Challenges and Solutions for Future Modeling Data Analysis Systems

Tsengdar Lee – tsengdar.j.lee@nasa.gov

NASA Headquarters

Dan Duffy, NASA GSFC

Seungwon Lee, JPL

Rama Nemani, NASA ARC

Duane Waliser, JPL

Jia Zhang, CMU

And Many Collaborators

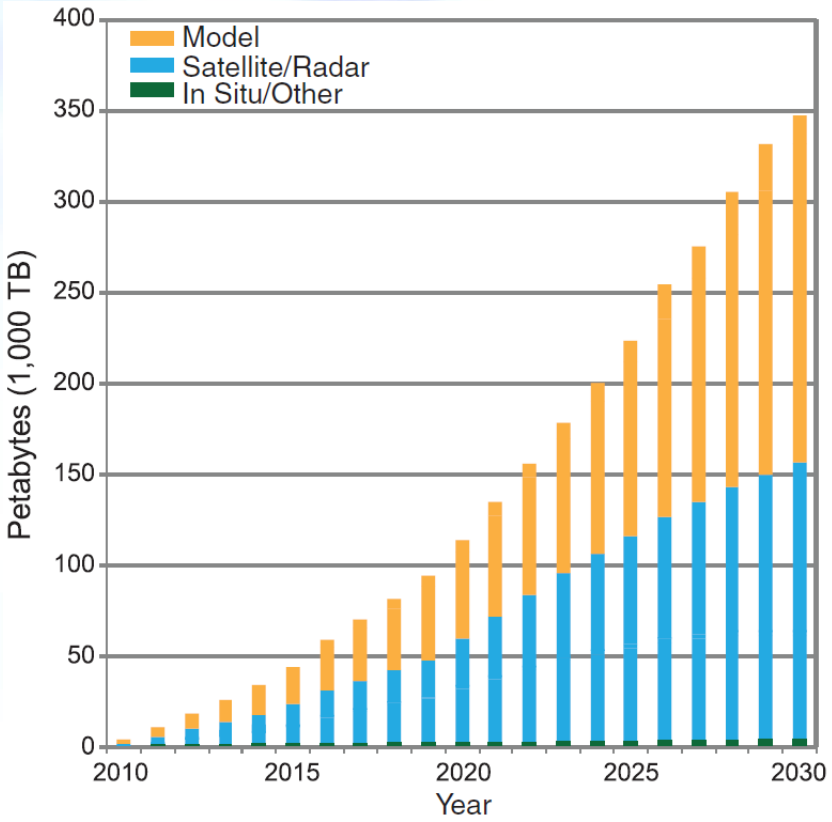
NAS CESAS – October 4, 2016

Projected Data Holding

•By 2020 it is estimated that all climate data holdings, including simulation, observation, and reanalysis sources, will grow to hundreds of exabytes in a worldwide-federated network [CKD Workshop, 2011 and CCDC Workshop, 2011].

– CCDC Workshop, International Workshop on Climate Change Data Challenges, June 2011, http://www.wikiprogress.org/index.php/Event:International_Workshop_on_Climate_Change_Data_Challenges.

– CKD Workshop, Climate Knowledge Discovery Workshop, March 2011, DKRZ, Hamburg, Germany, https://redmine.dkrz.de/collaboration/projects/ckd-workshop/wiki/CKD_2011_Hamburg.



- Climate Data Challenges in the 21st Century, Jonathan T. Overpeck, *et al.* *Science* 331, 700 (2011); DOI: 10.1126/science.1197869

Credit: LLNL/Dean Williams

2016 NASA Modeling Analysis & Simulation Product Plan

Forward Processing System	Satellite-Era Reanalysis 1979 - Present	EOS-Era Reanalysis 2000 – Present	Nature Runs (OSSEs)	Seasonal Forecast System	Coupled Simulations (Decadal, CMIP6)
3D-Hybrid Ensemble-Var (25km) 32 ensemble members Hydrostatic 1-Moment Cloud Microphysics *Current GEOS-5 FP system*	**MERRA (50km)** **Ending Feb. 2016** 3D-Var ~200 TB	**M2R12K (12km)** MERRA2 downscaled to 12 km Aerosols CO_2, CO, SO_2, O_3 Non-Hydrostatic 1-Moment Cloud Microphysics	**G5NR (7km)** Simulated 2005-2007 Aerosols, CO_2, CO, SO_2, O_3 Non-Hydrostatic 1-Moment Cloud Microphysics 4 PB	**GEOS SFS (50km)** MERRA-2 replay 50km, 40L ocean analysis 31 members per month Include aerosols, CO, CO_2 *M2-driven EnOI ocean analysis*	**GEOS CMIP (25km)** 25km Atmosphere 25km 50L ocean Include aerosols greenhouse gases Hydrostatic 2-Moment Cloud Microphysics
3D-Hybrid Ensemble-Var (12km) 32 ensemble members Atmosphere, ocean surface Hydrostatic 2-Moment Cloud Microphysics *Parallel FP stream in 1Q-2016*	**MERRA-2 (50km)** 3D-Var Aerosols and CO, SO_2, O_3 1-Moment Cloud Microphysics ~400 TB	**IESA (12km)** 3D-Hybrid Ensemble-Var 32 ensemble members atmosphere, land, ocean surface Aerosols, CO_2, CO, SO_2, O_3 Non-Hydrostatic 2-Moment Cloud Microphysics 5,000 cores ; 40 simulation days/day 150 days total wallclock ~3 to 4 PB of data	**G5NR-CHEM (12km)** Simulated 2013-2014 Replay to M2R12K Full Reactive Chemistry Non-Hydrostatic 1-Moment Cloud Microphysics 1 PB of data *4Q-FY2016*	**GEOS SFS (25km)** Alignment with "MERRA-3" 25km, 50L ocean analysis System design under review *FY2019 target*	*Planning/discussion and system evaluation in progress* *Will align with "MERRA-3" SFS and strategic direction o ESD*
4D Ensemble-Var (9km) ~100 ensemble members Atmosphere, ocean surface Non-Hydrostatic 2-Moment Cloud Microphysics (The first GEOS-6 system) *Parallel FP stream in 4Q-2016*	**MERRA-2 GMI replay (50km)** Replay GMI Chemistry 1 streams, 1,000 cores each 12 to 18 months ~ 1 PB	**IESAR4K (4km)** IESA Downscaled to 4km downscaling evaluation for NCA Aerosols, CO_2, CO, SO_2, O_3 Non-Hydrostatic 2-Moment Cloud Microphysics 5,000 cores ; 40 simulation days/day 150 days total wallclock ~3 to 4 PB of data	**G6NR (3km)** Simulated 2015 Aerosols CO_2, CO, SO_2, O_3, CH_4 Non-Hydrostatic 2-Moment Cloud Microphysics ~4 PB *Planning/evaluation*		
	Coupled Reanalysis ("MERRA-3") Atmosphere-land-ocean-cryosphere (alignment with SFS and CMIP6) *FY2019 target*				

- Core GMAO projects completed, in-progress
- Pathfinding projects toward GMAO core efforts.
- FY16 Projects
- Projects undergoing GMAO discussion/evaluatio
- Planned Future Projects

Past Foci

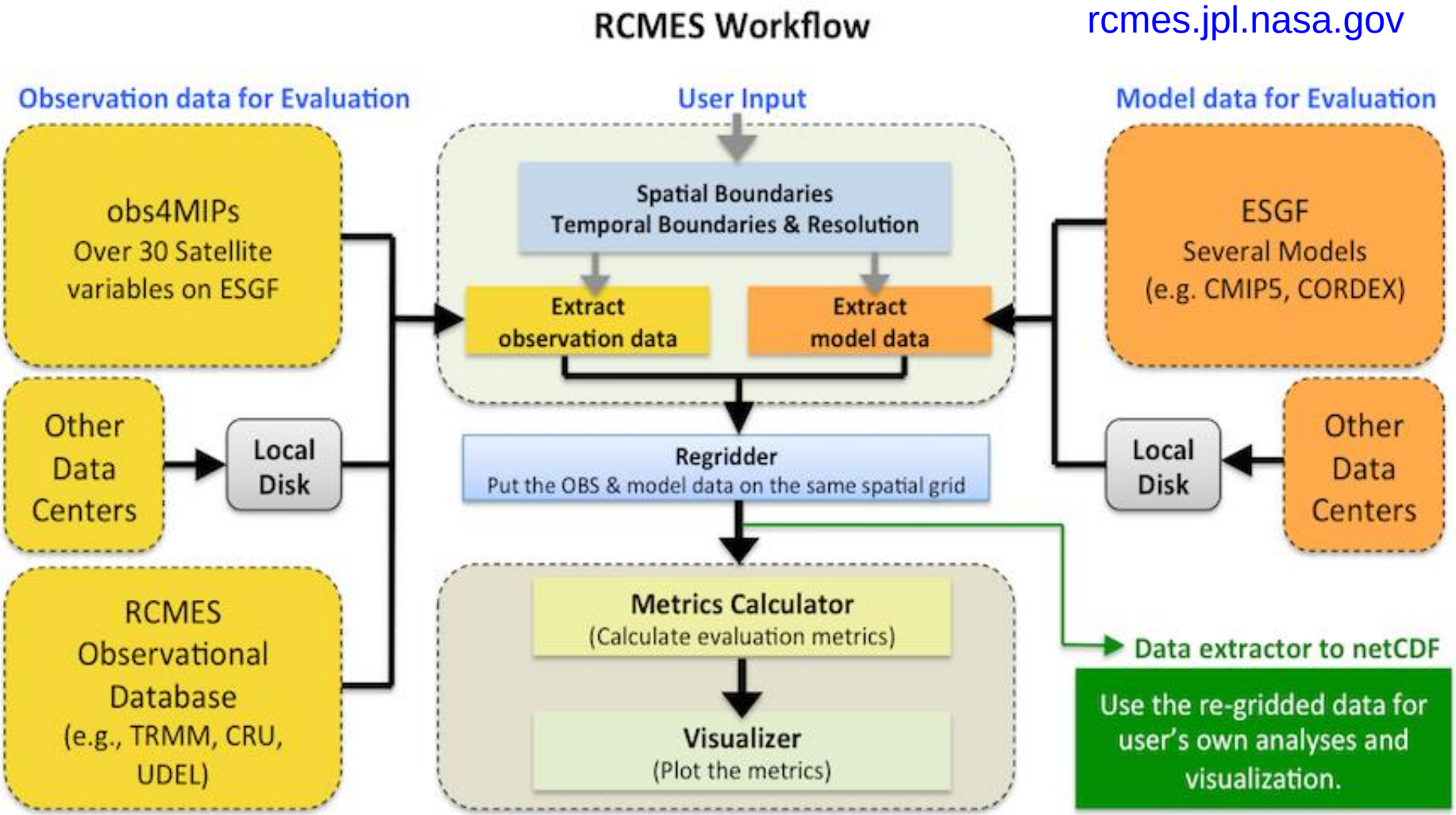
- Analogy:

- Challenges:
- Stewardship
- Curation
- Indexing
- Cataloging
- Searching
- Ordering
- Subsetting
- Provenance
- Lineage
- Data Mining
- Dissemination

Gearing up for Climate Modeling Data Analytics

- Traditional data center focuses on data archive, access and distribution
  - Scientists typically order and download specific data sets to a local machine to perform analysis
  - With large amount of observational and modeling data, downloading to local machine is becoming inefficient
  - Data centers are starting to provide additional services for data analysis
- NASA computing and computational science program is building “data analytics platforms” using “Climate Analytics as a Service” (CAaaS) such as NASA Earth Exchange (**NEX**), Regional Climate Modeling Evaluation Systm (**RCMES**), Climate Model Diagnostic Analyzer (CMDA) and Observation for Model Intercomparison Project (Obs4MIPs) using Earth System Grid Federation (ESGF)
  - Build on technologies
  - Enabled by a rule based data management system
  - Current research focuses on how to manage data movement from the archives to the analytical platforms

RCMES Architecture



Support – ARRA seed funding, NSF G8 ExArch Program, Mainly NASA NCA & Other

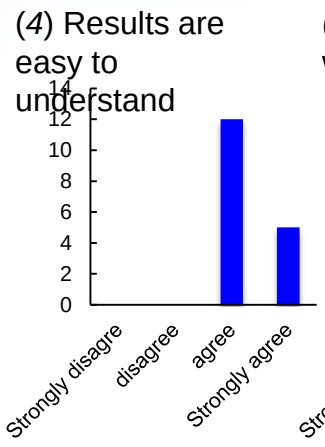
CORDEX-2016 Conference
Stockholm University, Sweden
May 19th Thursday, 2016

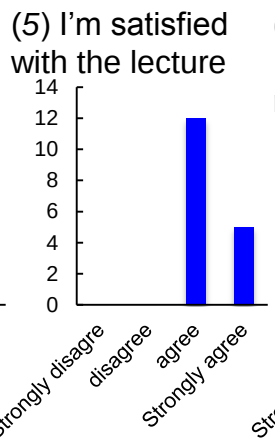
Attended by ~50 trainees from multiple countries, mostly in their early careers (students, postdocs, climate scientists, data scientists)

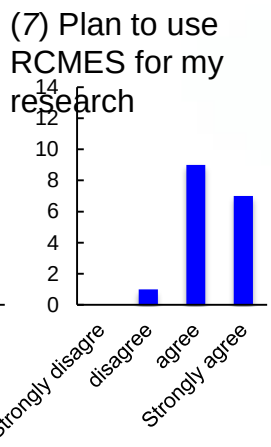
Among the 50 attendees, 20 responded to the survey. The responded attendees are: graduate students (7), postdocs (2), climate scientists (7), and data scientists (4).

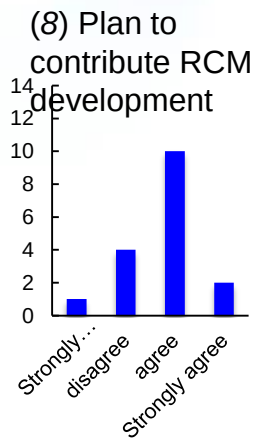
A eight-question survey was handed out at the end of the session. Responses to key items are compiled in the below:

- 17/17 agrees that RCMES results are easy to understand
- 17/17 were satisfied with the lecture
- 16/17 plan to use RCMES in their research
- 12/17 plan to contribute RCMES/OCW development



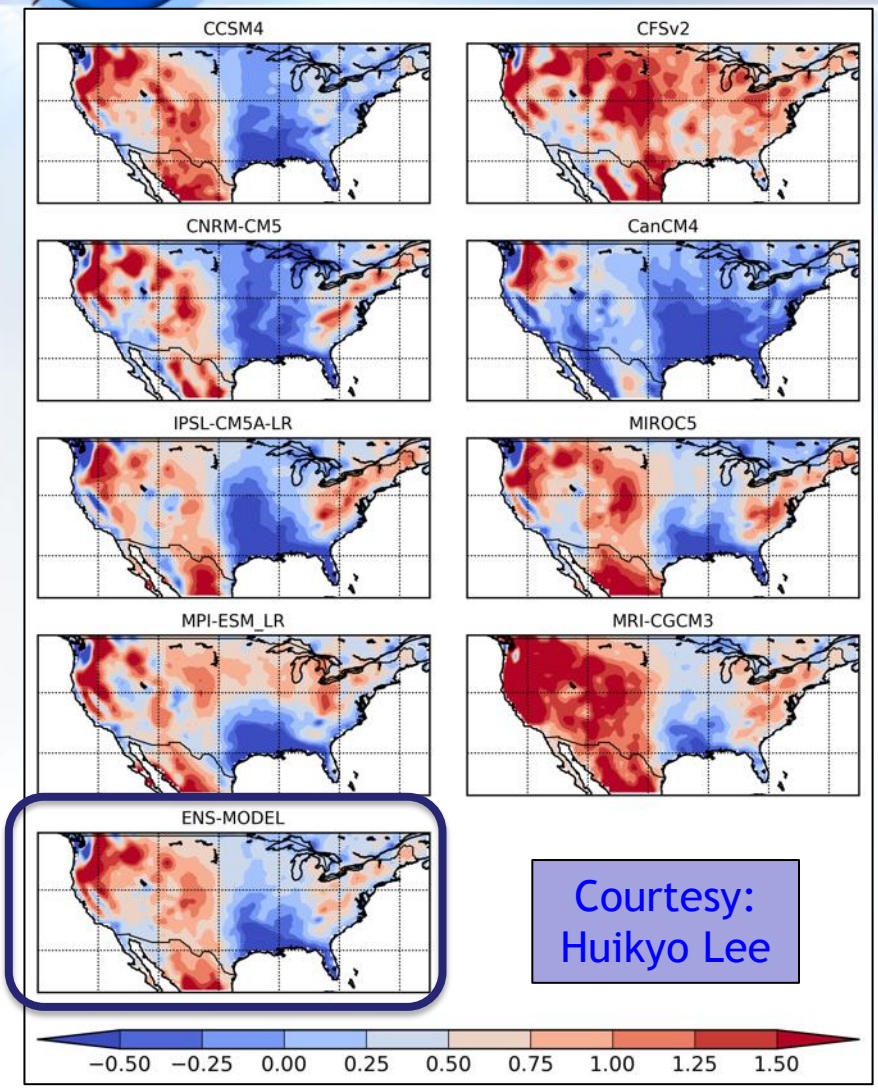






Direct Access to ESGF Available: Models & Observations

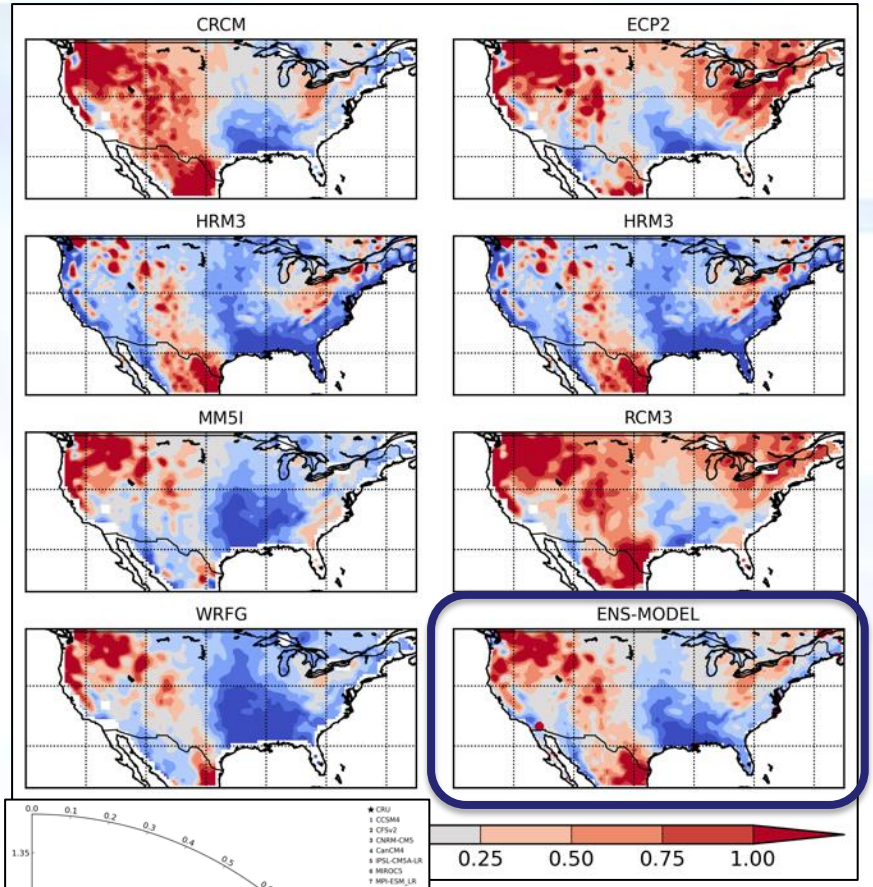
CMIP5 GCMs - observation

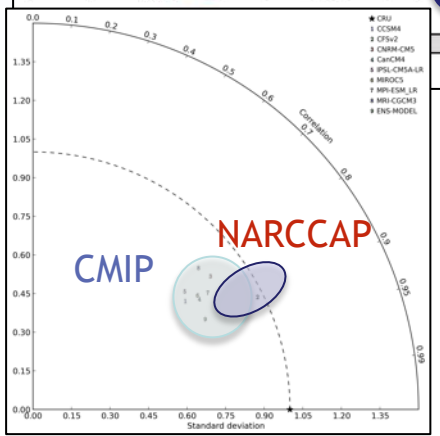


Courtesy: Huikyo Lee

Annual Precipitation Bias

NARCCAP RCMs - observation

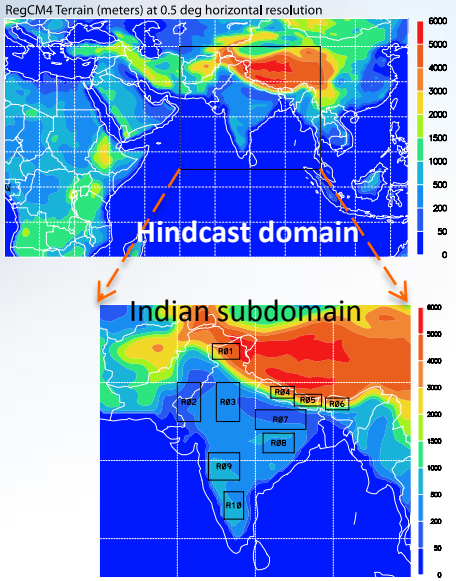




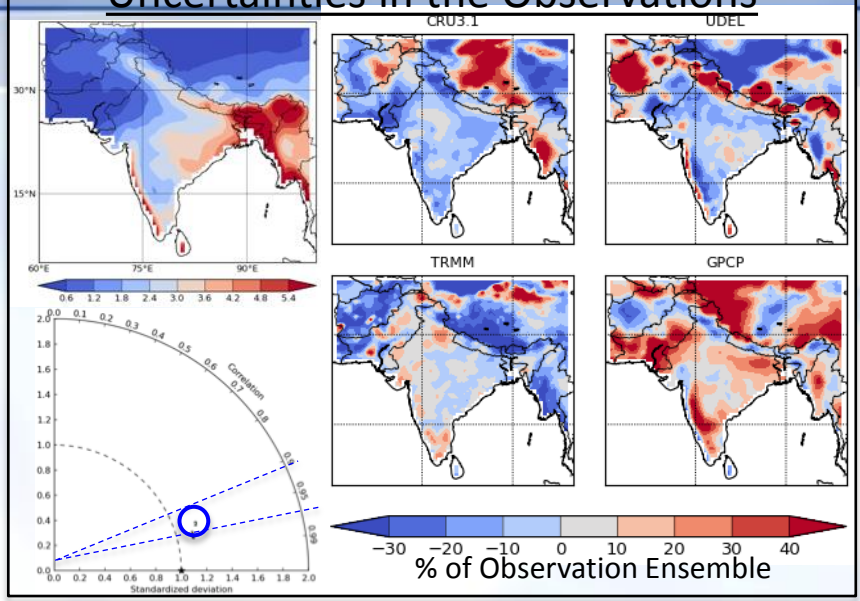
Capability to Perform Regional Evaluation of GCMs

Precipitation Evaluation of Multi-model Hindcast in the CORDEX South Asia – Indian subcontinent

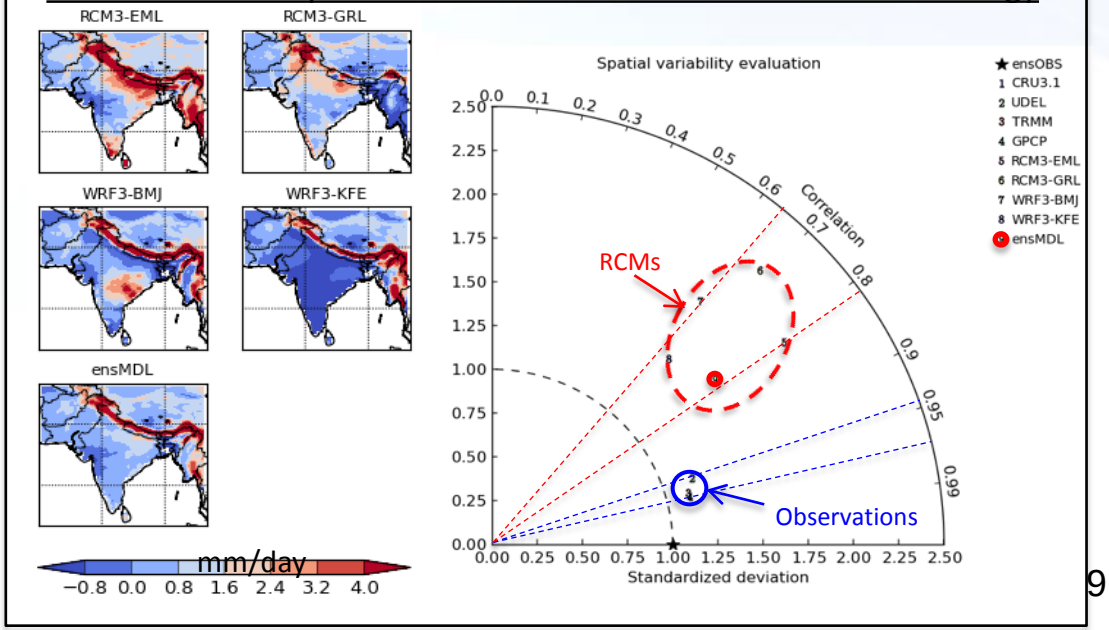
The hindcast and evaluation domain



Uncertainties in the Observations



Model biases in spatial variations in the annual-mean climatology

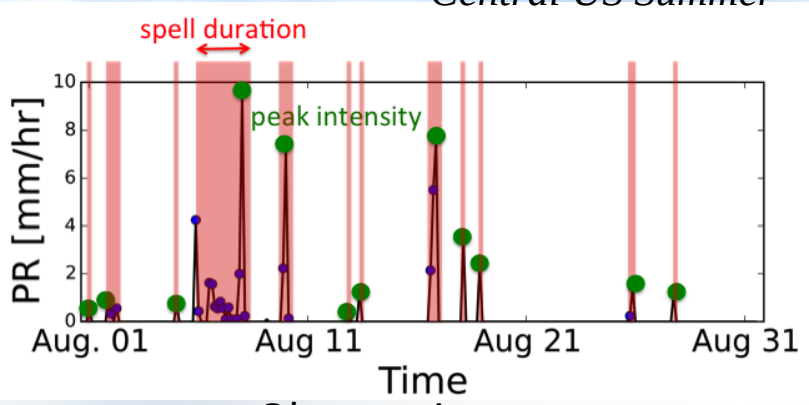


Kim, Sanjay Mattma, Boustani, Rao, Krishnan, Waliser, 2014, Uncertainties in Estimating Spatial and Interannual Variations in Precipitation Climatology in the India-Tibet Region from Multiple Gridded Precipitation Datasets, Int. J. Clim., Submitted.

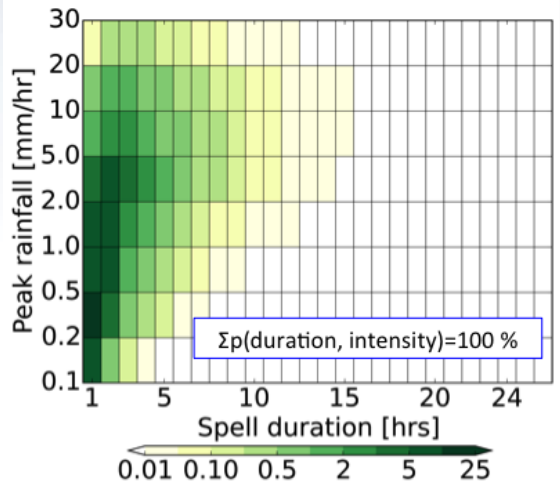
Precipitation Duration/Intensity Distributions

Reference : Stage IV 4km Gridded Observations

Central US Summer

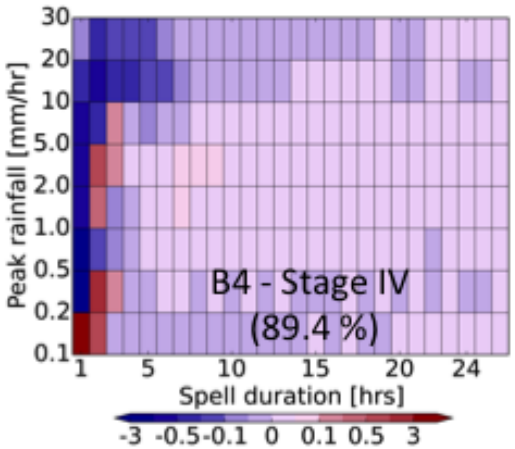


Observations

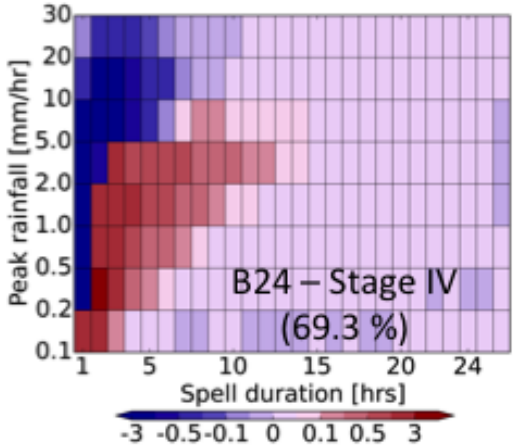


- From a model development as well as climate impacts assessment, the precipitation is important to represent correctly.
- Metrics are needed that limit data transfer or re-gridding needs.

4km nuWRF



24km nuWRF



- Higher resolution important for realistic precipitation extreme distributions
- Useful information can be obtained from native resolutions

Lee, Waliser, Braverman, 2016: In Prep

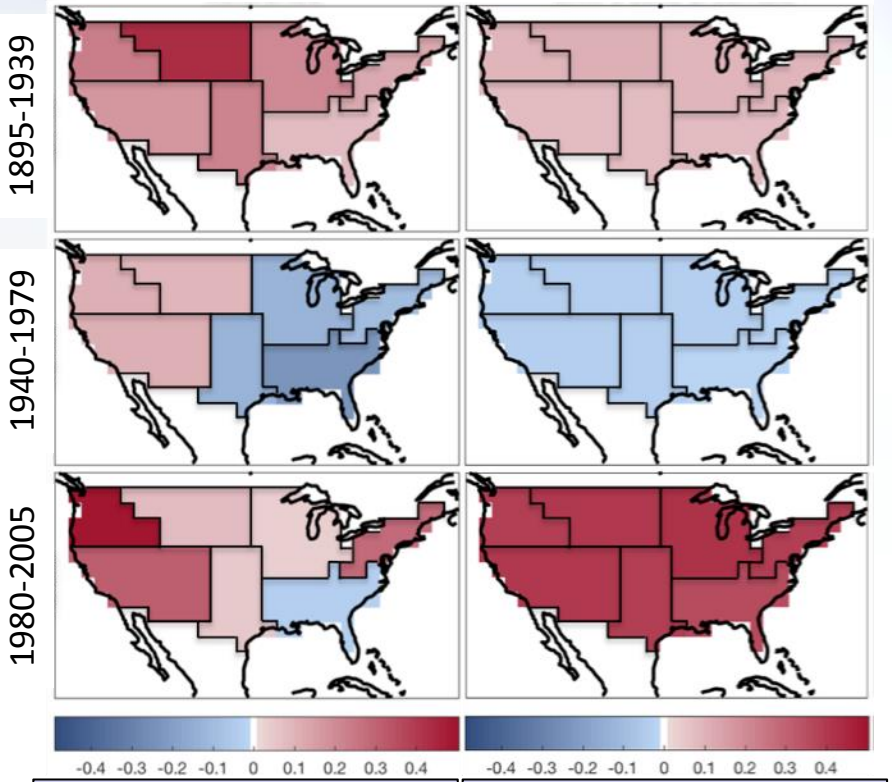
Regional Evaluation of the CMIP5 GCMs Ability to Reproduce Historical Surface Air Temperature and Precipitation Trends

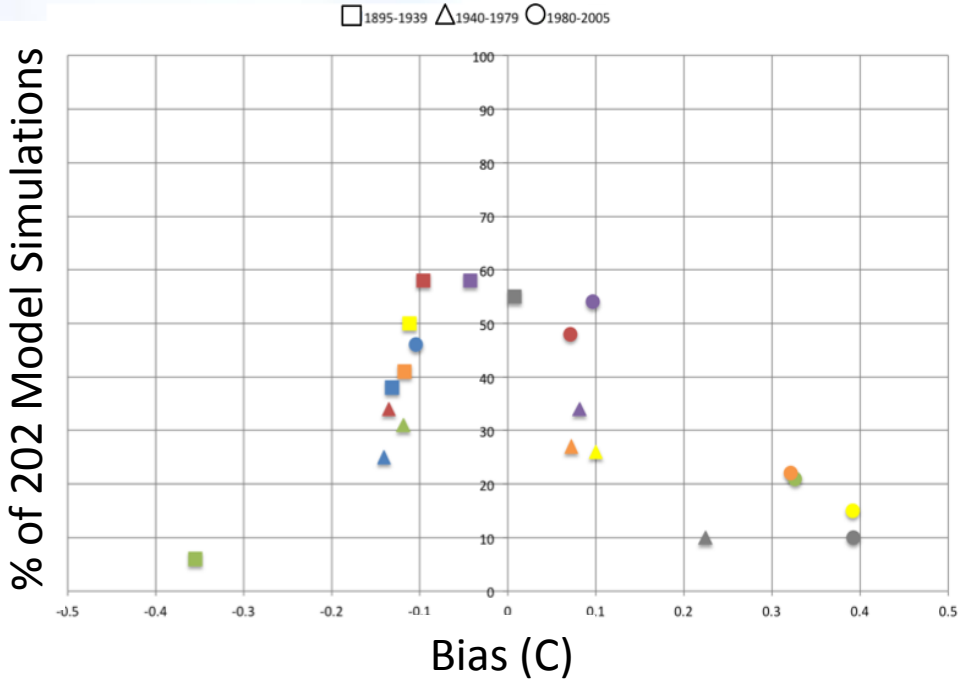
J. Lee, P. Loikith, K. Kunkel, H. Lee, D. Waliser

NCA-defined regional analysis

How well do CMIP5 models reproduce observed trends in precipitation and surface temperature (i.e. nClimDiv) over CONUS?

Observations **CMIP Ensemble**





Southeast "warming hole"

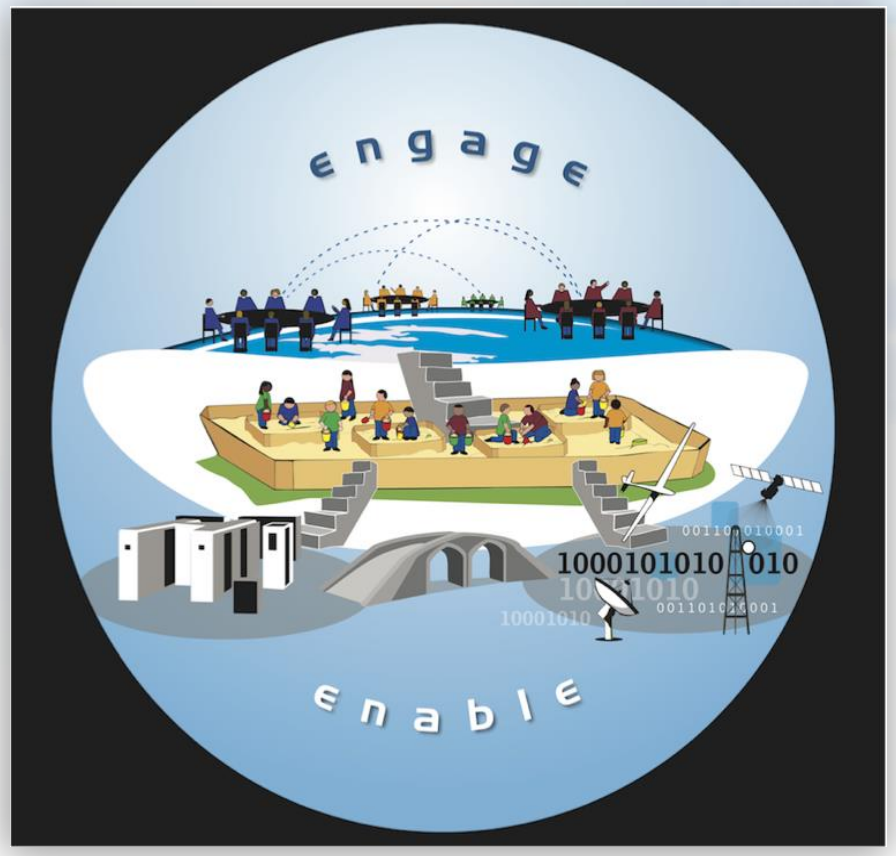
No Southeast "warming hole"

Paper development in progress
Metrics incorporated into RCMES over Fall

NASA Earth Exchange (NEX)

OVERVIEW

+ NEX is virtual collaborative that brings scientists together in a knowledge-based social network and provides the necessary tools, computing power, and access to bigdata to accelerate research, innovation and provide transparency.



VISION

To provide "science as a service" to the Earth science community addressing global environmental challenges

GOAL

To improve efficiency and expand the scope of NASA Earth science technology, research and applications programs

NEX Provides a Complete Work Environmentb

"Science As A Service"

COLLABORATION

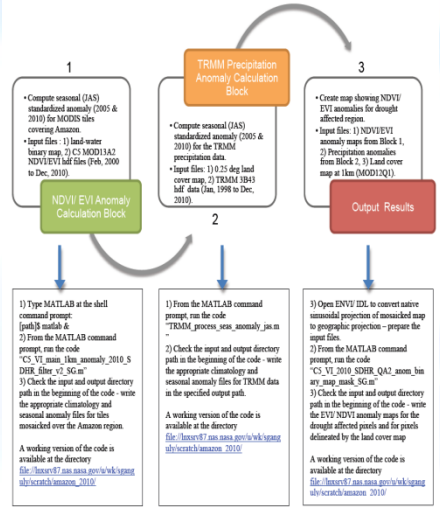
Over 400 Members

COMPUTING

Scalable
Diverse
Secure/Reliable

CENTRALIZED DATA REPOSITORY

Over 2300 TB of Data



KNOWLEDGE

Workflows
Machine Images
Model codes
Re-useable software

NEX Resources

Portal

- Web server
- Database server
- 503 registered members (up from 420)

Sandbox

- 96-core server, 264GB memory, with 320 TB storage
- 48-core server, 128 GB, 163 TB storage

HPC

- 720-core dedicated queue + access to rest of Pleiades
- 181 users/44 active (153/40 last year)
- 2.3 PB storage (from 850TB)

Models/Tools/Workflows used by NEX User Community

- GEOS-5
- CESM
- WRF
- RegCM
- VIC
- BGC
- LPJ
- TOPS
- BEAMS
- Fmask
- LEDAPS
- METRIC

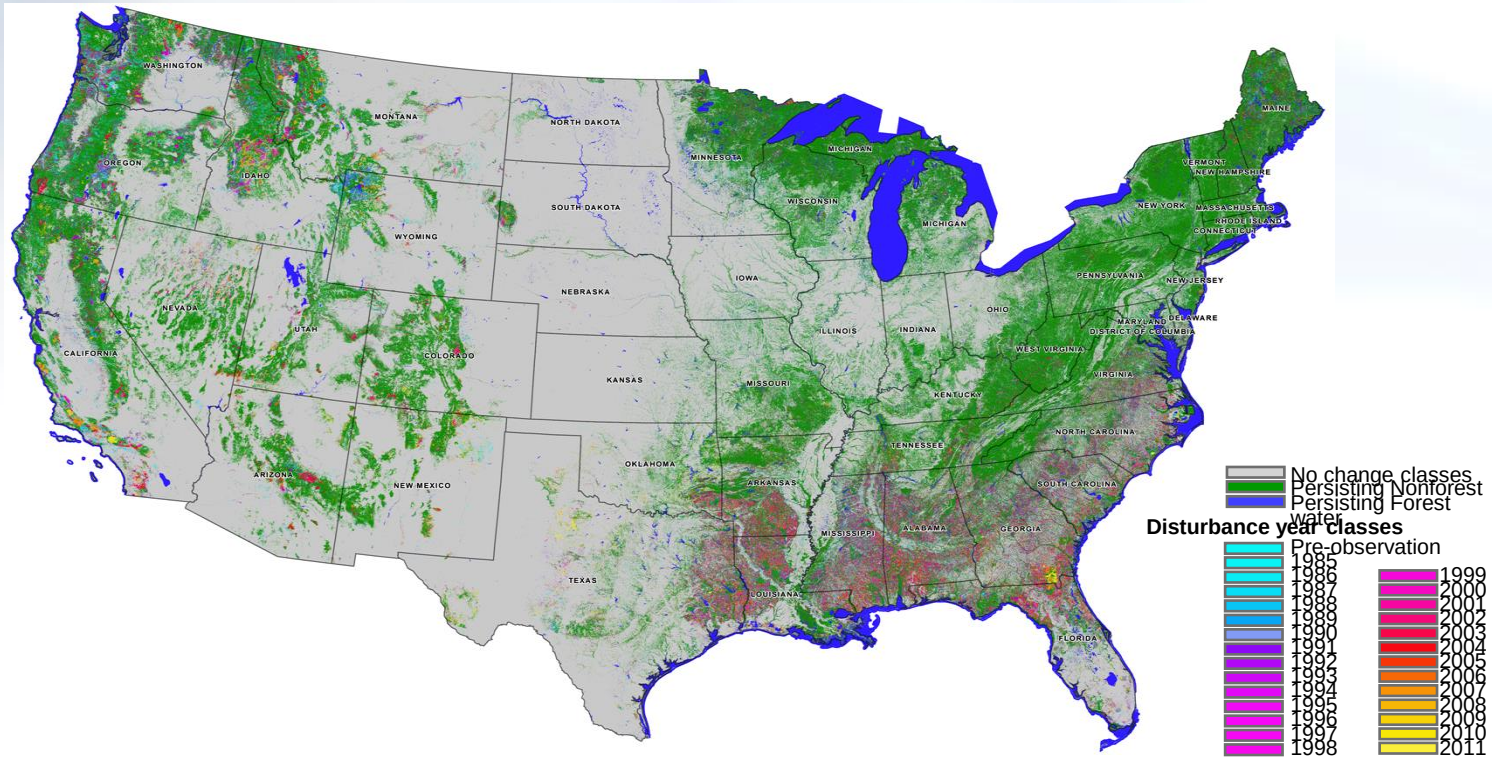
…

Data (>2 PB on & near-line)

- Landsat
- MODIS
- TRMM
- GRACE
- ICESAT
- CMIP5
- NCEP
- MERRA
- NARR
- PRISM
- DAYMET
- NAIP
- Digital Globe
- NEX-DCP30
- NEX-GDDP
- LOCA
- BCCA
- WELD
- NAFD-NEX

Long-term Satellite Data Analysis
Rates of Disturbance (LANDSAT)

Forest disturbance tracking with Landsat with implications for carbon cycle modeling



High Resolution Climate Projections

Climate Downscaling

DCP30 (Downscaled Climate Projections at 30arc sec)
Domain/Resolution: CONUS, ~800m
Frequency: Monthly
Variables: Tmax, Tmin, and Precip
No of CMIP5 models: 34
Baseline Data: Daly et al., 2002
Funding: NASA

BCCA (Bias Corrected Constructed Analogs)
Domain/Resolution: CONUS, ~12km
Frequency: Monthly
Variables: Tmax, Tmin, Precip
No of CMIP5 models: 21
Baseline Data: Maurer et al. 2002
Funding: USBR

LOCA (Localized constructed analogs)
Domain/Resolution: CONUS, ~6km
Frequency: Daily
Variables: Tmax, Tmin, Precip;
Humidity, Windspeed (in progress)
No of CMIP5 models: 32
Baseline Data: Livneh et al. 2013
Funding: USBR/CalEnergy

GDDP (Global Daily Downscaled Climate Projections)
Domain/Resolution: Global, ~25km
Frequency: Daily
Variables: Tmax, Tmin, and Precip
No of CMIP5 models: 21
Baseline Data: Sheffield et al. 2006
Funding: NASA

Community Engagement

Climate Model Diagnostic Analyzer

- Web-based tools running on Amazon cloud.
- Only requirement from a user machine is a web browser with an internet connection. No local installation needed.
- Provides datasets and analysis services.
- You can analyze the datasets using the services.
- You can download analyzed output datasets.
- You can download original input datasets.

Major Challenges Over Next 10 Years and What Can We Do Now

- Modeling and observational data will continue to grow exponentially
  - Major challenge in modeling data management, analysis, and collaboration
  - Tape archives will not meet these challenges
  - Network will not catch up
  - Library model will no longer work
  - Explore and adopt new storage technologies (e.g., object storage)
  - Build centralized data analytics systems
  - Data proximal analytic capabilities (move the analytics to the data)
  - Commoditize data storage and data analytics
- Large scale science informatics system will be needed to solve the future data challenbges