Data Producers Courting Data Reusers: Two cases from modeling communities

Jillian C. Wallis | UCLA

IDCC | 2014.02.26

Sharing Research Data is…

- Good for public, individual, academia
- Still not the norm
- Difficult

Dis/Incentives for Data Sharing

- Data Producers
  - Data sharing/management policy pressure
  - Intellectual property concerns
  - Provide complete metadata
  - Lack of credit for effort
  - No promise of reuse
- Data Reusers
  - In order to reuse data:
    - Assess relevance of data
    - Understand the data
    - Determine whether the data were trustworthy
  - Rely on formal training to provide data context
  - Trust the data producer in place of other context

Crossing Disciplinary Boundaries

- Concerns about misuse
- Lack of contextual knowledge/training
- Provide metadata for unknown reusers and reuses

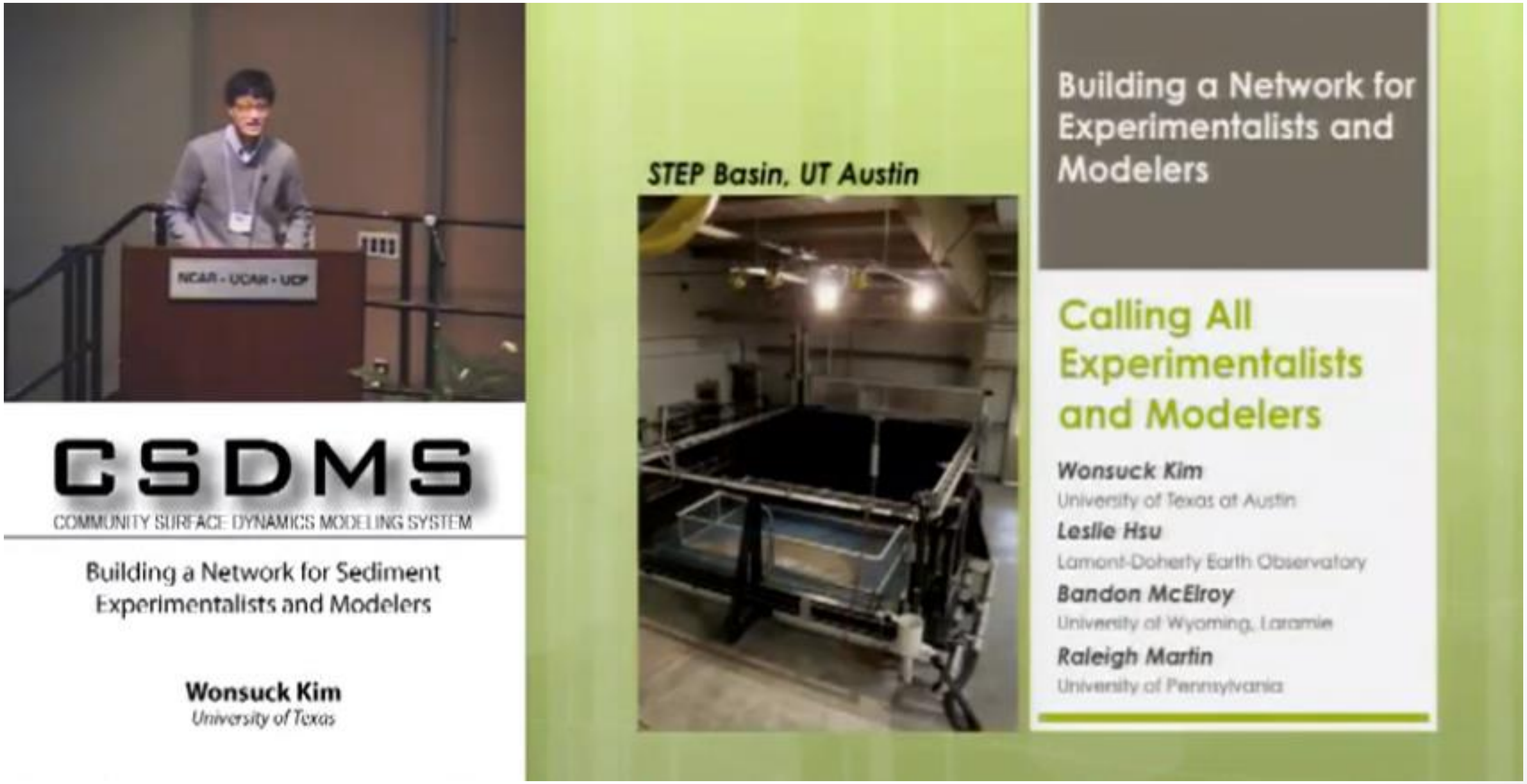
What happens when data producers work with data reusers?

Method

- Case Studies
  - Community Surface Dynamics Modeling System (CSDMS)
    - Annual Meeting
    - Keynote by Dr. Wonsuck Kim
    - Experimentalists courting modelers
  - National Climate Predictions & Projections Platform (NCPP)
    - Qualitative Evaluation of Downscaling (QED) workshop
    - Modelers courting policy makers from a variety of sectors
      - Agriculture, ecology, human health, water
- Participant observation
- Casual interviews and exit surveys

Case 1: CSDMS Annual Meeting

Case 1: CSDMS Keynote



CSDMS

COMMUNITY SURFACE DYNAMICS MODELING SYSTEM

Building a Network for Sediment Experimentalists and Modelers

Wonsuck Kim
University of Texas

STEP Basin, UT Austin

Building a Network for Experimentalists and Modelers

Calling All Experimentalists and Modelers

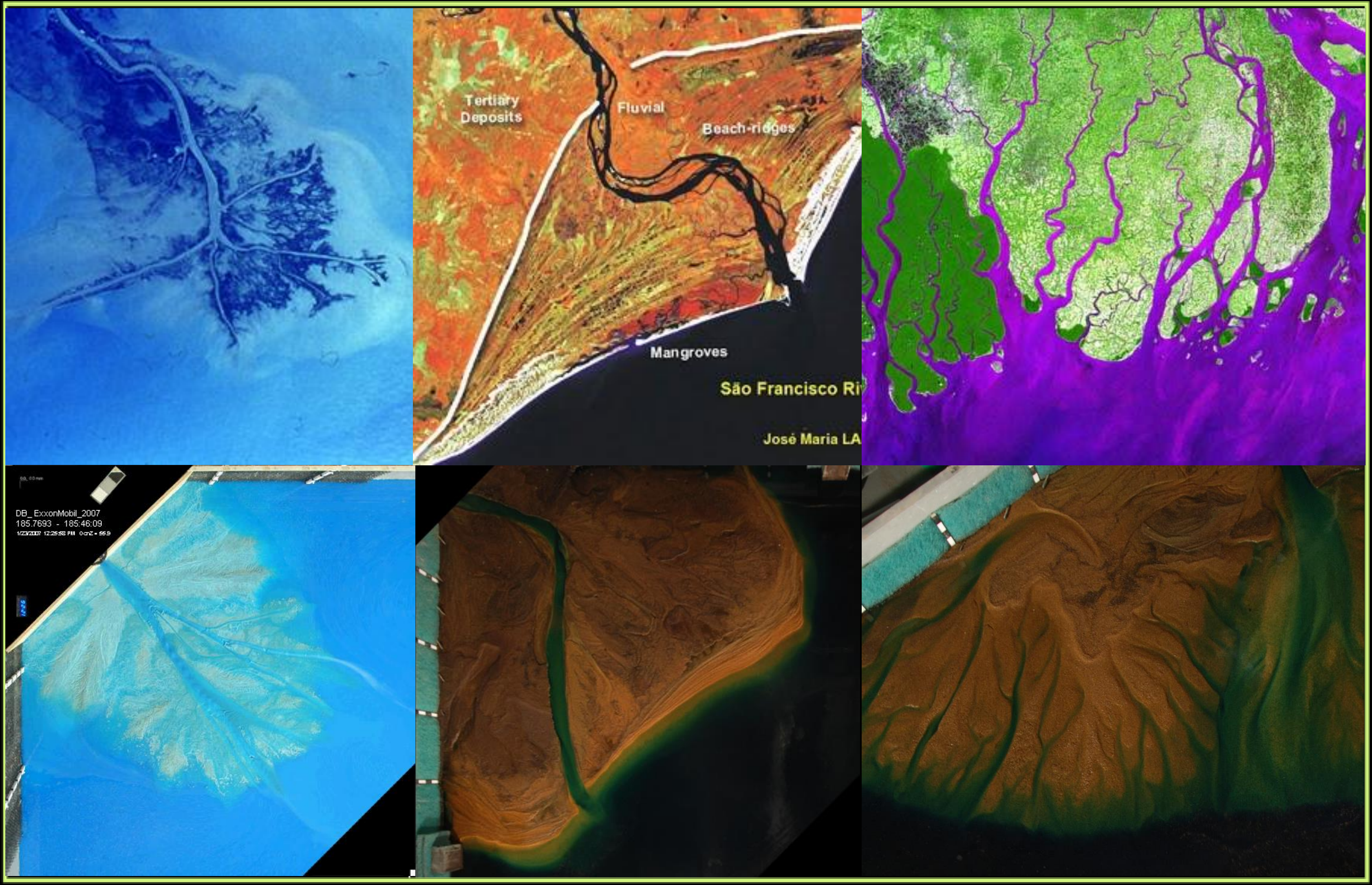
Wonsuck Kim
University of Texas at Austin

Leslie Hsu
Lamont-Doherty Earth Observatory

Bandon McElroy
University of Wyoming, Laramie

Raleigh Martin
University of Pennsylvania

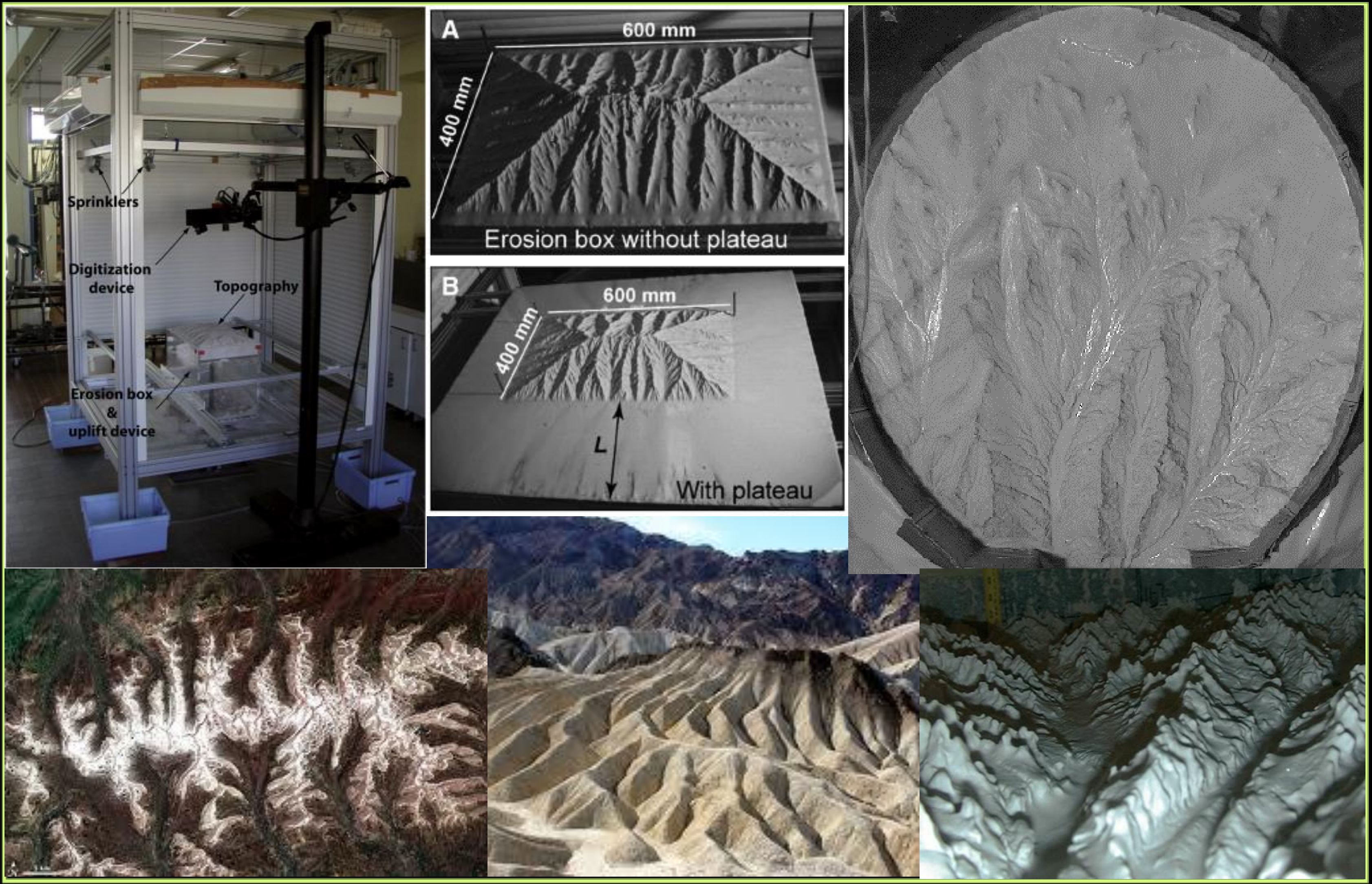
Experiments Deltas



Martin, Paola, Baumgardner, Cazanacli, & Abeyta

Experiments Erosional Landscape

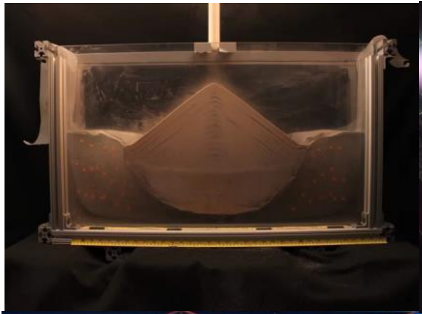
Bonnet, Crave, Hasbargen, Paola

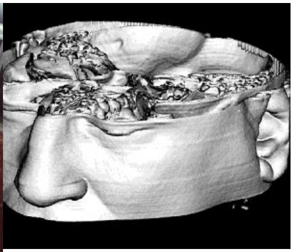


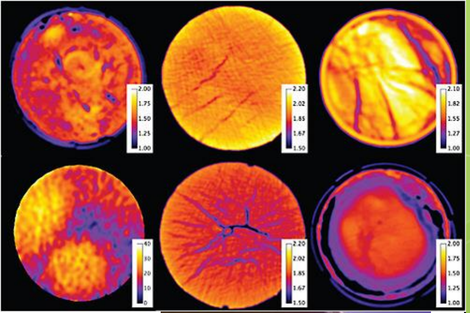
The Future of Experiment

- Slide stolen from Gary Parker (UIUC)

- **Tomography** - imaging internal stratigraphy
- **RFID and GPS tracking** - tracking of all particles
- Digital camera, topographic scanner - cheaper and better
- New materials in fluid, sediment, and substrate alternatives







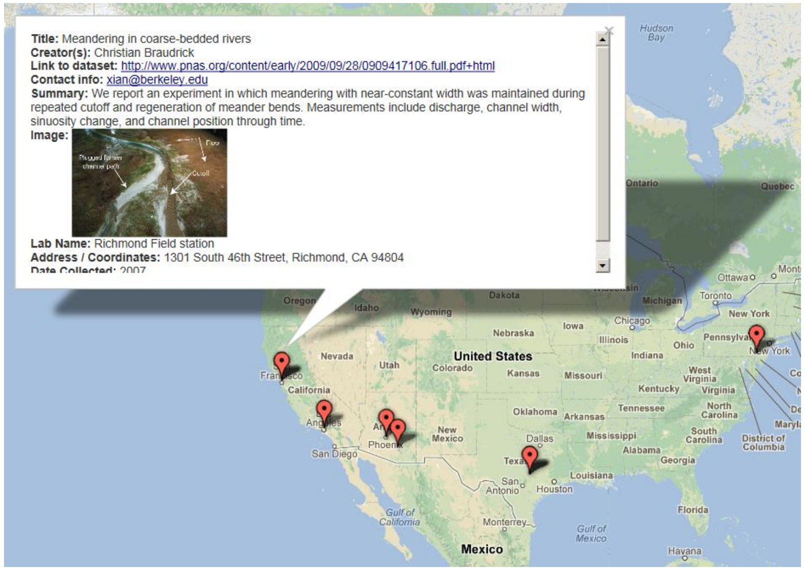
Calling All Experimentalists and Modelers

- 2012 AGU Town hall meeting and Workshop at UT Austin

- **NEEDS:**
  - Best practices in experimental methods and in the storage, archiving, and dissemination of experimental data
  - Need for a centralized place to deposit data or solicit information
  - Standards or guidelines to facilitate interoperability and reuse
  - More frequent communication between investigators will lead to rescue of data and knowledge from inaccessible dark data storage and will accelerate learning and production of results and analysis

Calling All Experimentalists and Modelers

- Building a Sediment Experimentalist Network (SEN)
- **SEN Knowledge Base (SEN-KB):**
  - a **data repository** leveraging and building on the existing National Center for Earth-surface Dynamics **(NCED) Data Repository**
  - synthesizes research activities of experimentalists by continuously aggregating existing and newly-collected experimental data.
  - **Modelers: We need your inputs for** best practices and metadata to effectively share data.



Title: Meandering in coarse-bedded rivers
Creator(s): Christian Braudrick
Link to dataset: http://www.pnas.org/content/early/2009/09/28/0909417106.full.pdf+html
Contact info: xian@berkeley.edu
Summary: We report an experiment in which meandering with near-constant width was maintained during repeated cutoff and regeneration of meander bends. Measurements include discharge, channel width, sinuosity change, and channel position through time.
Image:
Lab Name: Richmond Field station
Address / Coordinates: 1301 South 46th Street, Richmond, CA 94804
Date Collected: 2007

Fusion table: prototype database showing locations and products of active laboratory research. Currently 6 laboratories and 14 data sets

Calling All Experimentalists and Modelers

- **SEN Experimental Collaboratories (SEN-EC), creating a new form of research collaboration in our community:**
  - Modelers: Participate in **Community experiment**
  - **Formulate and address grand challenges together**
  - **Community experiment in STEP basin at UT Austin**
    - Survey to all participants with some guidelines before the workshop
    - Conduct a community experiment together with onsite and virtual participants
    - **Upload all data (images, topography, sliced deposit sections) through the Fusion Table**
    - **Provide a web resource to discuss "how to use the data?"**

CSDMS Keynote Outcomes

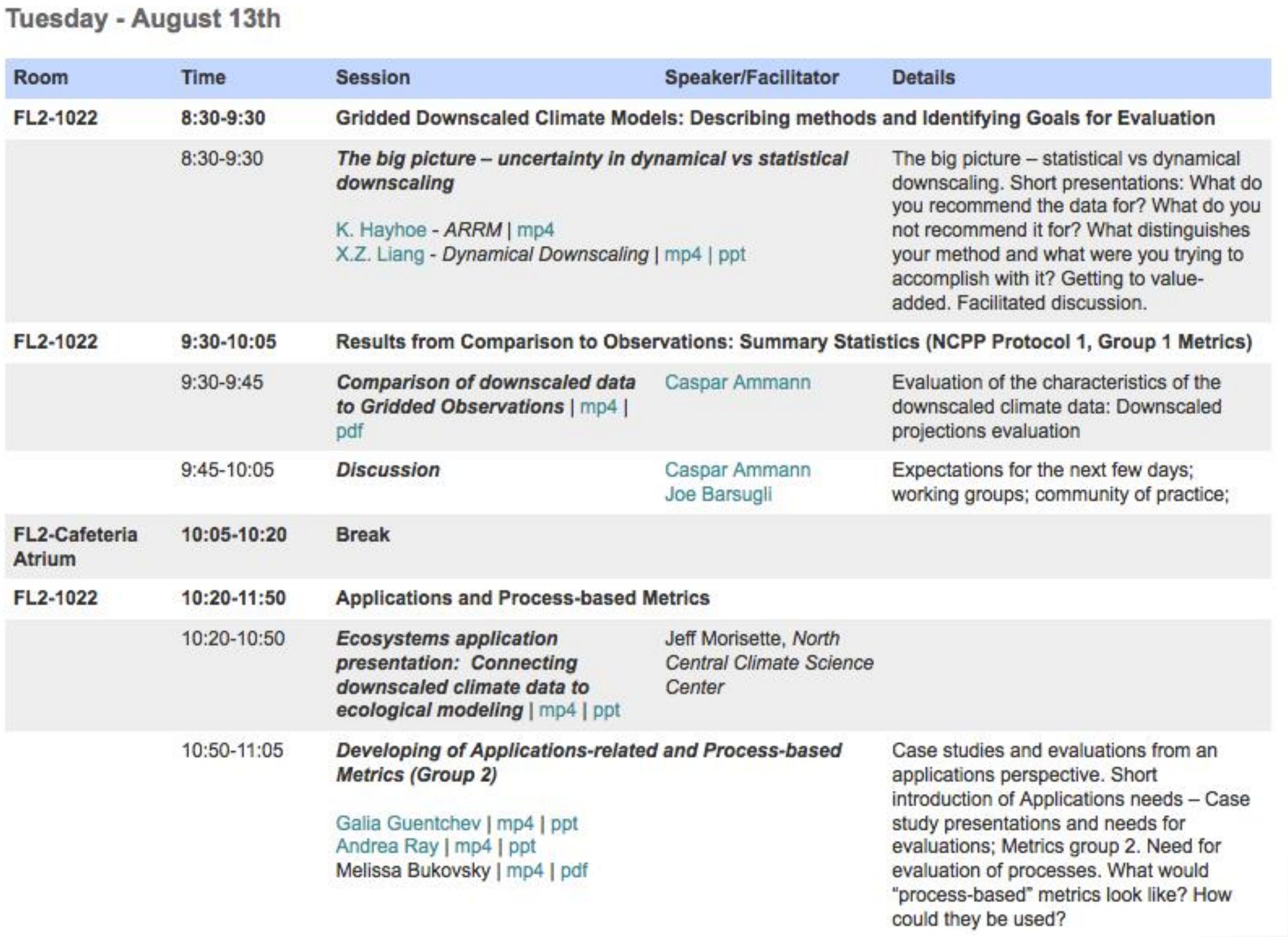
- The keynote fit with the over-arching theme of “tracking uncertainty”
- Ended with an open invitation to collaborate
- Dr. Kim’s talk created a lot of buzz during the meeting
- He has since had modelers approach him to collaborate

Case 2: NCPP QED Workshop

- Climate model outputs are not easy to "just use"
- NCPP Goal:
  - *The National Climate Predictions & Projections (NCPP) Platform works to advance the provision of regional and local information about the evolving climate and to accelerate its use in adaptation planning and decision making.*
- NCPP Qualitative Evaluation of Downscaling Workshop Goal:
  - *A goal of our evaluation workshop this summer is to assist the targeted communities in planning. This may include determining what type of information is needed by practitioners, defining how to choose between different data sets that can be used to obtain the necessary information, and/or providing narratives on past and future impacts.*

NCPP QED Interactions

Workshop Agenda



Tuesday - August 13th

Room	Time	Session	Speaker/Facilitator	Details
FL2-1022	**8:30-9:30**	**Gridded Downscaled Climate Models: Describing methods and Identifying Goals for Evaluation**		
	8:30-9:30	***The big picture – uncertainty in dynamical vs statistical downscaling*** K. Hayhoe - *ARRM* | mp4 X.Z. Liang - *Dynamical Downscaling* | mp4 | ppt		The big picture – statistical vs dynamical downscaling. Short presentations: What do you recommend the data for? What do you not recommend it for? What distinguishes your method and what were you trying to accomplish with it? Getting to value-added. Facilitated discussion.
FL2-1022	**9:30-10:05**	**Results from Comparison to Observations: Summary Statistics (NCPP Protocol 1, Group 1 Metrics)**		
	9:30-9:45	***Comparison of downscaled data to Gridded Observations*** | mp4 | pdf	Caspar Ammann	Evaluation of the characteristics of the downscaled climate data: Downscaled projections evaluation
	9:45-10:05	***Discussion***	Caspar Ammann Joe Barsugli	Expectations for the next few days; working groups; community of practice;
FL2-Cafeteria Atrium	**10:05-10:20**	**Break**		
FL2-1022	**10:20-11:50**	**Applications and Process-based Metrics**		
	10:20-10:50	***Ecosystems application presentation: Connecting downscaled climate data to ecological modeling*** | mp4 | ppt	Jeff Morisette, *North Central Climate Science Center*	
	10:50-11:05	***Developing of Applications-related and Process-based Metrics (Group 2)*** Galia Guentchev | mp4 | ppt Andrea Ray | mp4 | ppt Melissa Bukovsky | mp4 | pdf		Case studies and evaluations from an applications perspective. Short introduction of Applications needs – Case study presentations and needs for evaluations; Metrics group 2. Need for evaluation of processes. What would "process-based" metrics look like? How could they be used?

Example Climate Modeler Talk

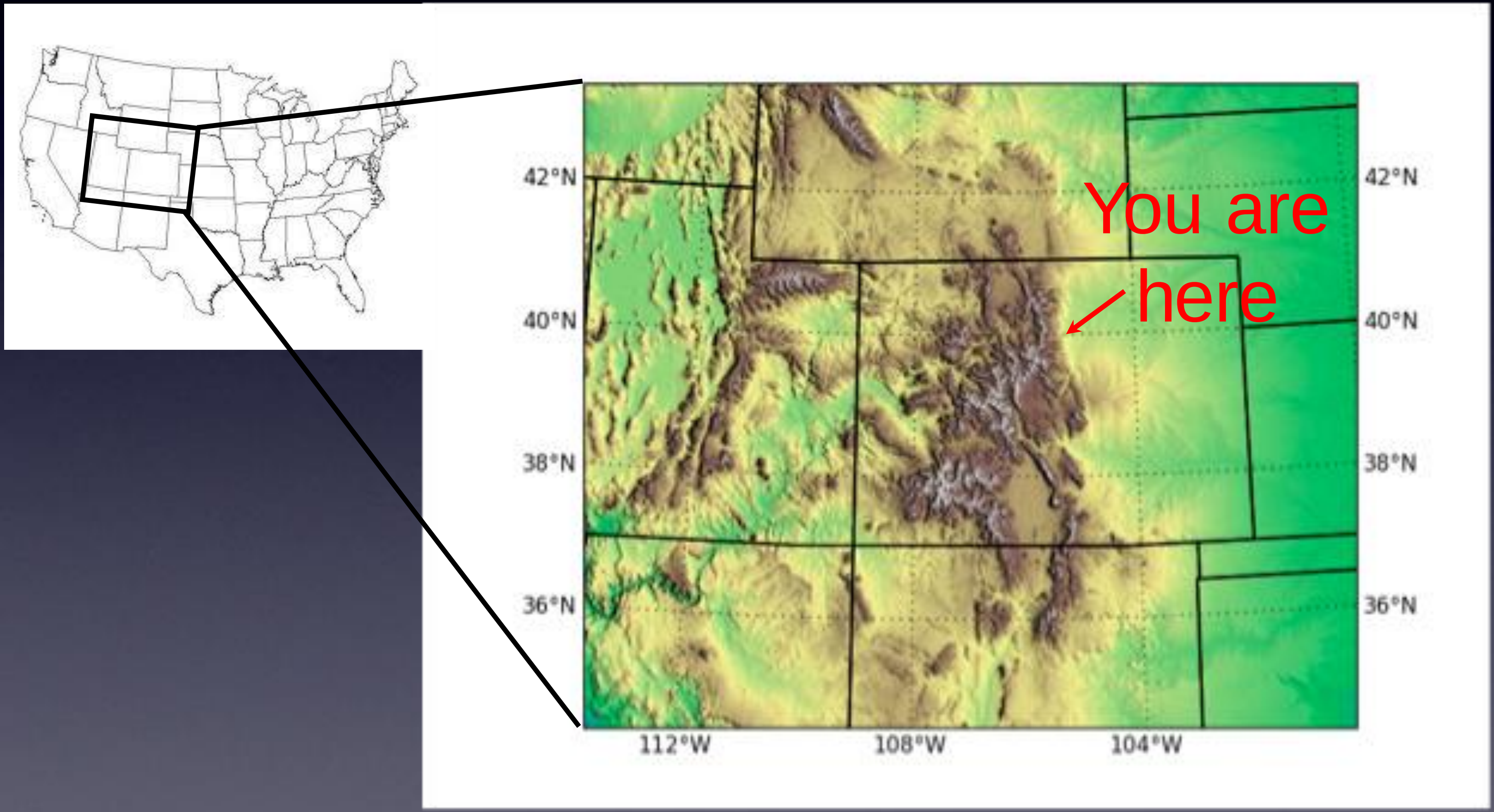
Climate Data from Observations, Statistics, and Physics

Ethan Gutmann, Tom Pruitt, Naoki Mizukami, Martyn Clark, Levi Brekke, Jeffrey Arnold, Changhai Liu, Roy Rasmussen

8/12/2013

NCPP - Quantitative Evaluation of Downscaling Workshop

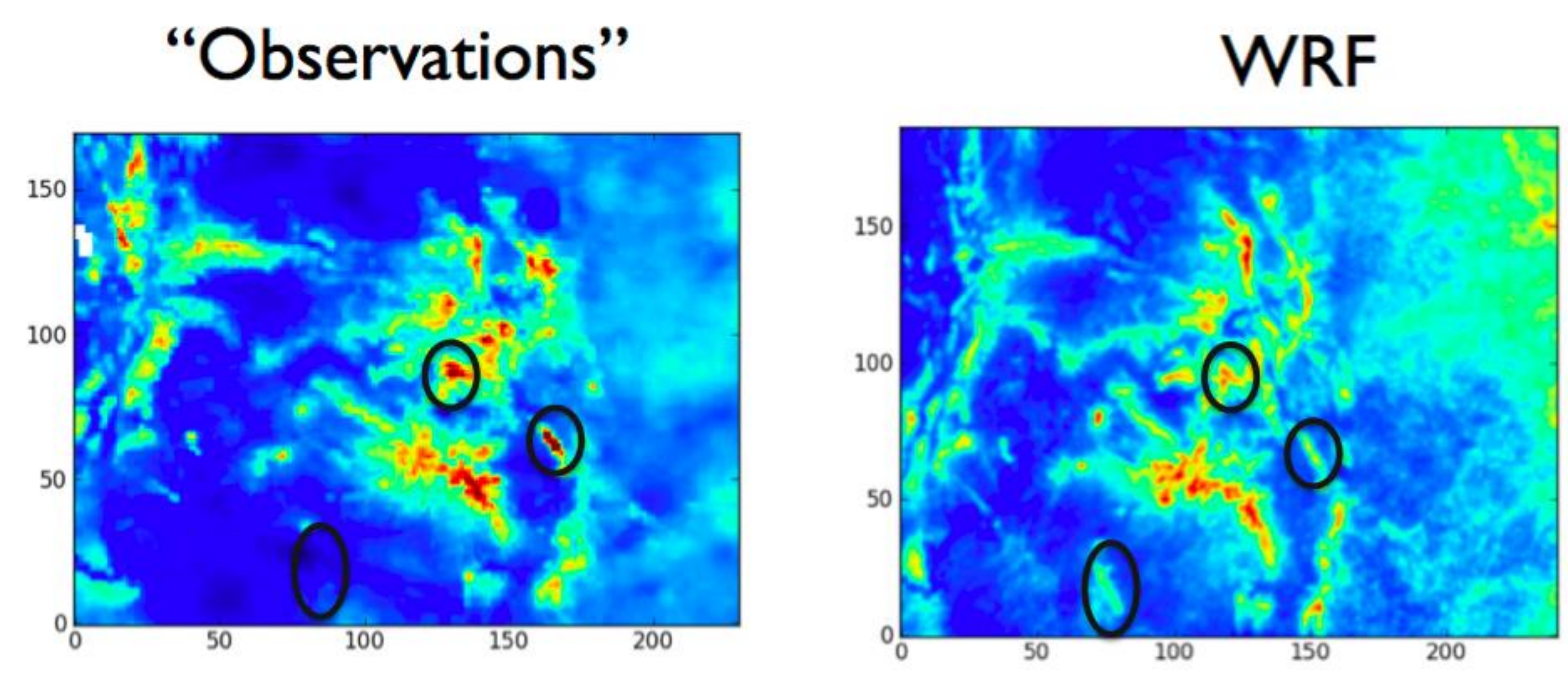
Before we begin



You are here

Observations

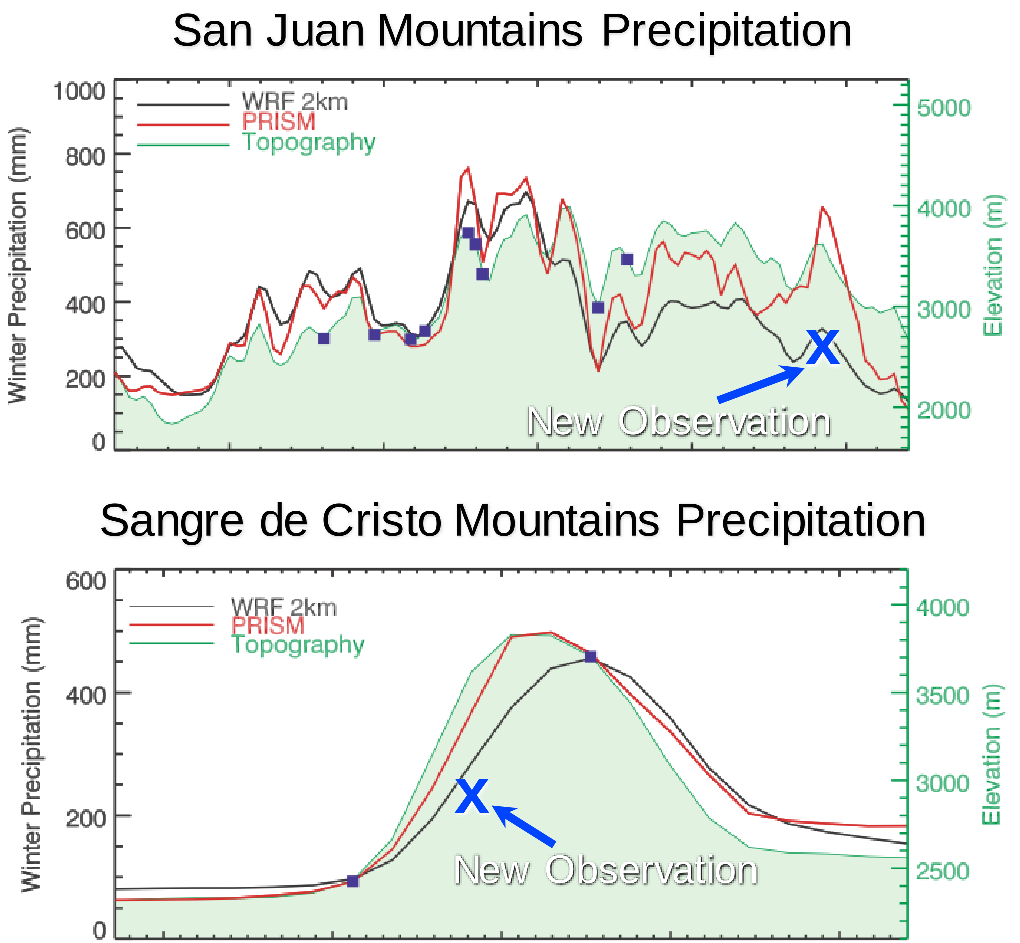
Mean Annual Precipitation



Ethan Gutmann, *Climate Data from Observations, Statistics, and Physics*

Observations

- Statistically derived "observations" don't get spatial distributions right even in current climate.



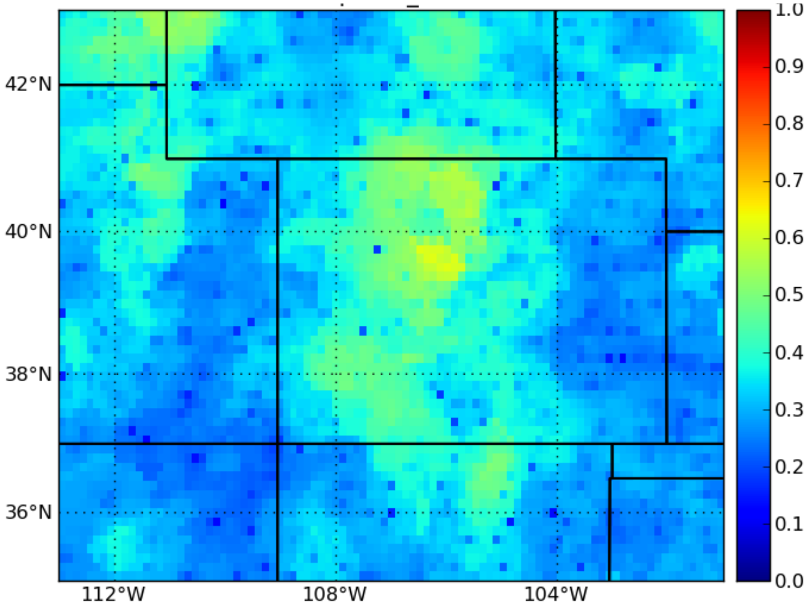
Gutmann et al. (2012)

Ethan Gutmann, *Climate Data from Observations, Statistics, and Physics*

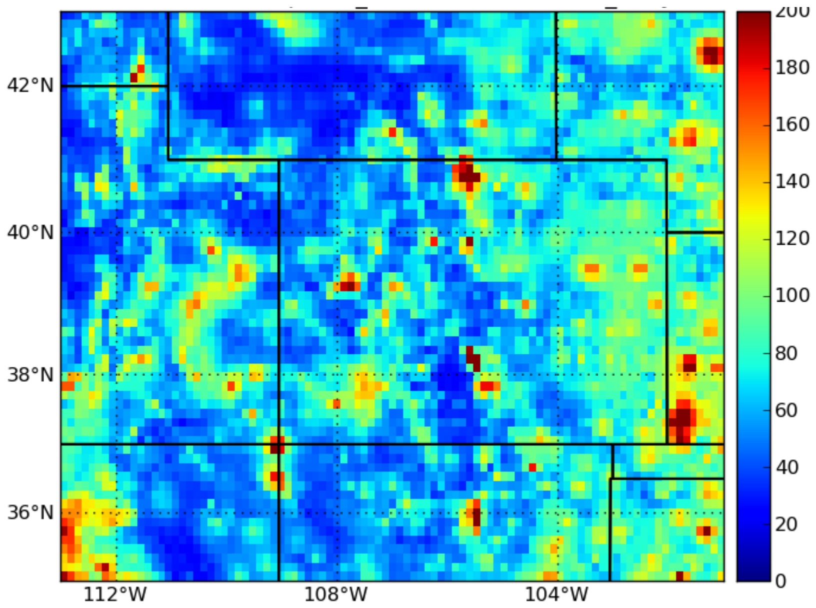
Observations

Polka-dot features due to the interpolation between station observations

Wet Day Fraction

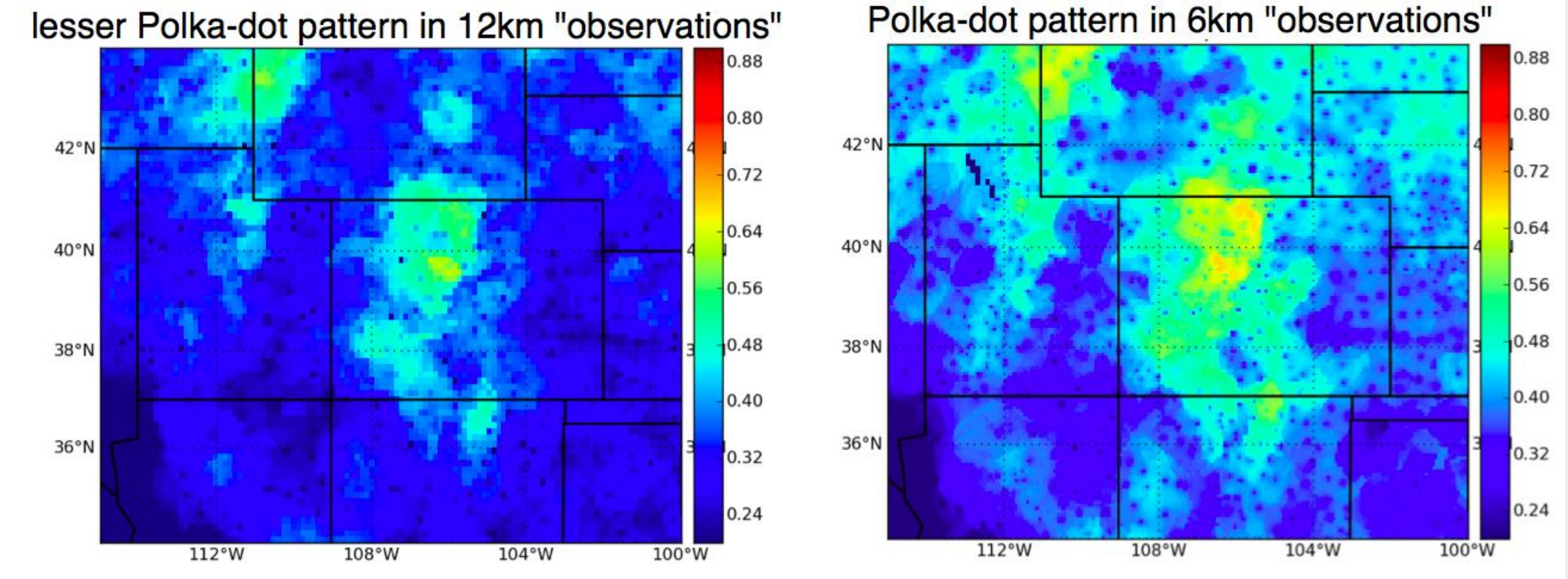


Extreme Events



Observations

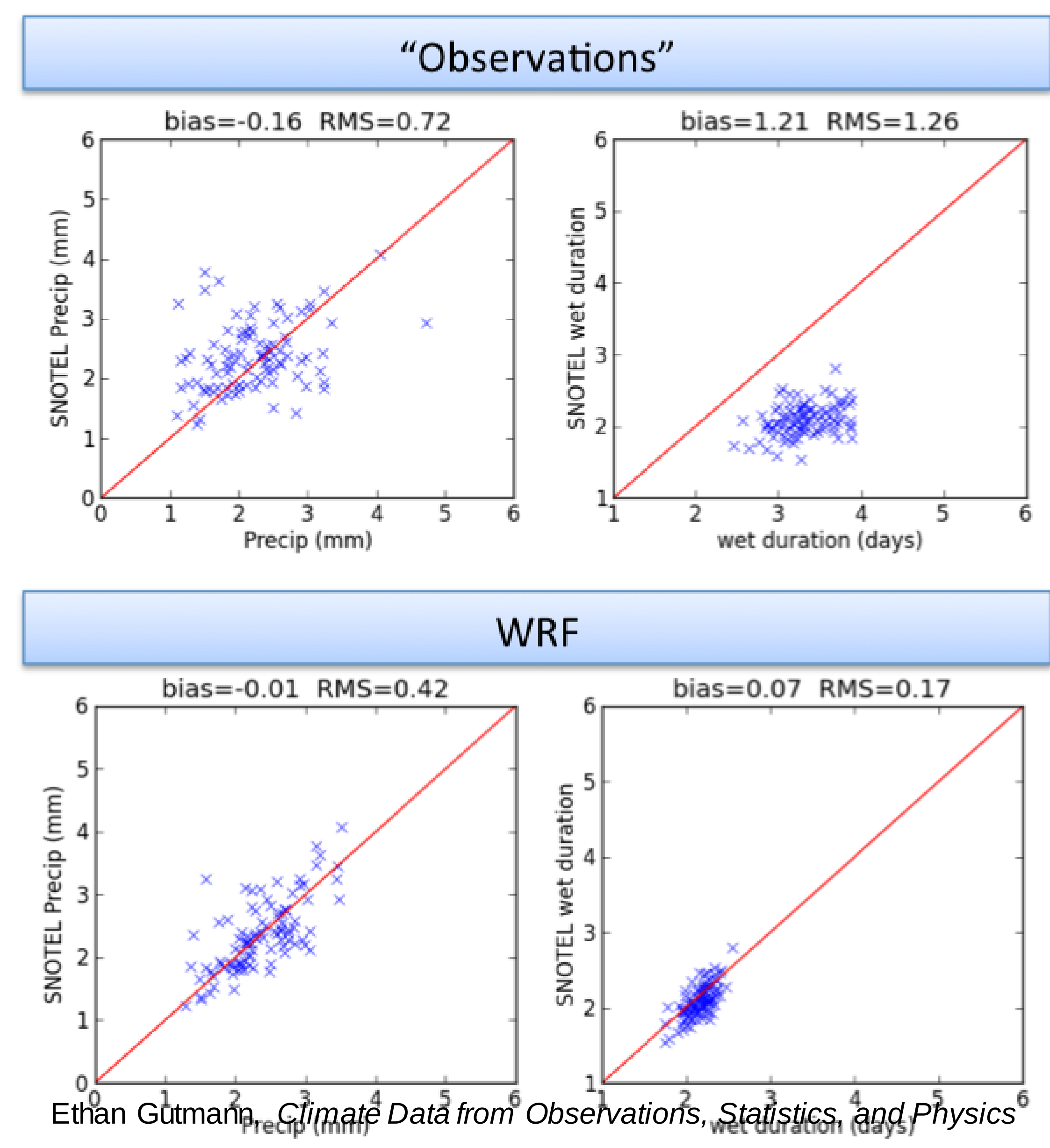
Smaller grid spacings are not necessarily better


lesser Polka-dot pattern in 12km "observations"
Polka-dot pattern in 6km "observations"
42°N
40°N
38°N
36°N
112°W
108°W
104°W
100°W
0.88
0.80
0.72
0.64
0.56
0.48
0.40
0.32
0.24

Gutmann et al. (submitted)

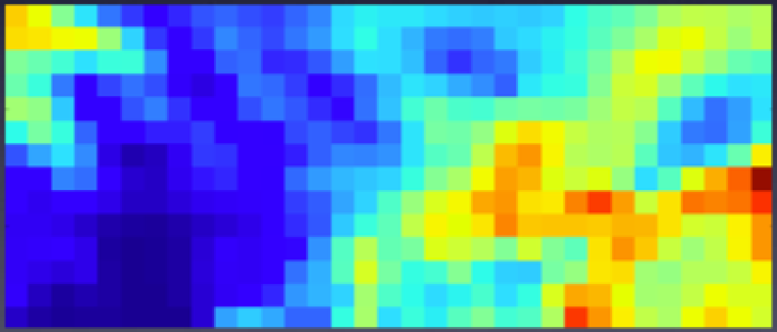
“Observations” vs Observations

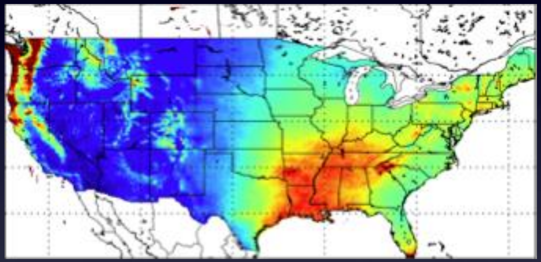
- “Observations” have too many wet days, and large errors in Precip totals
- WRF appears better in both regards...

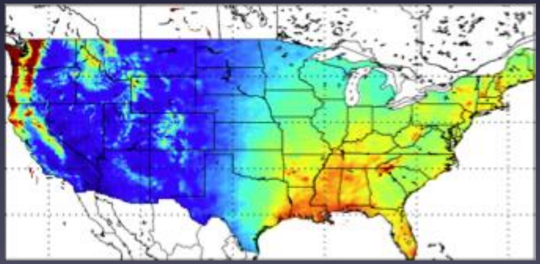


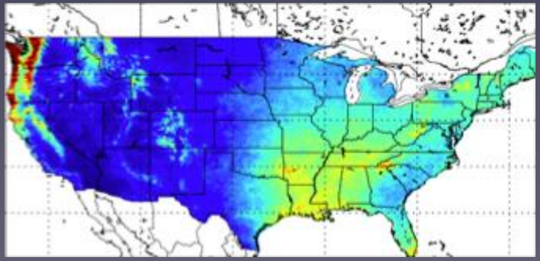
Statistics

- Climate model outputs are too coarse
- Dynamical downscaling is too expensive (for now)



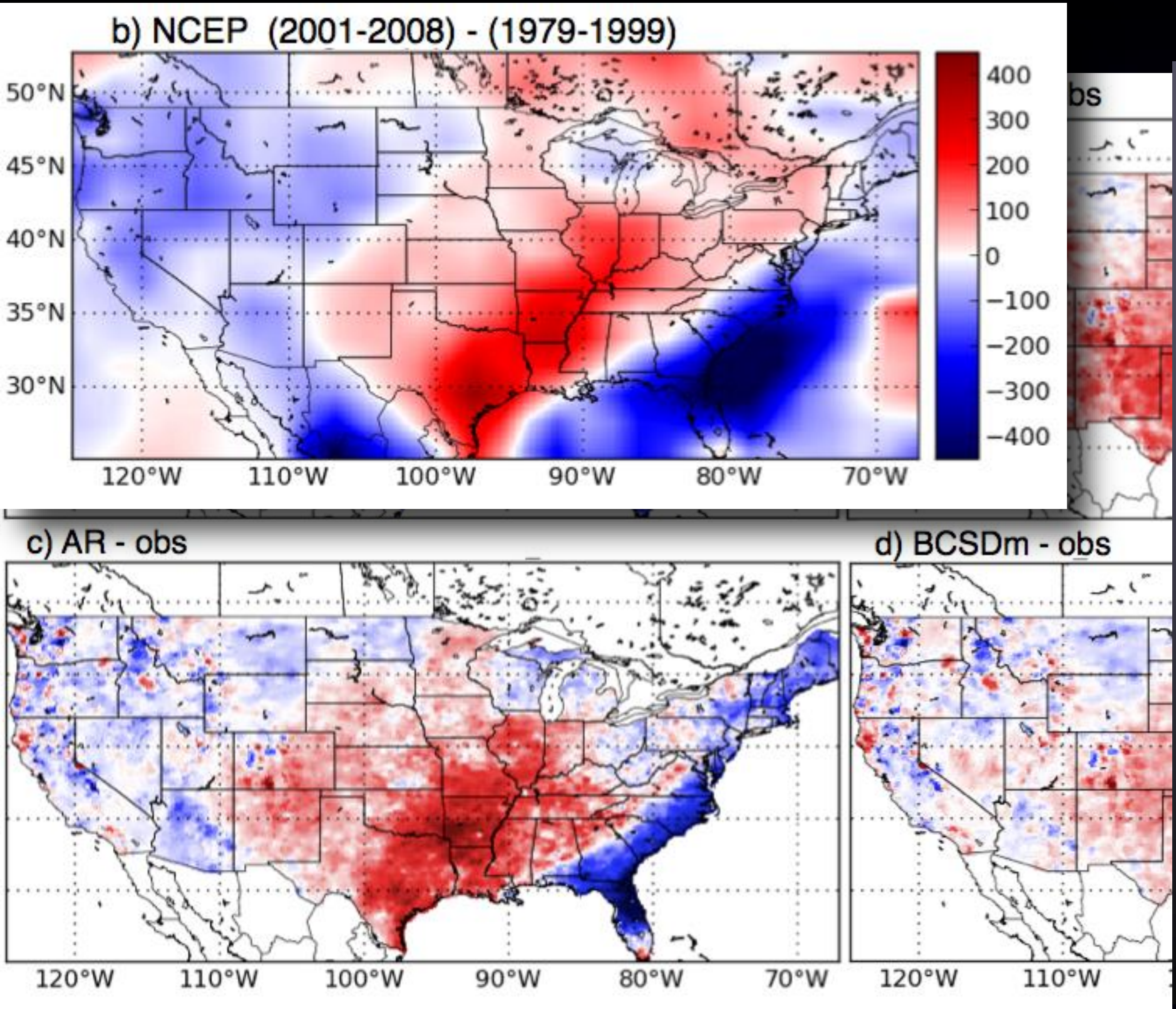






- Statistical downscaling is common
  … but what does that do to the data?

Bias : Large scales



- BCCA is biased low
- Other large scale biases due to changes in NCEP
- Reanalyses are not stable over time (Trenberth et al., 2011)
- Satellites and other assimilated datasets come and go

Gutmann et al. (submitted)

Example Output User Talk

TAMPA
BAY
WATER
Supplying Water To The Region

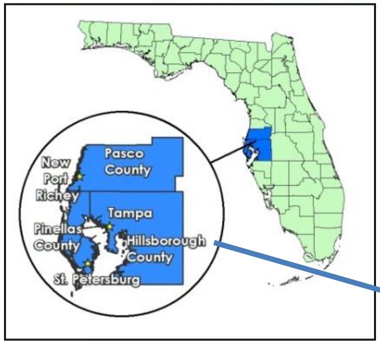
Use of Climate Projections for Water Supply Planning

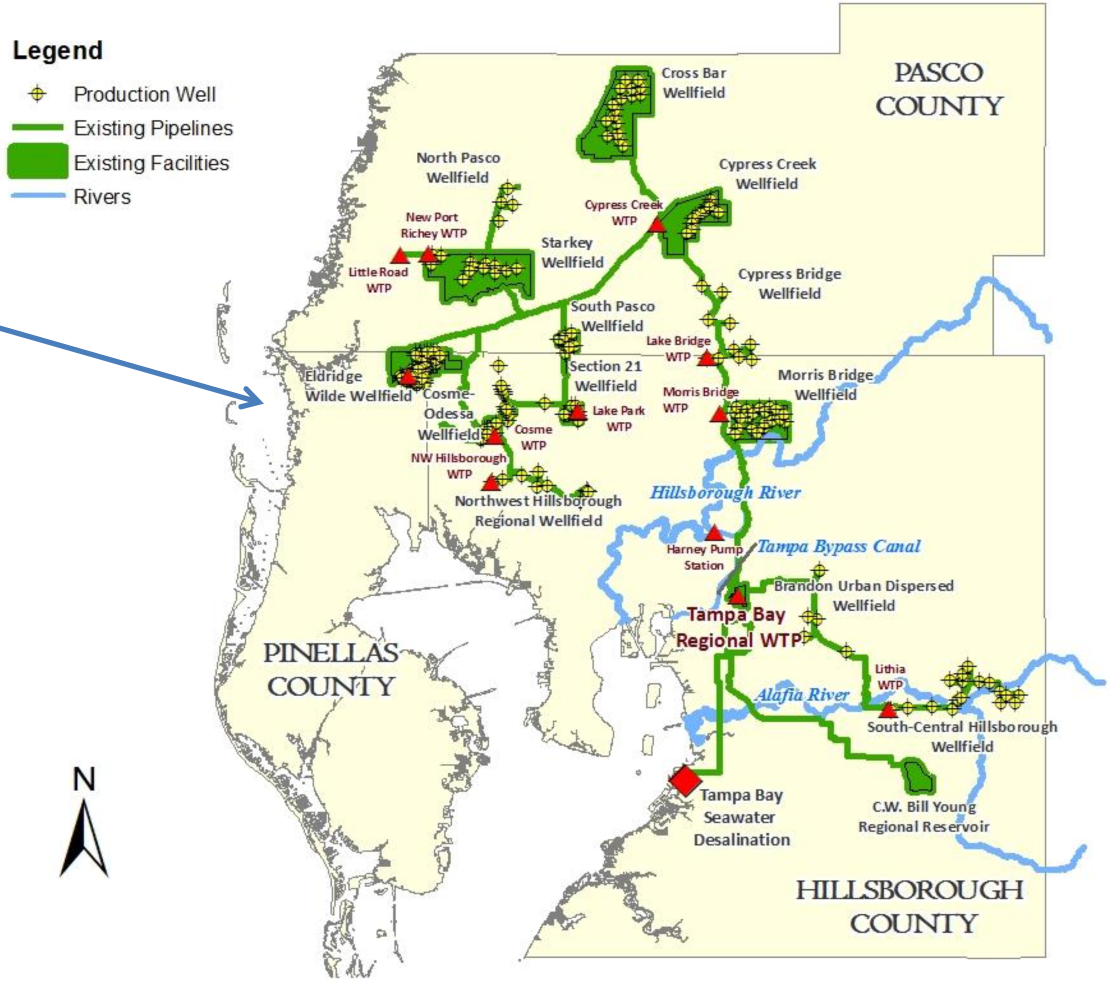
Alison Adams, Ph.D., P.E.

NCPP Workshop

August 12-16, 2013

Florida's Largest Regional Public Water Supplier





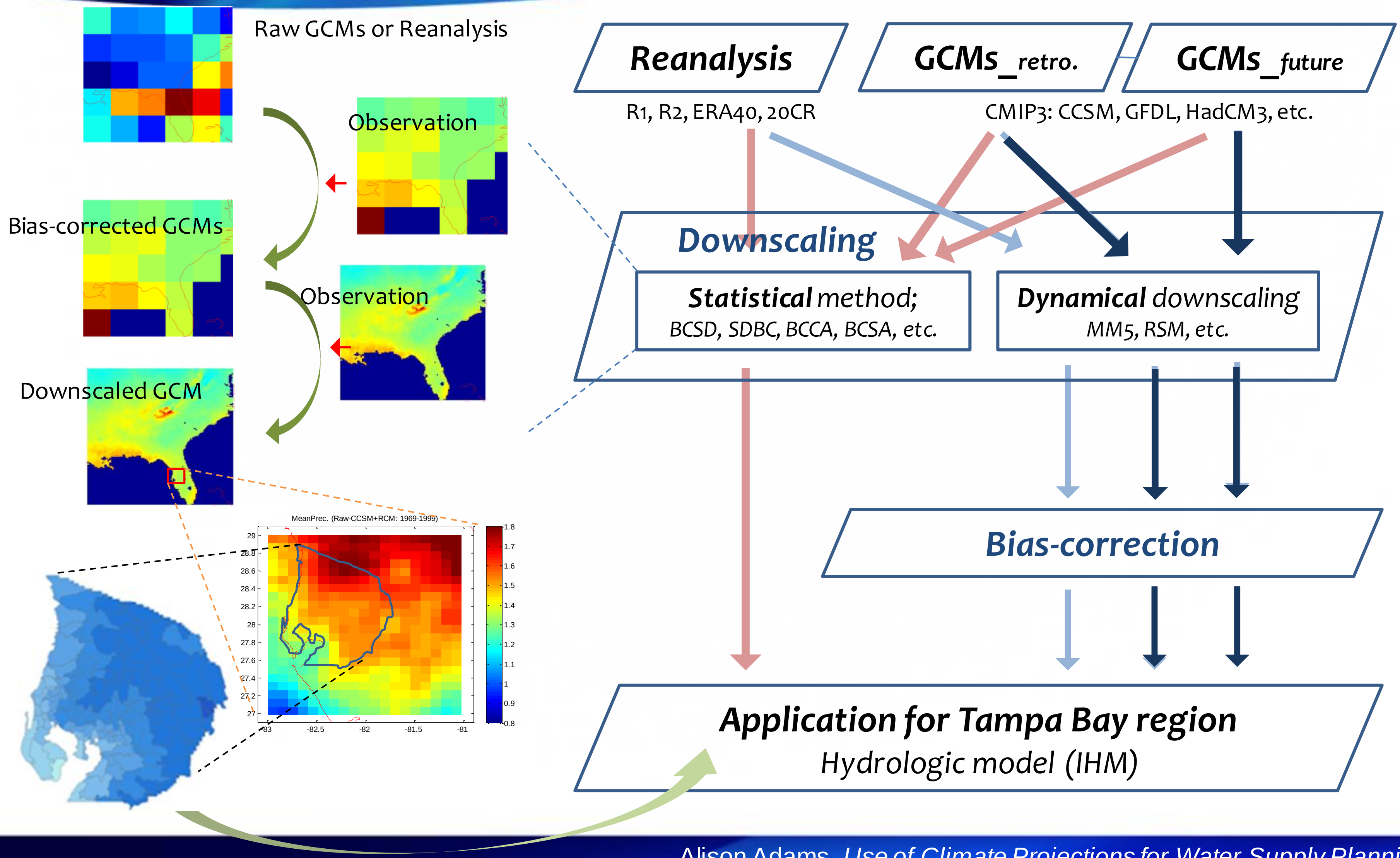
Wholesale drinking water to six governments

2.4 Million Residents

220-250 mgd annual average

Seasonal to multi-year variable climate

Water Institute Research



Reanalysis

R1, R2, ERA40, 20CR

GCMs_*retro.*

GCMs_*future*

CMIP3: CCSM, GFDL, HadCM3, etc.

Downscaling

Statistical *method;*
BCSD, SDBC, BCCA, BCSA, etc.

Dynamical *downscaling*
MM5, RSM, etc.

Bias-correction

Application for Tampa Bay region
Hydrologic model (IHM)

What we have done so far

1. Statistical downscaling
   - Comparative evaluation of 4 methods (BCSD_daily, BCCA, SDBC, BCSA)
     - Hwang and Graham (2013) Hydro. Earth Syst. Sci
   - Hydrologic simulation
     - Submitting to ASABE transaction
2. Evaluation of downscaled reanalysis data
     - R1+MM5 (Hwang et al., 2011)
     - R2+RSM (Stefanova et al., 2011)
     - ERA40+RSM (Stefanova et al., 2011)
     - 20CR+RSM (DiNapoli and Misra, 2012)
   - Hwang et al 2013 Reg. Environ Change

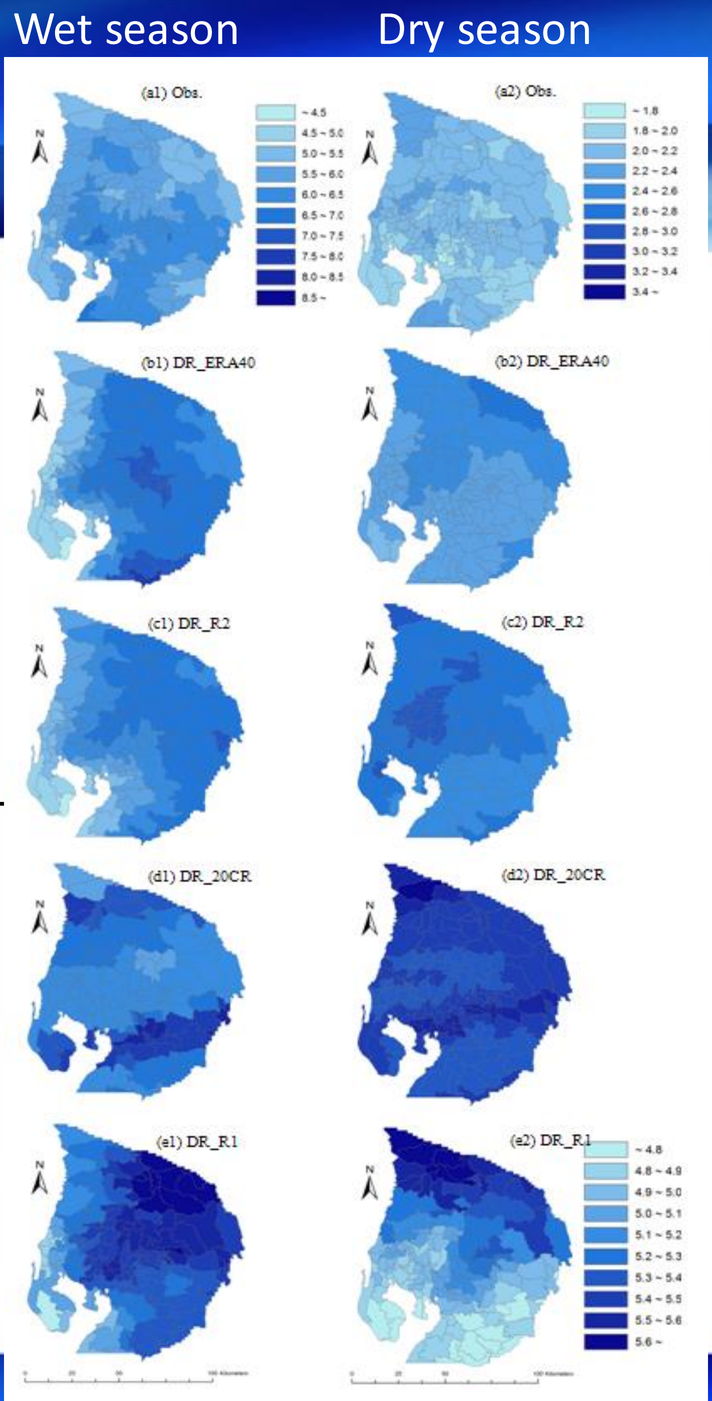
Bias-corrected reanalysis data for hydrologic model

- Study period from 1989 to 2001

1. R1+MM5 (Hwang et al., 2011)
   1986-2008
2. **R2+RSM (Stefanova et al., 2011)**
   1979-2001
3. ERA40+RSM (Stefanova et al., 2011
   1979-2001
4. 20CR+RSM (DiNapoli and Misra, 2012)
   1903-2008

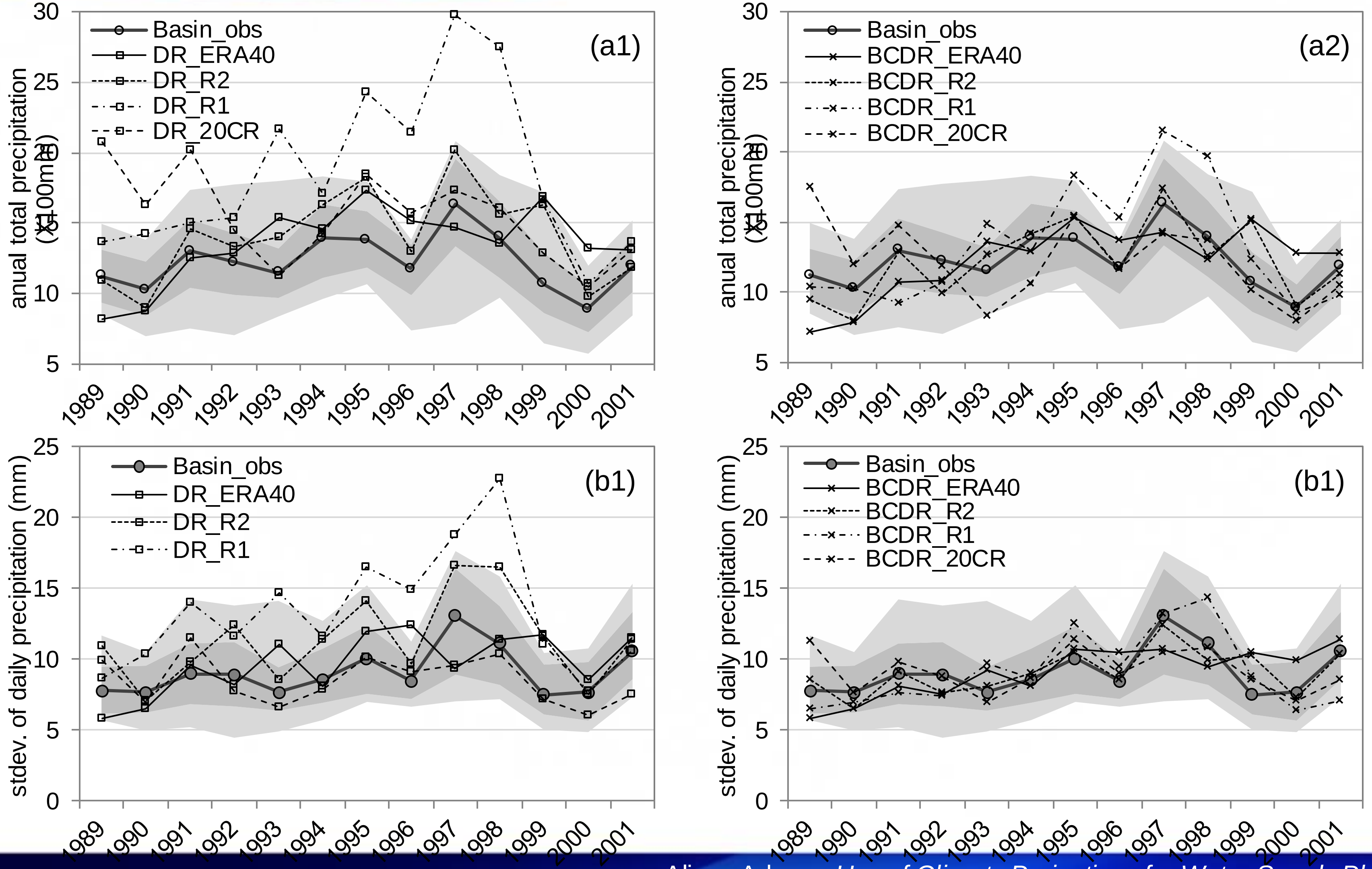
IHM calibration/verification period

1989-2006



Comparison of time series of (a) annual total precipitation and (b) standard deviation of daily precipitation over the year

Raw results ***Bias-corrected results***



NCPP QED Outcomes

- Next steps to make models and model outputs useful:
  - Domain specific “nutrition label” metadata standards
  - Education of output users
    - Basic rules of thumb
      - “Use a dozen models if you can”
      - “Don't pick the best models, cull the worst”
      - “Don’t go it alone”
    - Domain specific instruction
  - Foster a climate translator workforce
- Response
  - Modelers came to learn about how model outputs were used and user needs
  - Policy makers came to learn about model ouputs and limitations
  - Everyone went home happy

Benefits of Courting Reusers

- Data Producers
  - Effort goes towards a known reuser
  - Able to negotiate intellectual property concerns prior to data reuse
  - Averting misuse of data through education strategies
  - Ability to report reuse to funding agencies
- Data Reusers
  - Metadata to support access and interpretation
  - Trust via known data producers

How to Woo Data Reusers

- Identify a community to reuse data
- Engage with the community
- Establish common ground
  - Language
  - Assumptions
  - Incentives
  - Metadata
- Collaborate!

How can we help?

- Recommend this process
- Point them to Dr. Kim's presentation as an example
- We are well positioned to provide match-making services
  - Data producers come to us for data management help
  - Researchers come to us to find data

Thank You!

- The work presented here was generously funded by NSF award #0941386, *“Scaling Up: Introducing Commoditized Governance into Community Earth Science Modeling”*
- Members of CSDMS and NCPP, specifically Dr. James P. Syvitski and Dr. Richard B. Rood, for inviting and supporting participation in their respective events
- Dr. Wonsuck Kim, Dr. Ethan Gutmann, and Dr. Alison Adams of their slides.
- Dr. Paul N. Edwards and Dr. Christine L. Borgman for their research support

NSF

NCPP
National Climate
Predictions & Projections Platform

Earth System CoG

CSDMS
COMMUNITY SURFACE DYNAMICS MODELING SYSTEM