NSF

BIG WEATHER WEB

A common and sustainable big data infrastructure in support of weather prediction research and education in universities

Unidata Modeling Research in the Cloud Workshop, 5/31/17

UNIVERSITY OF CALIFORNIA SANTA CRUZ

UNIVERSITY OF NORTH DAKOTA

TEXAS TECH UNIVERSITY

SD MINES

UNIVERSITY of WISCONSIN UWMILWAUKEE

UNIVERSITY AT ALBANY State University of New York

PennState

Colorado State University

UCAR COMMUNITY PROGRAMS unidata Data Services and Tools for Geoscience

NCAR UCAR National Center for Atmospheric Research

Carlos Maltzahn Background

- **Adjunct Professor**, Computer Science, UC Santa Cruz
- **Director**, UCSC Systems Research Laboratory (SRL)
- **Director**, Center for Research in Open Source Software (CROSS) cross.ucsc.edu
- **Director**, UCSC/LANL Institute for Scalable Scientific Data Management (ISSDM)
- 1999-2004: **Performance Engineer**, Netapp
- **Advising** 6 Ph.D. students.
- **Graduated** 5 Ph.D. students
- **I do this 100% of my time!**

- Current Research
  - High-performance ultra-scale storage and data management
  - End-to-end Performance management and QoS
  - Reproducible Evaluation of Systems
  - Network Intermediaries
- Other Research
  - Data Management Games
  - Information Retrieval
  - Cooperation Dynamics

Baskin Engineering UC SANTA CRUZ

BIG WEATHER WEB

Project

- NSF-funded Scientific Software Integration project (SSI) of the Software Infrastructure for Sustained Innovation (SI2) program
- Goal: sustainable community SW framework

Collaborators

Carlos Maltzahn, Ivo Jimenez (UC Santa Cruz),

Josh Hacker, John Exby, Kate Fossell (NCAR),

Mohan Ramamurthy (Unidata),

Gretchen Mullendore, Timothy See (UND),

Brian Ancell (Texas Tech),

William Capehart (SDSM),

Clark Evans (UW Milwaukee),

Robert Fowell, Kevin Tyle (U Albany),

Steven Greybush (Penn State),

Russ Schumacher (CSU).

Problem

Informed by EarthCube Users workshops

- Poor reproducibility of data-intensive science
  - Impact on education and research
- Impaired availability of intermediate results
  - Unnecessary duplication of work, steep learning curves
- Communities of practice are falling behind
  - Limited ability to adopt new technologies

Domain:
Numerical Weather Prediction

- NWP groups at universities use supercomputing time to create large ensembles
- Current practice:
  - keep ensembles in scratch space or download to local infrastructure
  - Don’t share ensemble products, don’t share tools
  - Rewards for results, not data

General Approach

- Establish “nuclei”: pieces of technology that
  - Are easily shareable
  - Have the ability to grow & improve over time
  - Ensure “buy-in” from researchers and students
- Examples:
  - Wikipedia
  - Linux kernel
- Infrastructures to enable community-driven review and improvement

Big Weather Web Nuclei

1. Large ensemble distributed over 7 universities:

   Gretchen Mullendore (UND), Brian Ancell (Texas Tech), William Capehart (SDSM), Clark Evans (UW Milwaukee), Robert Fowell (SUNY Albany), Steven Greybush (Penn State), Russ Schumacher (CSU).

2. Common storage, linking, and cataloging methodology: Data Investigation and Sharing Environment
   - Permanent naming and high availability of data and experiments
   - Connecting data, platform, tools, analysis
3. Software Container technologies for easy deployment and reproducibility
   - Self-contained: software can be instantly deployed in common environments
   - Naming and versioning: compact reference mechanisms for complex environments
   - Good for reproducibility and education

BIG WEATHER WEB

Nucleus 1: Large, distributed ensembles

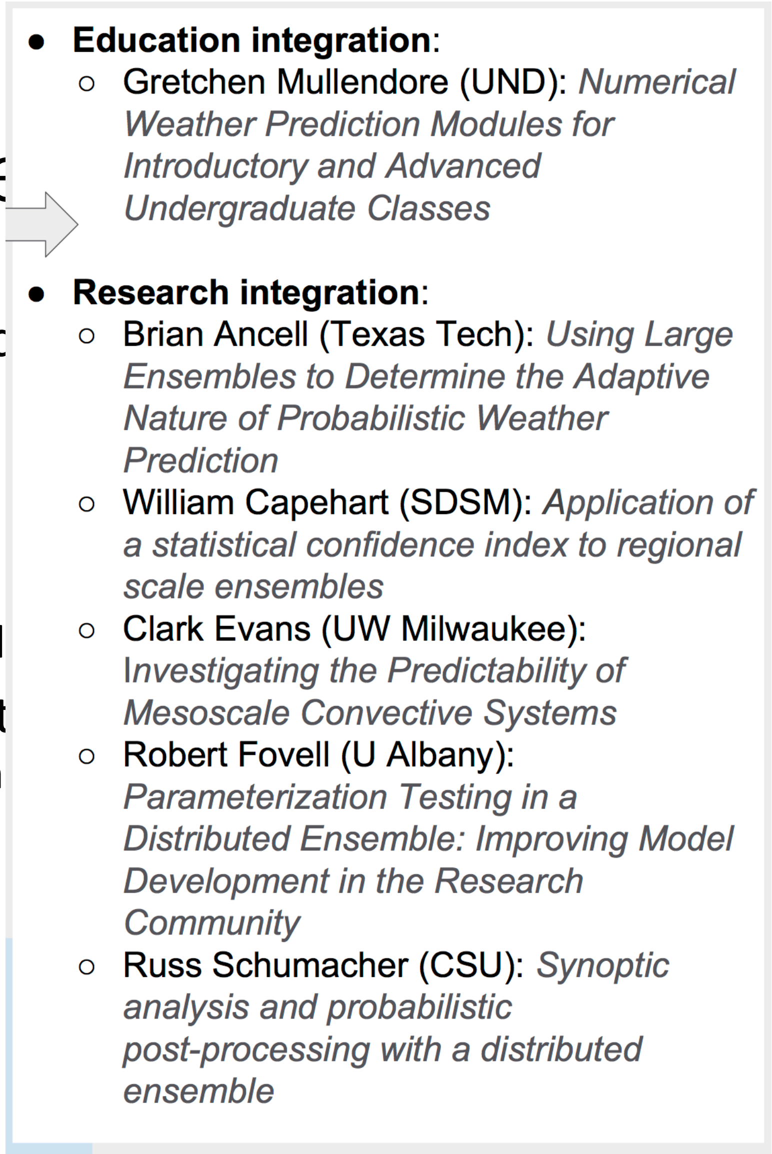
- Testing the distributed ensemble framework and tools
- Sharing of “knowledge products”
  - Initialization methods
  - Physics options
  - Workflow scripts for producing & analyzing data
  - Success: BWW Pis are using the BWW ensemble to do science
- Tracking data authorship and community impact
  - We have a DOI but access has to be managed (expense of data egress, see below)
  - Ensemble is evolving over time
- Dissemination of framework & tools
  - See NCAR’s “WRF in a box” work

BIG WEATHER WEB

Nucleus 1: Large, distributed e

- Testing the distributed ensemble framework anc
- Sharing of “knowledge products”
  - Initialization methods
  - Physics options
  - Workflow scripts for producing & analyzing data
  - Success: BWW PIs are using the BWW ensemble to d
- Tracking data authorship and community impact
  - We have a DOI but access has to be managed (expen
  - Ensemble is evolving over time
- Dissemination of framework & tools
  - See NCAR’s “WRF in a box” work

BIG WEATHER WEB



- **Education integration**:
  - Gretchen Mullendore (UND): *Numerical Weather Prediction Modules for Introductory and Advanced Undergraduate Classes*
- **Research integration**:
  - Brian Ancell (Texas Tech): *Using Large Ensembles to Determine the Adaptive Nature of Probabilistic Weather Prediction*
  - William Capehart (SDSM): *Application of a statistical confidence index to regional scale ensembles*
  - Clark Evans (UW Milwaukee): *Investigating the Predictability of Mesoscale Convective Systems*
  - Robert Fovell (U Albany): *Parameterization Testing in a Distributed Ensemble: Improving Model Development in the Research Community*
  - Russ Schumacher (CSU): *Synoptic analysis and probabilistic post-processing with a distributed ensemble*

Nucleus 2: Common storage, linking, and cataloging methodology

- Enable figures in publications & teaching materials to link to environments, tools, and data that produced them
- Provided in a form that is reusable
  - Easy install of environment and tools
  - Creation and access to data products without need to download everything
  - Data products by themselves link back to their antecedents in a reusable way.

BIG WEATHER WEB

Nucleus 2: Common storage, linking, and cataloging methodology

Use cloud services instead of on-premise installations

- Converts hard technical, management, and funding questions into just funding questions
- Started with commercial cloud: AWS ($9k/month credit)
  - 50TB so far on S3: $800/month
  - THREDDS server on EC2: $160/month
  - Particular thanks to John Exby and Kevin Tyle
- Challenges:
  - Commercial cloud: cost of data egress. Planning move to XSEDE/TACC/Wrangler
  - Long-term management of storage commons (with better-than-scratch-space policies)
  - Long-term naming: getting a DOI is the easy part -- long-term availability?

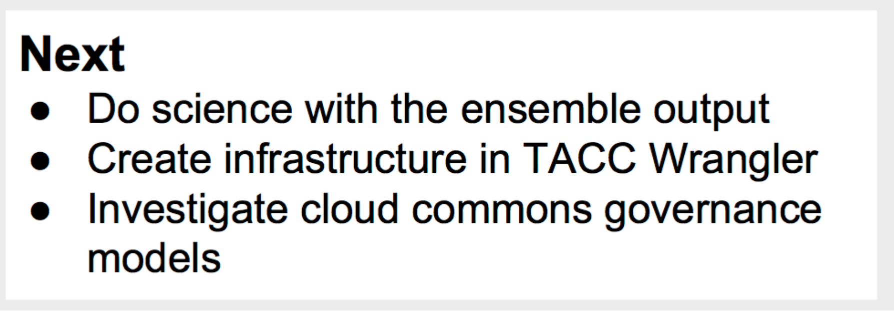
Nucleus 3: Software Containers

- See earlier talk ”Collaborative WRF-based research & education, enabled by software containers” by *Josh Hacker, John Exby, and Kate Fossell*
- Practical Falsifiable Research (Popper, falsifiable.us, see poster)
    - Apply open-source software community practices to experiment management
        - Script everything, leverage workflow systems & DevOps tools
            - See also Eric Klavins’ Aquarium Project: klavinslab.org/aquarium.html
        - Keep everything in git repositories
        - Use software containers for all software
    - Naming convention to automate running and validating experiments
    - Conventions for compact computing environment description
    - See poster by Ivo Jimenez

Popper
falsifiable.us

BWW Outreach

- 2015 Unidata Users Meeting
- 2015 AGU Townhall
  - ~50 attendees
  - 10 new bww-users subscribers
- 2016 Presentation at AMS
- 2016 Unidata Workshops
- WRF in a box in the class room
  - 2016 UND class by Tim See (UND)
  - 2017 UND class by Gretchen Mullendore (UND)
  - Wiliam Capehart (SDSM) using BWW ensembles (2 papers)
- Popper
  - Jimenez et al. VarSys'16, Chicago, IL
  - Jimenez et al. USENIX ;login:, Winter'16
  - Guest lecture in 2017 UND class by Gretchen Mullendore



Next

- Do science with the ensemble output
- Create infrastructure in TACC Wrangler
- Investigate cloud commons governance models

BIG WEATHER WEB

BIG WEATHER WEB

- Websites:
  - bigweatherweb.org
  - www.ral.ucar.edu/projects/ncar-docker-wrf
  - falsifiable.us
- Email list: bww-users@unidata.ucar.edu
- Slack: bigwxweb.slack.com, invitations: Kevin Tyle, ktyle@albany.edu
- Contact: Carlos Maltzahn, carlosm@ucsc.edu

UNIVERSITY OF CALIFORNIA SANTA CRUZ

UND UNIVERSITY OF NORTH DAKOTA

TEXAS TECH UNIVERSITY

SD MINES

UNIVERSITY of WISCONSIN UWMILWAUKEE

UNIVERSITY AT ALBANY State University of New York

PennState

Colorado State University

UCAR COMMUNITY PROGRAMS unidata Data Services and Tools for Geoscience

NCAR UCAR National Center for Atmospheric Research