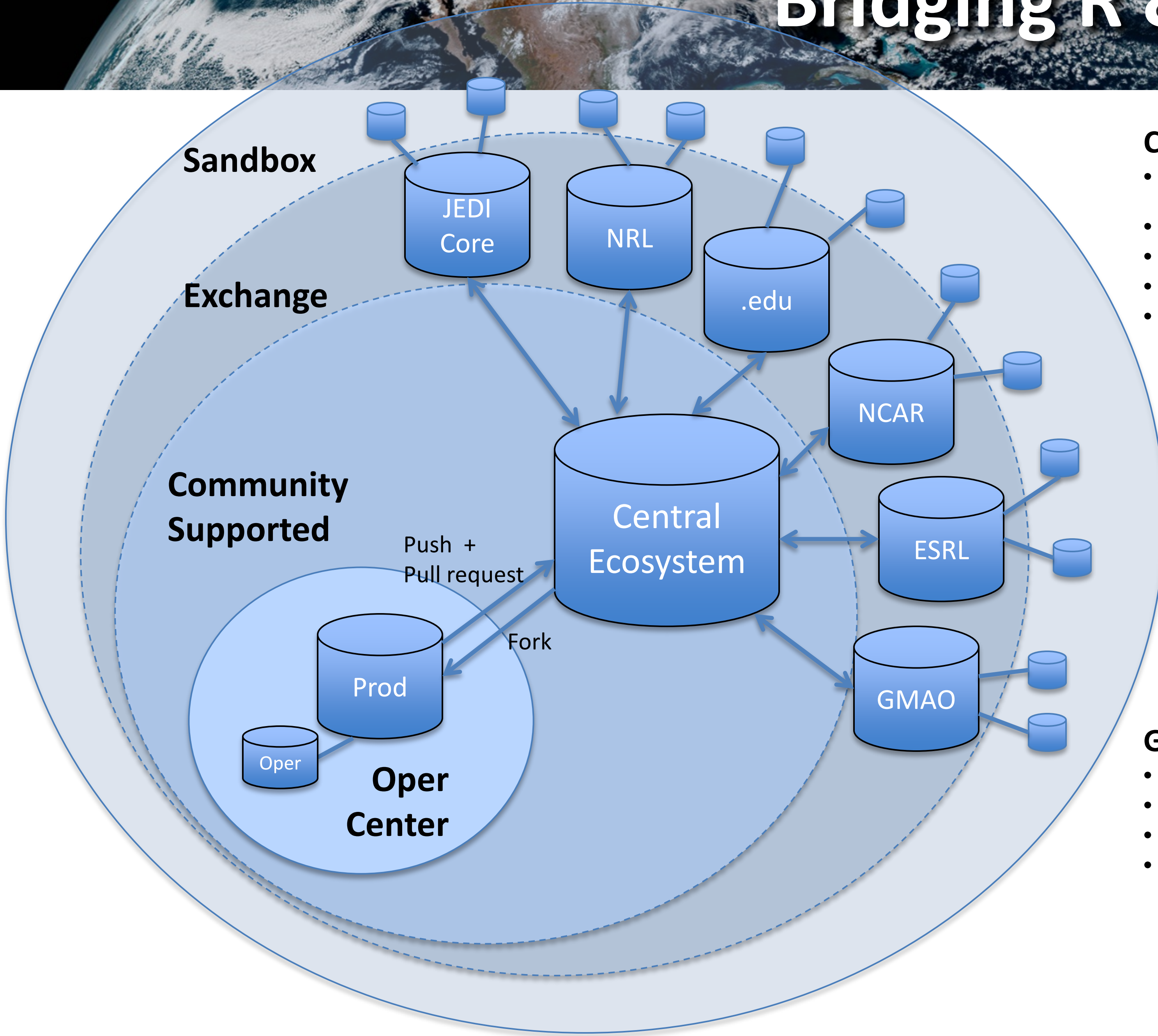


Using Pipelines

- Pull the Docker container image for JEDI
- Download the controllable test cases
- Automatically building and testing on Cloud with every commit
- Bitbucket Pipelines can send notifications to your team's chat room, and also via email



Bridging R & O



Collaborative ecosystem

- Code repository & reviews *(GitHub, Bitbucket)*
- Issue tracking *(JIRA)*
- Automatic tests *(Pipelines, Docker)*
- Documentation *(Confluence)*
- Support *(JIRA Helpdesk)*

Governance

- Specify roles + authorities
- Define interfaces
- Identify code utility
- Allocate resources

Cloud Computing for Operations

- Not widely used because
  - Production schedules
    - Tight operational timing constraints (#1 requirement)
    - Codes often optimized for single machine architecture (portability issue)
  - Cost: dedicated machine cheaper when almost 100% usage
  - Inter-processor communications (large jobs ~200 nodes x 24 cores)
  - I/O and data transfer
    - Large volume of input/output data analyzed and stored in house
  - Community *mainframe* mindset
    - Security concerns, egos (Top-500 HPCs with >100,000 cores @5% peak)

- Potential
  - Increased reliability (less single points of failure)
  - Cost: reduced need to buy redundancies, more regular expense

Cloud Computing for Development

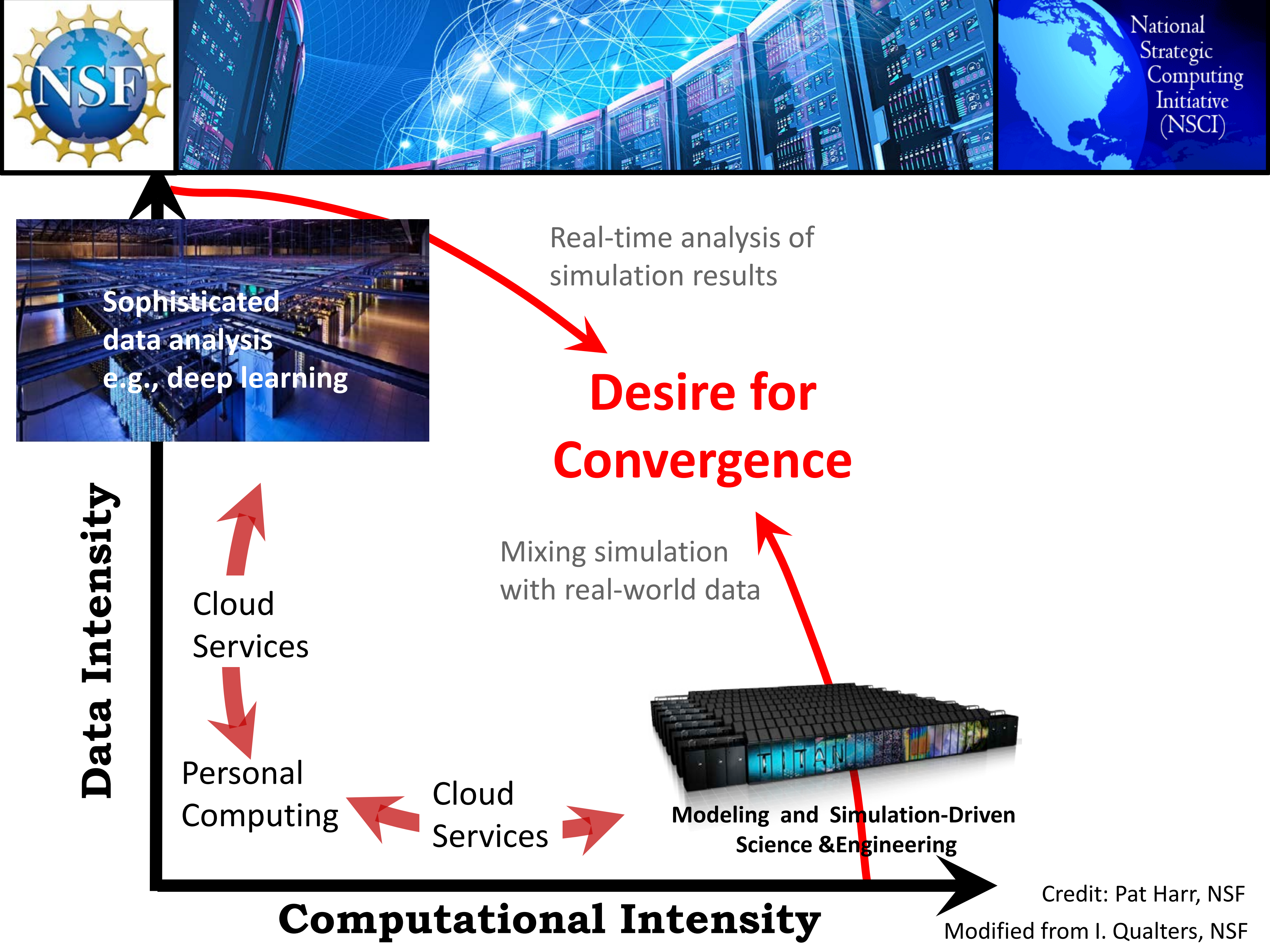
- Not widely used because
  - Cost
    - Every incremental science improvement (~ 1-2% in skill) requires huge computing resources for evaluation beyond chaotic nature
    - Projects and/or scientists often get free allocations on large research machines already "pre-paid", often as backup for operations
  - Easy R2O transitions: reproduce operations (and share infrastructure)

- Potential
  - More flexible for irregular computational load
  - Current providers of computing resources for research could provide "cloud-like" interfaces for access and tools

Cloud Computing for Research

- Not widely used because
  - Scientists fear not being able to log in and hack experiments
  - The *technically-aware* fear losing their ability to jump the queues
  - Technical debt
    - Large legacy code, lack of software engineers to modernize
    - Manpower cost of poor code infrastructure not properly considered

- Potential
  - Cost for low intensity use
  - Test portability and explore potential of new hardware
  - Ease of access (diversity and inclusion)
  - Facilitator for community collaborations


NSF
National Strategic Computing Initiative (NSCI)
Sophisticated data analysis e.g., deep learning
Real-time analysis of simulation results
Desire for Convergence
Mixing simulation with real-world data
Cloud Services
Personal Computing
Cloud Services
TITAN
Modeling and Simulation-Driven Science &Engineering
Data Intensity
Computational Intensity
Credit: Pat Harr, NSF
Modified from I. Qualters, NSF

Final Comments

- JCSDA starting pilot project exploring cloud solutions as community development environment
  - For collaborations across federal agencies
  - For research funded by JCSDA and private/academic partners

- Goal: collaborative Data Assimilation code repository, building, automated testing, datasets, deployment, documentation & tutorials

- Only scratching the surface of potential benefits from cloud computing (big data analytics)

- Welcome partnerships!

VIIRS Earth Day 20160422

Discussion…

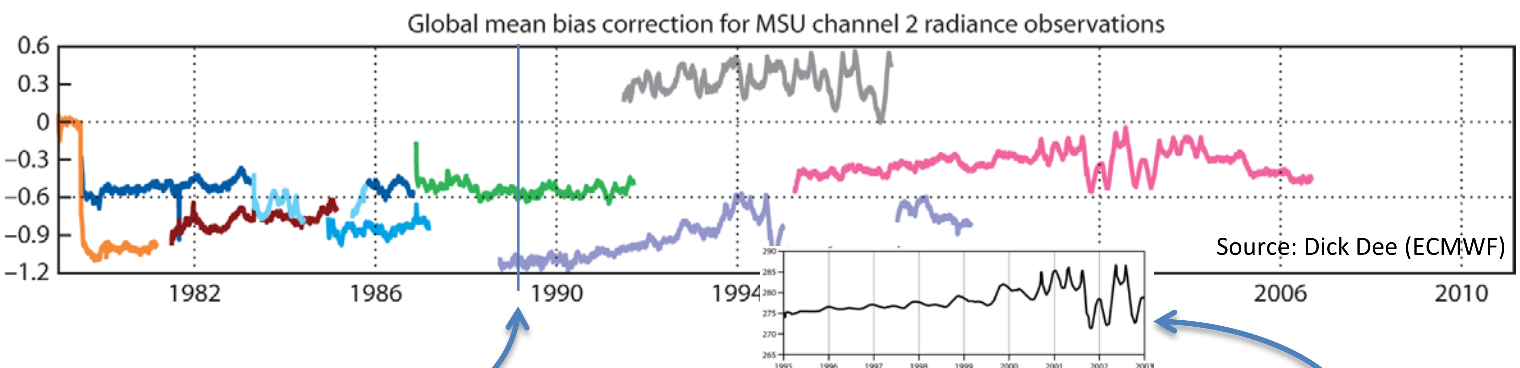
Introduction to Data Assimilation

Socio-economic benefit of NWP forecast: estim. \$100B-\$1T per year *(Riishojgaard, 2014)*

Contributions to NWP forecast: Initial Conditions = Model *(Magnusson and Källen, 2013)*

Initial Conditions: Satellites dominate global observation impact

Climate: key role of reanalyses



Jan 1989: Transition between two production streams

NOAA-14 recorded warm-target temperature changes, due to orbital drift (Grody et al. 2004)