

# **HPC and Big Data: COLA's Experience in the Advanced Scientific Discovery Program**

**Jim Kinter, COLA**

***Computing in the Atmospheric Sciences  
Annecy, France      12 September 2013***



# COLA News

## COLA Moves to GMU Fairfax Campus

*We are pleased to announce that the Center for Ocean-Land-Atmosphere Studies (COLA) will become an integral part of George Mason University (GMU) in 2013-14. COLA staff and the COLA computing facility will be collocated with Climate Dynamics faculty and students and the AOES Department on the GMU main campus in Fairfax, Virginia.*



GMU Research Hall – Home of AOES and COLA.

**James L. Kinter III,**  
Director, COLA

**Barry A. Klinger,**  
Graduate coordinator, AOES

**Jagadish Shukla,**  
Director, Climate Dynamics Program

**David M. Straus,**  
Chair, AOES



# GMU Ph.D. Program in Climate Dynamics

Affiliated with the Department of  
Atmospheric, Oceanic,  
and Earth Sciences

## Faculty

- **T. DelSole**; Ph.D., Harvard Univ.
- **P. Dirmeyer**; Ph.D., Univ. of Maryland
- **E. Jin**; Ph.D., Seoul National Univ.
- **B. Huang**; Ph.D., Univ. of Maryland
- **V. Krishnamurthy**; Ph.D., M.I.T.
- **J. Lu**; Ph.D., Dalhousie Univ.
- **J. Kinter**; Ph.D., Princeton Univ.
- **B. Klinger**; Ph.D., M.I.T./Woods Hole Ocean. Inst.
- **E. Schneider**; Ph.D., Harvard Univ.
- **P. Schopf**; Ph.D., Princeton Univ.
- **J. Shukla** (director); Ph.D., B.H.U.; Sc.D., M.I.T
- **C. Stan**; Ph.D., Colorado State Univ.
- **D. Straus** (chair, AOES); Ph.D., Cornell Univ.

# 5 Myths About Big Data

1. **There is a clear definition of Big Data.** (I don't know what it is, but I've got it!)
2. **Big Data is new.**
  - **Science has been using Big Data for a long time**, e.g., Kepler “mining” the obs of Brahe.
  - Statisticians: Big Data = Statistics, albeit sexier, more broadly applied
3. **Big Data is revolutionary.**
  - More likely to have modest, gradual impact.
  - Large effects are easy to recognize (small data), but handling subtleties require Big Data
4. **Bigger data is better.**
  - Big data sets are hard to work with, even using automated methods.
  - Bias can still be present in big data sets.
5. **Big Data means the end of science.**
  - Can't go fishing for correlations and explain the world, e.g., spurious correlations or conflated cause and effect
  - Still need hypotheses, ideas and theories: “If you don't ask good questions, your results can be silly and meaningless”

**“Having more data won't substitute for thinking hard, recognizing anomalies, and exploring deep truths.”**

Samuel Arbeson, *Wash. Post* (18 Aug. 2013)

# Predictability\* of the Physical Climate System

## Overarching Scientific Questions

What **limits predictability** at all time scales **from days to decades**? Is there a fundamental limit? What is the role of model error? Initial conditions error?

How do the **initial state**, the **coupling** of system components, and the **changes in external forcing contributes to predictability** at *different* time scales?

What aspects of the **total climate system** (troposphere, stratosphere, world oceans, land surface, vegetation, sea ice, land ice, snow) **are predictable** in which geographic regions, for which seasons, and how does that change in the future? For the current and future generation of climate models and observing systems?

What is the **optimal combination** of models to predict means? Extremes?  
No current models are perfect, e.g. for regional water cycle

\* **Note: Predictability is a necessary (but not sufficient) condition for attribution**

# Why does climate research need HPC and Big Data?

- **Societal demand for information** about weather-in-climate and climate impacts on weather
- **Seamless days-to-decades prediction & unified weather/climate modeling**
- **Multi-model ensembles and Earth system prediction**
- **Requirements for data assimilation**

# Driver: Societal Demand for Climate Information

- America's Climate Choices



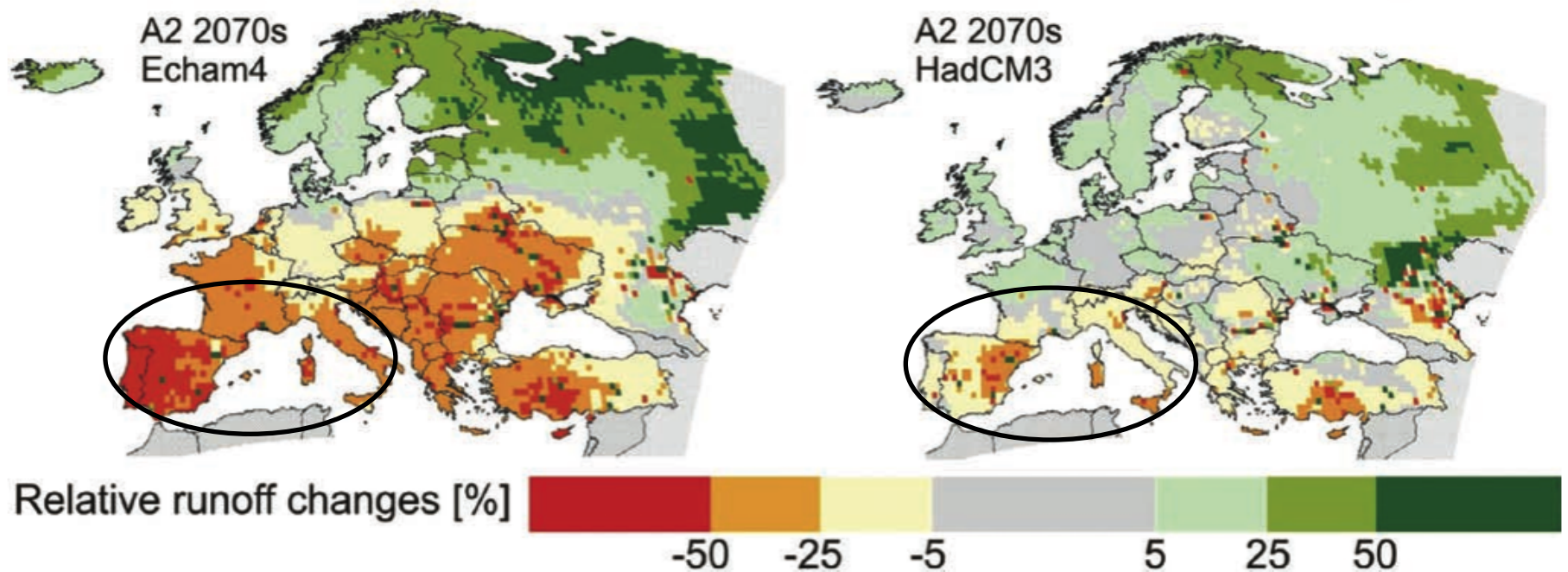
- (USGCRP) National Climate Assessment



- Intergovernmental Panel on Climate Change



# Regional Climate Change – Beyond CMIP3 Models' Ability?



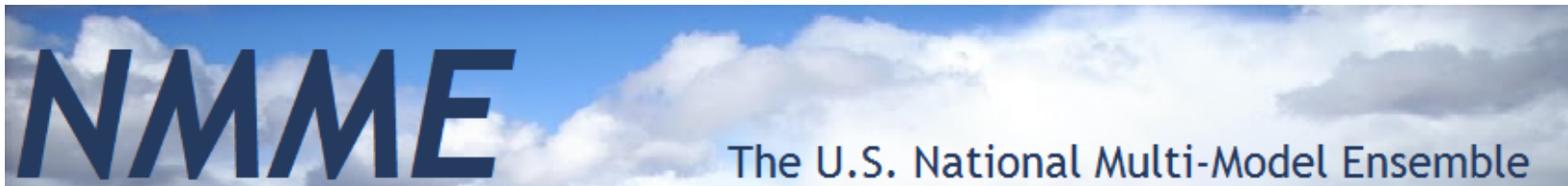


# Driver: Seamless Prediction, Unified Modeling

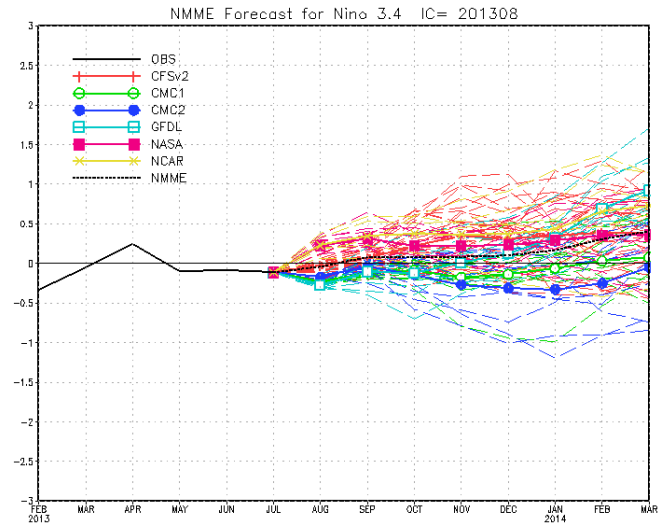
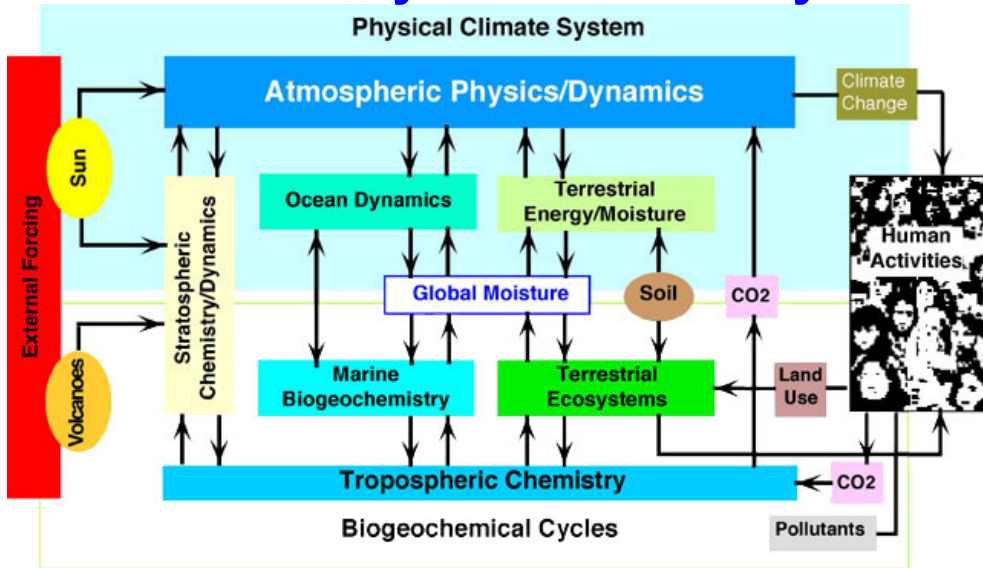
- **Seamless prediction**\*: Viewing **weather and climate prediction as initial-boundary value problems** that share common processes and dynamics and that can be addressed **using unified models** with common methods across a broad range of time scales and spatial resolutions.

\* Note: Prediction implies starting from an observed initial state, which in turn implies data assimilation

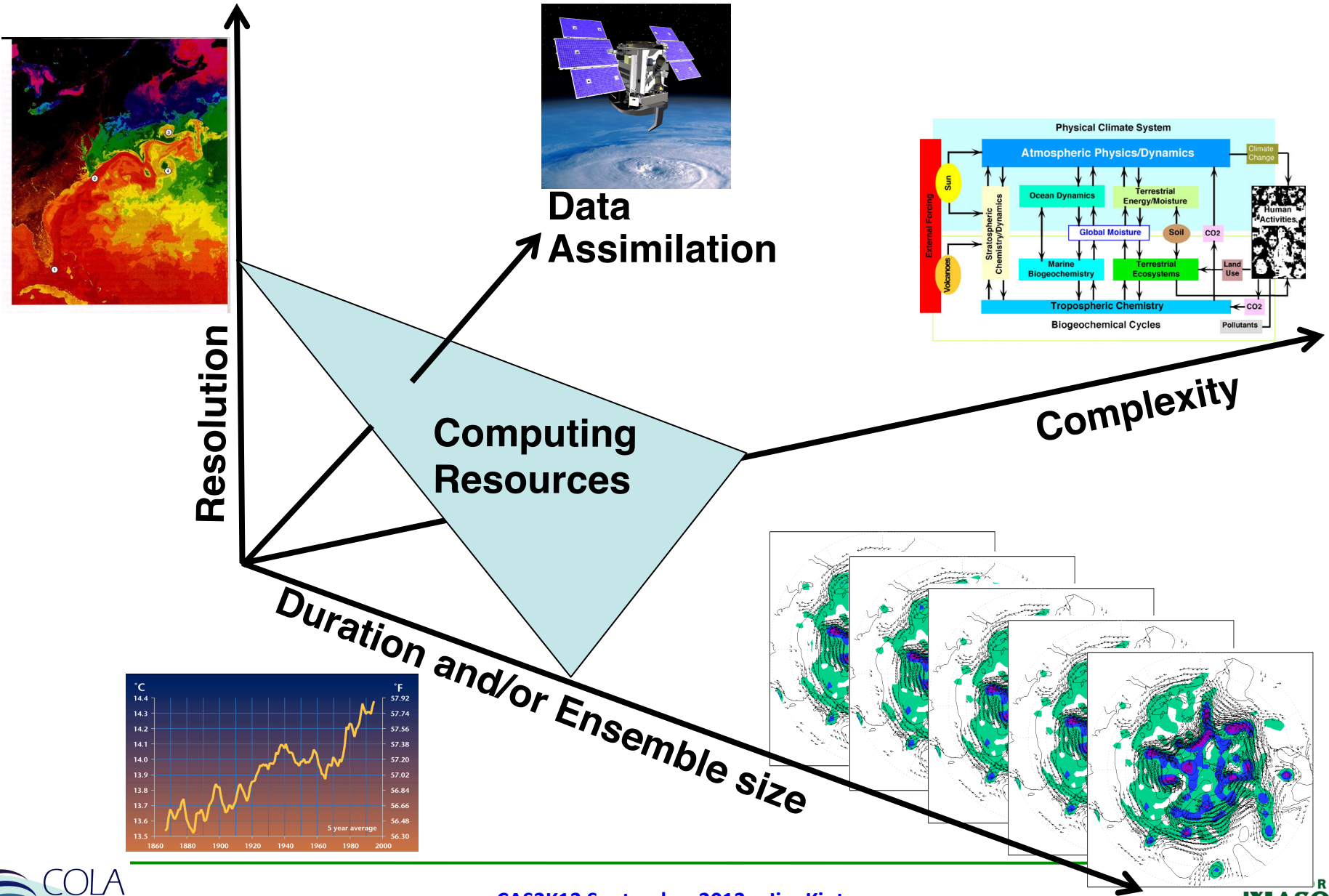
# Driver: Multi-Model Ensembles & Total Climate System Prediction



## Total Climate System – Earth System



# Balancing Demands on Resources



# HPC & Big Data at COLA

## Representative Projects

**Project Athena**: An International, Dedicated High-End Computing Project to **Revolutionize Climate Modeling** (Dedicated XT4 at NICS)

→ Update on CAS2K11 briefing from Martin Miller

**Project Minerva**: Exploring High Spatial Resolution for **Seasonal Climate Prediction** (Dedicated Advanced Scientific Discovery on NCAR Yellowstone)

**PetaApps Team**: Climate Models' Representation of **Unpredictable Noise** in the Atmosphere, Ocean or Sea Ice (TeraGrid Ranger and Kraken)

# Origins of *Project Athena*

- 2008 World Modeling Summit: **dedicate petascale supercomputers to climate modeling**
- U.S. National Science Foundation **offered to dedicate the Athena supercomputer for 6 months** in 2009-2010 as a pilot study
- An **international collaboration** (*Project Athena*) was formed by groups in the U.S., Japan and the U.K. to use Athena to take up the challenge



# Project Athena

## Collaborating Groups

**COLA** - Center for Ocean-Land-Atmosphere Studies, USA (NSF-funded)

**ECMWF** - European Center for Medium-range Weather Forecasts, UK

**JAMSTEC** - Japan Agency for Marine-Earth Science and Technology, Research Institute for Global Change, Japan

**University of Tokyo**, Japan

**NICS** - National Institute for Computational Sciences, USA (NSF-funded)

**Cray** Inc.

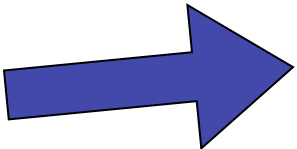
## Codes

**NICAM:** Nonhydrostatic Icosahedral Atmospheric Model

**IFS:** ECMWF Integrated Forecast System

# NICS Resources for *Project Athena*

- The Cray XT4 – **Athena** – the first NICS machine in 2008
  - 4512 nodes w/ AMD 2.3 GHz quad-core CPUs + 4 GB RAM
  - **18,048 cores** + 17.6 TB aggregate memory
  - **165 TFLOPS peak** performance
  - Dedicated to this project during October 2009 – March 2010 → 72 million core-hours!
- Other resources made available to project:
  - **85 TB Lustre file system**
  - **258 TB auxiliary Lustre file system** (called *Nakji*)
  - *Verne*: **16-core** 128-GB system (data analysis) during production phase (2009-2010)
  - *Nautilus*: SGI UV with **1024 Nehalem EX cores**, 8 GPUs, 4 TB memory, 960 TB GPFS disk (data analysis) in 2010-11



Many thanks to  
NICS for resources  
and sustained  
support!

# Project Athena Experiments

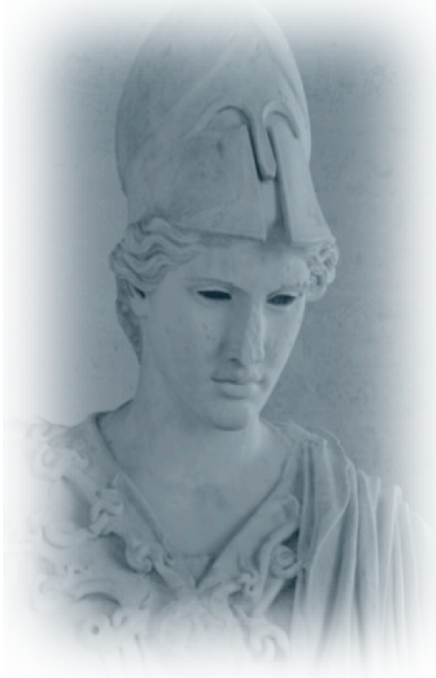
Model/Exp.	Resolution	# Cases	Period	Notes
NICAM / Hindcasts	7 km	8	103 days	21 May - 30 Aug 2001 - 2009
IFS / Hindcasts	125 km 39 km 16 km	48	395 days	1 Nov - 30 Nov (following year) 1960 - 2007
IFS / Hindcasts	10 km	20		1 Nov - 30 Nov (following year) 1989 - 2007
IFS / Hindcasts	125 km 39 km 16 km 10 km	9	103 days	21 May - 30 Aug 2001 - 2009 NICAM analogs
IFS / Summer Ensembles	39 km 16 km	6	132 days	21 May - 30 Sep selected years
IFS / Winter Ensembles	39 km 16 km	6	151 days	1 Nov - 31 Mar selected years
IFS / AMIP	39 km 16 km	1	47 years	1961 - 2007
IFS / Time Slice	39 km 16 km	1	47 years	2071 - 2117

<http://wxmaps.org/athena/home/>



# Project Athena Publications

- Dawson, A., T. N. Palmer and S. Corti, 2012: **Simulating regime structures in weather and climate prediction models**. *Geophys. Res. Lett.*, 39, doi:10.1029/2012GL053284
- Dirmeyer, P. A. and Co-Authors, 2012: **Evidence for enhanced land-atmosphere feedback in a warming climate**. *J. Hydrometeor.*, 13, 981-995.
- Dirmeyer, P. A. and Co-Authors, 2011: **Simulating the diurnal cycle of rainfall in global climate models: Resolution versus parameterization**. *Climate Dyn.* doi: 10.1007/s00382-011-1127-9.
- Jung, T. and Co-Authors, 2011: **High-Resolution Global Climate Simulations with the ECMWF Model in the Athena Project: Experimental Design, Model Climate and Seasonal Forecast Skill**. *J. Climate*, doi:10.1175/JCLI-D-11-00265.1.
- Kinter III, J. L. and Co-Authors, 2013: **Revolutionizing Climate Modeling – Project Athena: A Multi-Institutional, International Collaboration**. *Bull. Amer. Meteor. Soc.*, 94, 231-245.
- Manganello, J. V. and Co-Authors, 2012: **Tropical Cyclone Climatology in a 10-km Global Atmospheric GCM: Toward Weather-Resolving Climate Modeling**. *J. Climate* 25, 3867-3893.
- Miyamoto, Y., M. Satoh, H. Tomita, K. Oouchi, Y. Yamada; C. Kodama, J. L. Kinter III, 2013: **Gradient wind balance in tropical cyclones in high--resolution--global experiments**. *Mon. Wea. Rev.* (submitted).
- Palipane, E. and Co-Authors, 2013: **Improved Annular Mode Variability in a Global Atmospheric General Circulation Model with 16-km Resolution**. *J. Climate* (submitted).
- Satoh, M. and Co-Authors, 2011: **Intra-Seasonal Oscillation and its control of tropical cyclones simulated by high-resolution global atmospheric models**. *Climate Dyn.*, doi10.1007/s00382-011-1235-6.
- Solomon, A. and Co-Authors, 2013: **The distribution of U.S. tornado risk in a changing climate**. *J. Climate* (submitted).

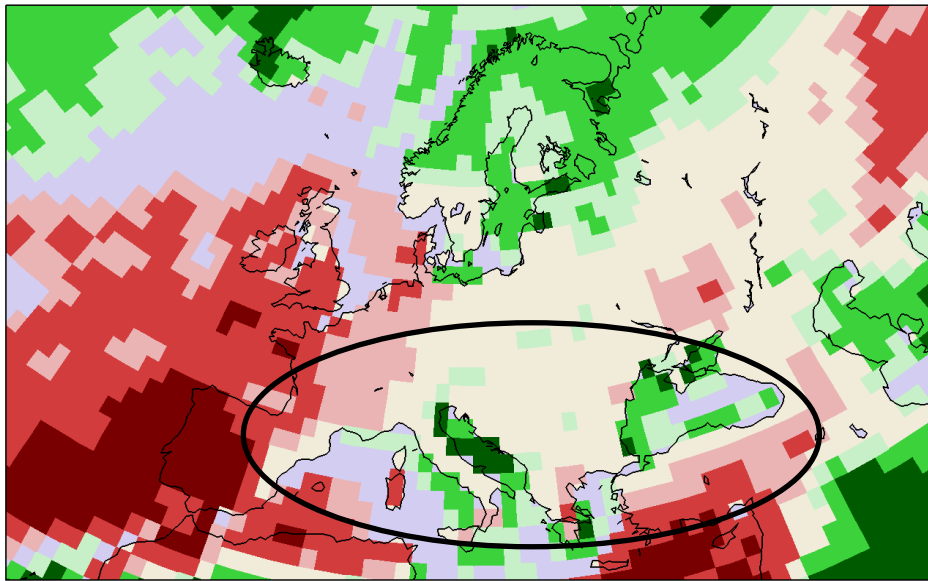


# Sample Results

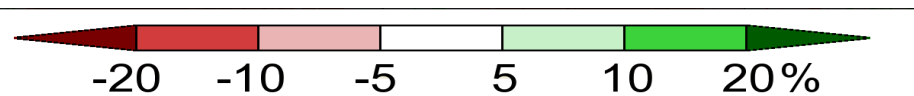
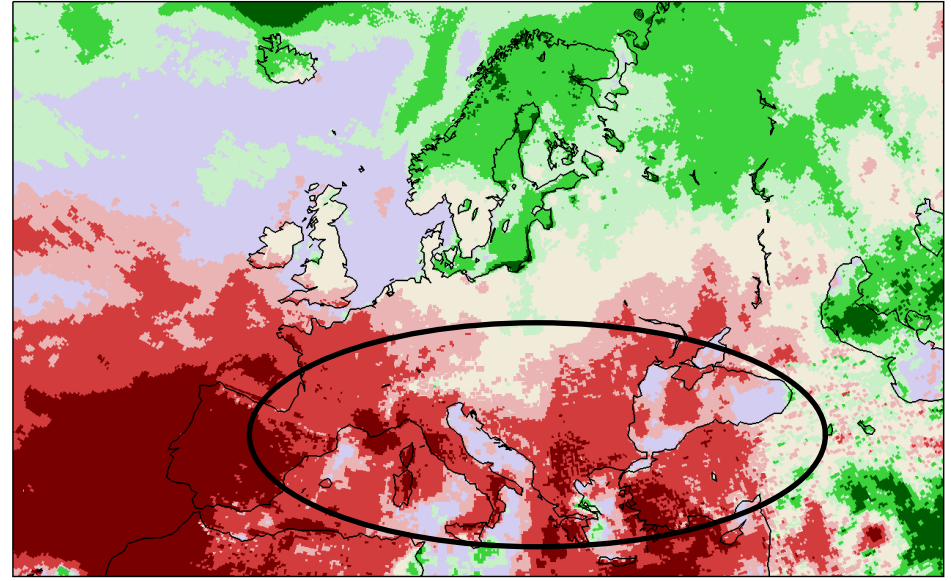
- Basics of model climate
- South Asian monsoon
- Resolution dependence of snow
- Diurnal cycle of precipitation
- **Projection of climate change**
- Tropical cyclones
- Tornadoes in climate simulation

# Europe Growing Season (Apr-Oct) Precipitation Change: 20<sup>th</sup> C to 21<sup>st</sup> C

T159 (125-km)



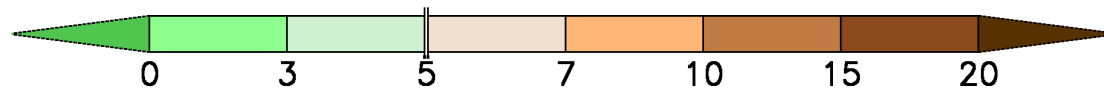
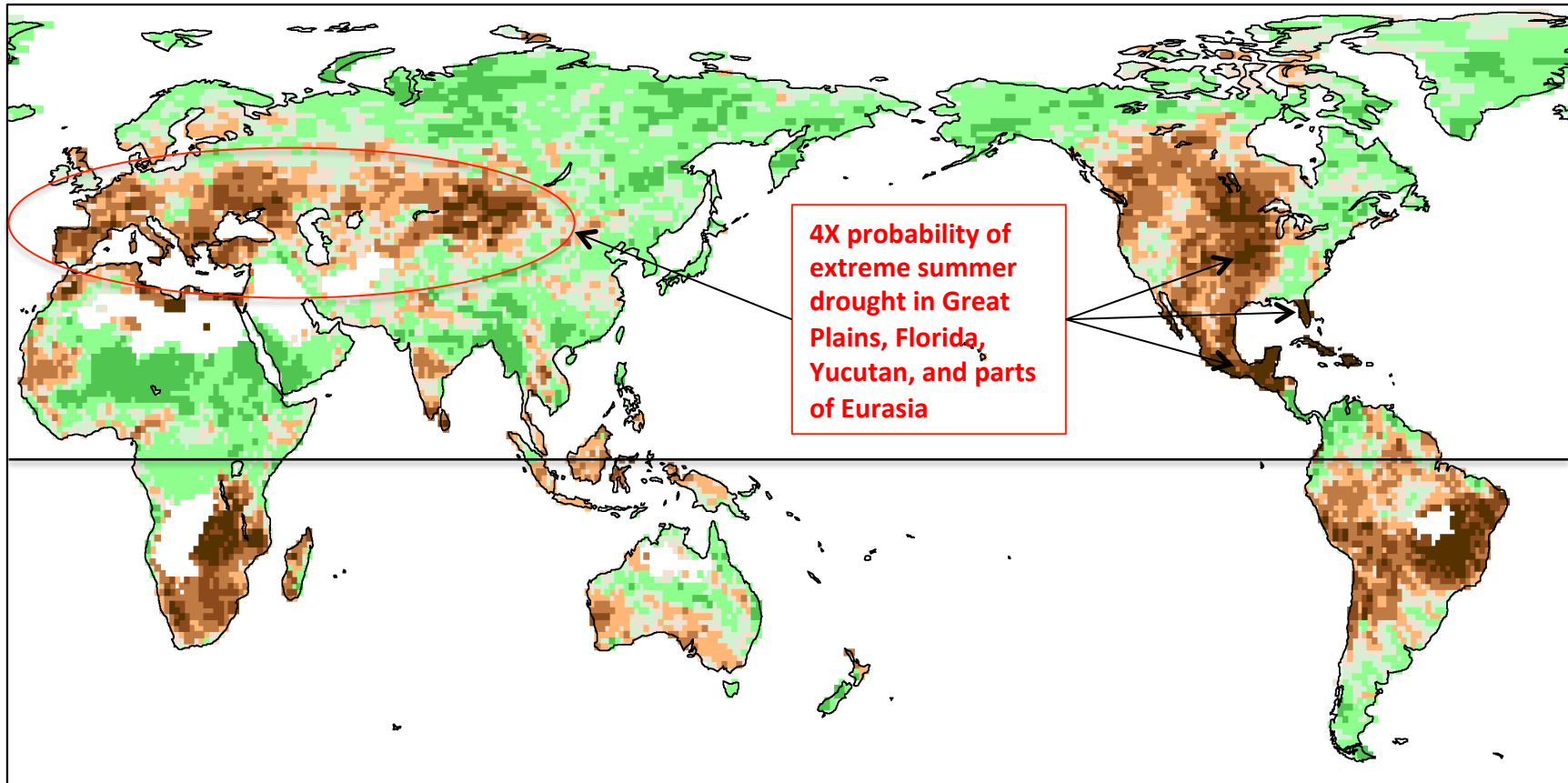
T1279 (16-km)



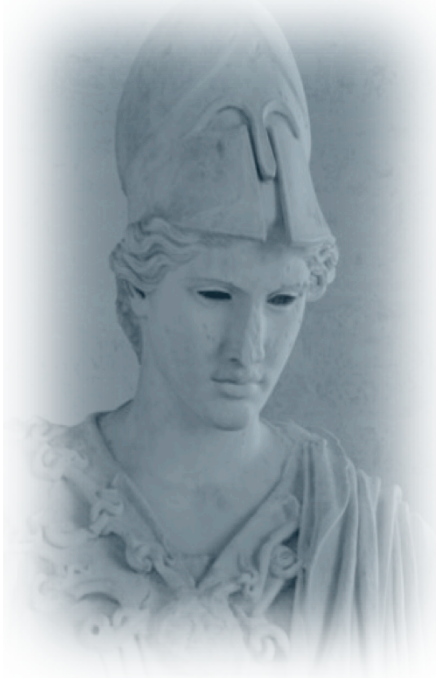
“Time-slice” runs of the ECMWF IFS global atmospheric model with observed SST for the 20<sup>th</sup> century and CMIP3 projections of SST for the 21<sup>st</sup> century at two different model resolutions

The continental-scale pattern of precipitation change in April – October (growing season) associated with global warming is similar, but the regional details are quite different, particularly in southern Europe.

# Future Change in Extreme Summer Drought Late 20<sup>th</sup> C to Late 21<sup>st</sup> C



**10<sup>th</sup> Percentile Drought:** Number of years out of 47 in a simulation of future climate (2071-2117) for which the June-August mean rainfall was less than the 5<sup>th</sup> driest year of 47 in a simulation of current climate (1961-2007).



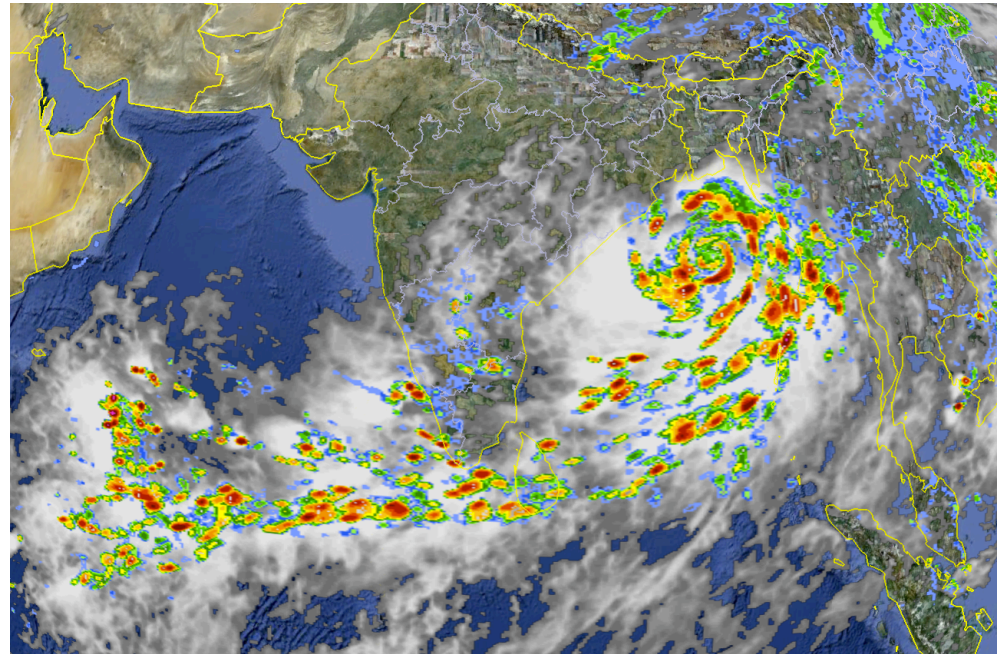
# Sample Results

- Basics of model climate
- South Asian monsoon
- Resolution dependence of snow
- Diurnal cycle of precipitation
- Projection of climate change
- **Tropical cyclones**
- Tornadoes in climate simulation

# Athena – Clouds and Precipitation

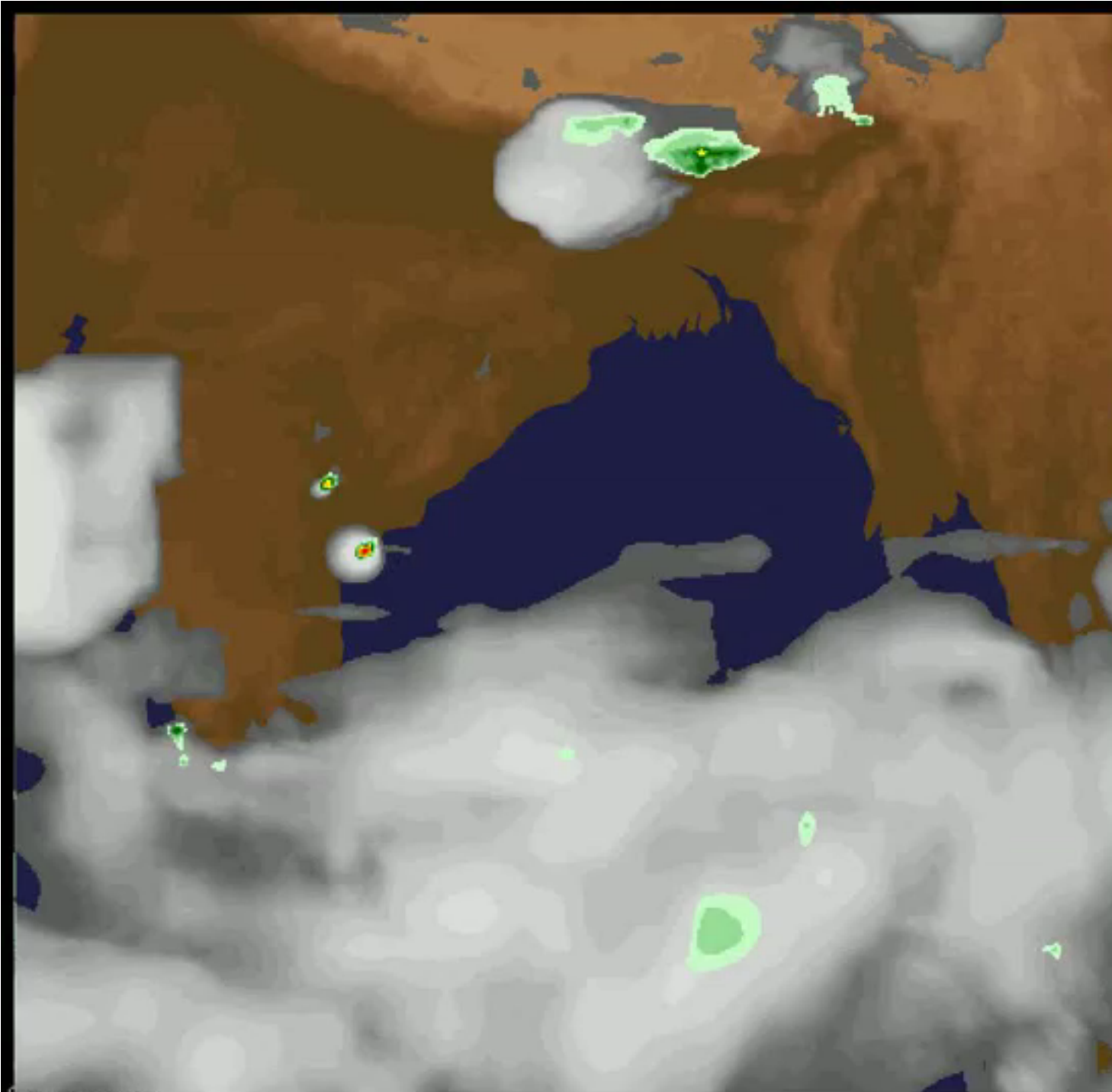
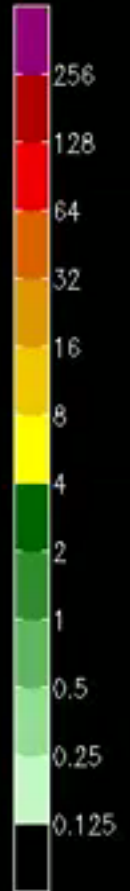


Boreal Summer 2009  
Brian Doty  
COLA



01Z MAY 21, 2009

mm/day





# *Project Athena: Summary*

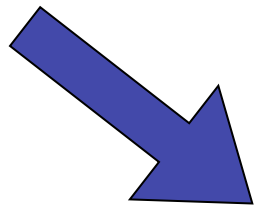
- **Good news:** Extreme spatial resolution **improves many of the qualitative features of large-scale climate simulation**
- **As expected:** High spatial resolution provides **higher fidelity representation of features sensitive to orography or geography**
- **Unexpected: Nonlinear dynamical effects** can alter simulation changes due to spatial resolution improvements much more and possibly in different ways than we might have expected
- **Bad news** (as expected?): **Large biases remain** in hard-to-simulate fields like tropical precipitation → still need to understand and properly represent the effects of subgrid-scale physical processes





# Project Minerva

- Explore the impact of **increased atmospheric resolution** on model fidelity and prediction skill in a ***coupled, seamless framework*** by using a state-of-the-art coupled operational long-range prediction system to systematically evaluate the prediction skill and reliability of a robust set of hindcast ensembles at low, medium and high atmospheric resolutions
- **NCAR Advanced Scientific Discovery Program** to inaugurate *Yellowstone* (72 K-core IBM iDataPlex)
- Allocated 21 M core-hours on Yellowstone
- **Used ~28 M core-hours** (Our jobs squeaked in under core size that “broke” the system)



**Many thanks to  
NCAR for  
resources and  
sustained  
support!**



# Project Minerva

## ECMWF team:

- Frederic Vitart, lead
- Roberto Buizza
- Erland Kallen
- Franco Molteni
- Tim Stockdale
- Peter Towers
- Nils Wedi


## COLA team:

- Ben Cash, lead
- Rondro Barimalala
- Paul Dirmeyer
- Mike Fennessy
- V. Krishnamurthy
- Julia Manganello
- David Straus

## University of Oxford:

- Tim Palmer
- Andrew Dawson

# ECMWF Coupled Ensemble Systems


System	Atmosphere model cycle	Atmosphere spectral truncation	Atmosphere vertical levels	Ocean model	Ocean horizontal res, equatorial refinement	Ocean vertical levels
 MINERVA	IFS cy 38r1	T319 / T639 / T1279	91 levels, top = 1 Pa	NEMO v 3.0/3.1	1 degree, ~ 0.3 deg. Lat	42 levels
System 4	IFS cy 36r4	T255	91 levels, top = 1 Pa	NEMO v 3.0/3.1	1 degree, ~ 0.3 deg. Lat	42 levels
ENS (current)	IFS cy 38r2	T639 (d 0-10), T319	62 levels, top = 5 hPa	NEMO v 3.0/3.1	1 degree, ~ 0.3 deg. Lat	42 levels
ENS (end 2013)	IFS cy 40r1	T639 (d 0-10), T319	91 levels, top = 1 Pa	NEMO v 3.4	1 degree, ~ 0.3 deg. Lat	42 levels

**System 4:** Operational seasonal prediction system

**ENS:** Operational medium-range/monthly prediction system

Courtesy Franco Molteni & Frederic Vitart, ECMWF

# ECMWF Coupled Ensemble Systems

System	Coupler	Time range of ocean-atmosphere coupling	Coupling frequency	Unperturbed initial cond. for re-forecasts	Atmospheric perturbations	Ocean perturbations	Stochastic model perturbations
 MINERVA	OASIS-3	from start	3 hours	ERA-Interim + ORA-S4	SV, EDA from 2011 dates	5 ocean analyses + SST perturbations	3-timescale SPPT + KE backscatter
System 4	OASIS-3	from start	3 hours	ERA-Interim + ORA-S4	SV	5 ocean analyses + SST perturbations	3-timescale SPPT + KE backscatter
ENS (current)	OASIS-3	from day 10	3 hours	ERA-Interim + ORA-S4	SV, EDA from current or recent date	generated by ENS member fluxes during day 1 to 10	2-timescale SPPT + KE backscatter
ENS (end 2013)	sequential, single executable	from start	3 hours	ERA-Interim + ORA-S4	SV, EDA from current or recent date	5 ocean analyses	2-timescale SPPT + KE backscatter

**ORA-S4** : Ocean Re-Analysis for ECMWF System-4

**EDA** : Ensemble of Data Assimilations (low-res 4D-var)

**SV** : Singular Vectors of 48-hour linear propagator

**SPPT** : Stochastic Perturbation of Physical Tendencies scheme

**Courtesy Franco Molteni & Frederic Vitart, ECMWF**



# Minerva Prediction Experiments

Experiment	Years	Ens. Size	Initial Months	Duration (mon)
T319_base	1980-2011	51	May, Nov	7
T319_2_year_extension	1980-2011	15	May	24
T639_base	1980-2011	15	May, Nov	7
T639_extended_ensemble	1980-2011	36	May, Nov	May: 5 mo Nov: 4 mo
T639_2_year_extension	1980-2011	15	Nov	24
T1279_base	2000-2011	15	May	7

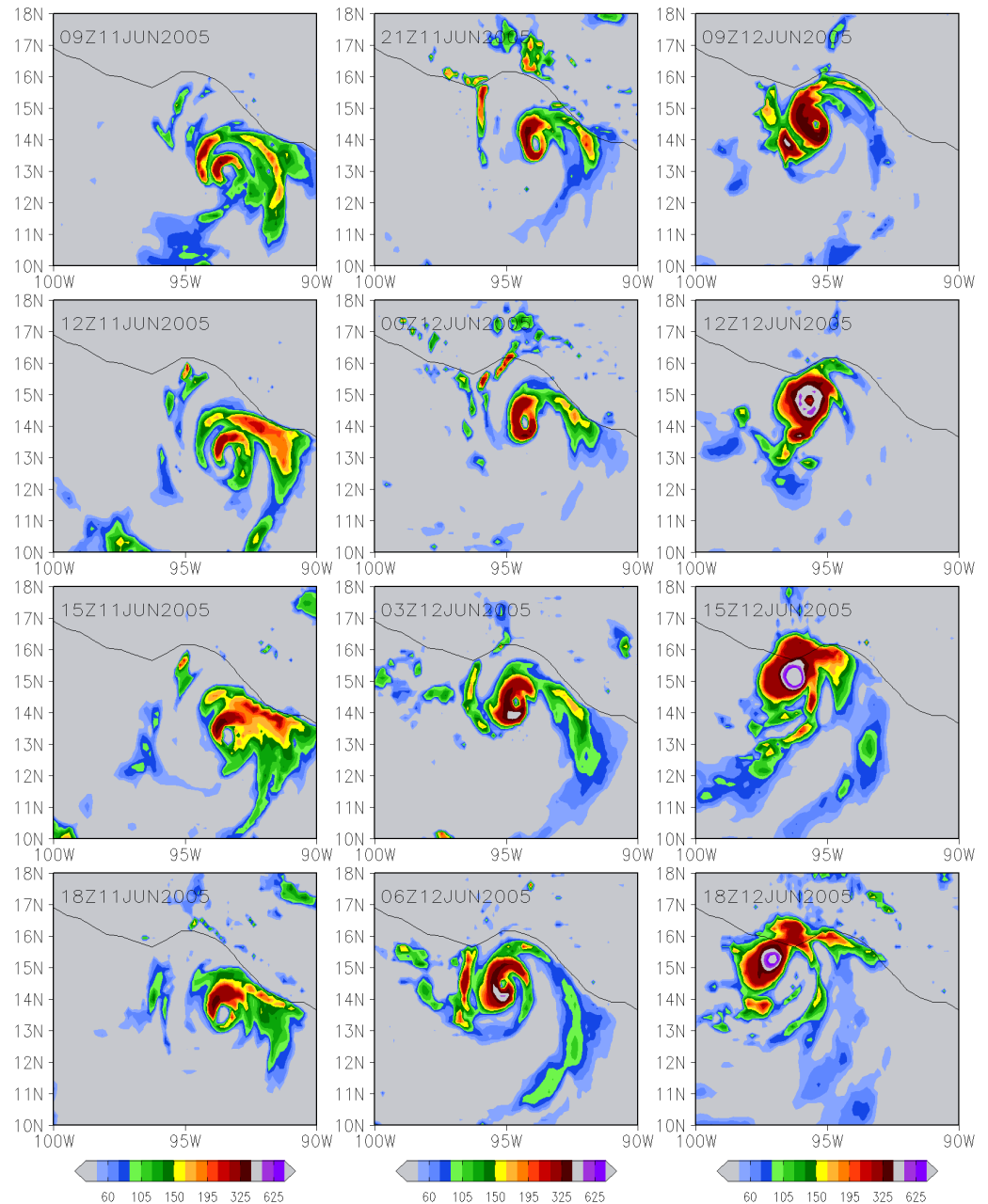


# Minerva: Coupled Prediction of Tropical Cyclones

11-12 June 2005 hurricane off west coast of Mexico: precipitation in mm/day every 3 hours (T1279 coupled forecast initialized on 1 May 2005)

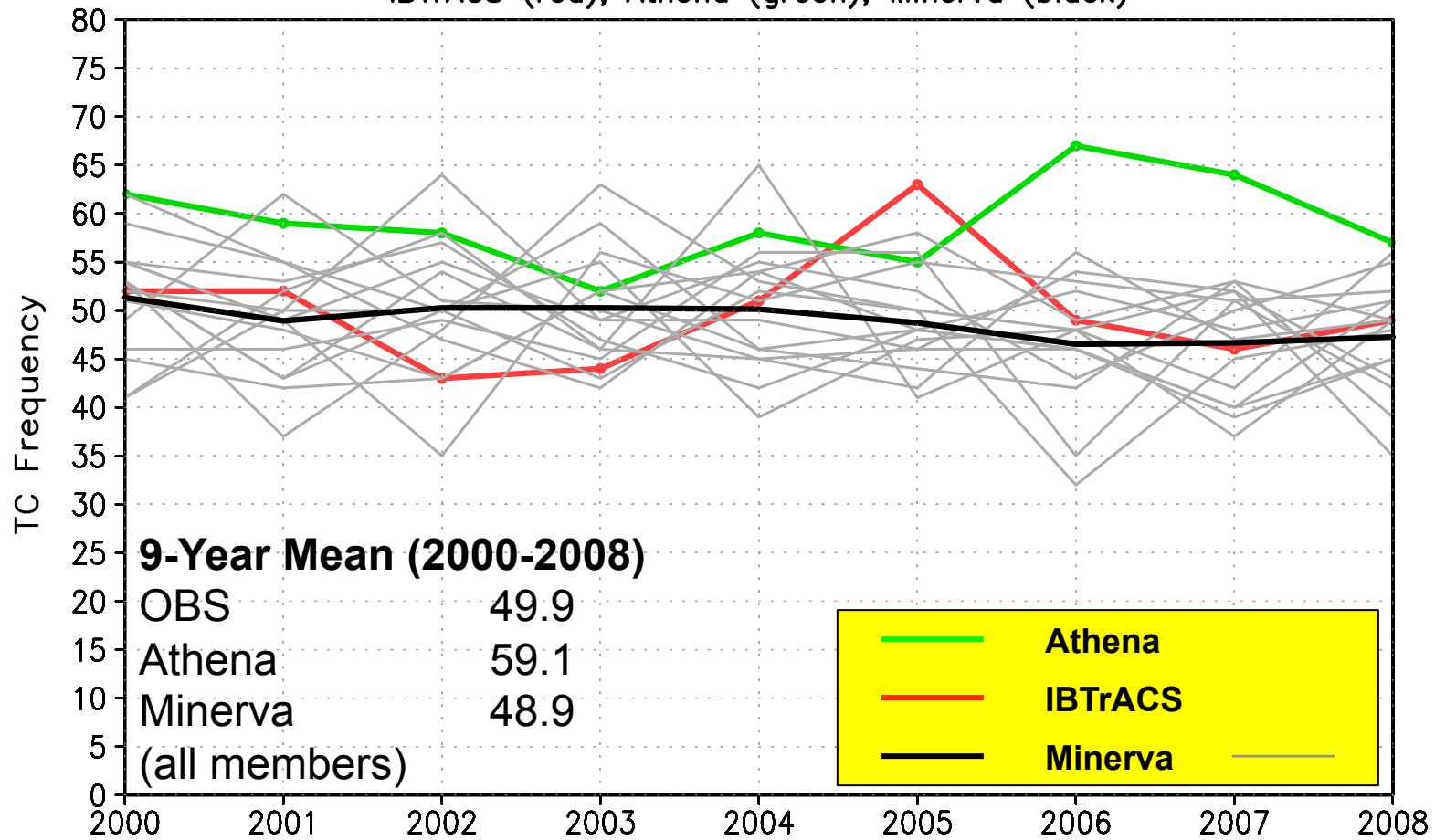
The predicted maximum rainfall rate reaches 725 mm/day (30 mm/hr)

Based TRMM global TC rainfall observations (1998-2000), the frequency of rainfall rates exceeding 30 mm/hr is roughly 1%



# Minerva vs. Athena – TC Frequency (NH; JJASON; T1279)

TC Frequency for the NH, JJASON 2000–2008, T1279, Ident. II  
IBTrACS (red), Athena (green), Minerva (black)



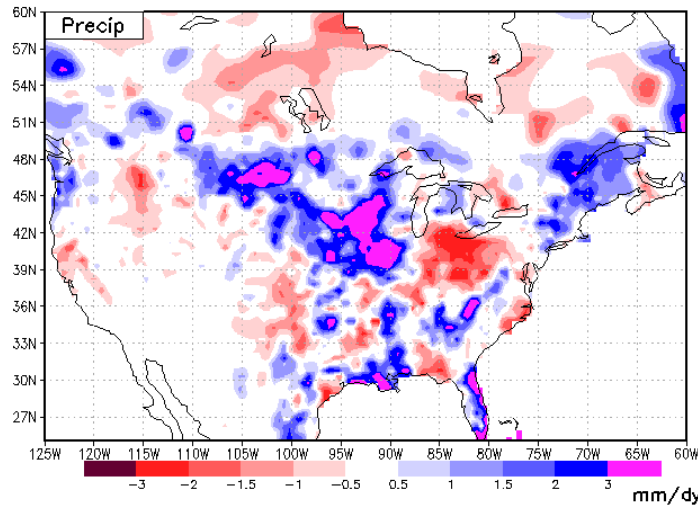
# Individual Forecast Anomalies



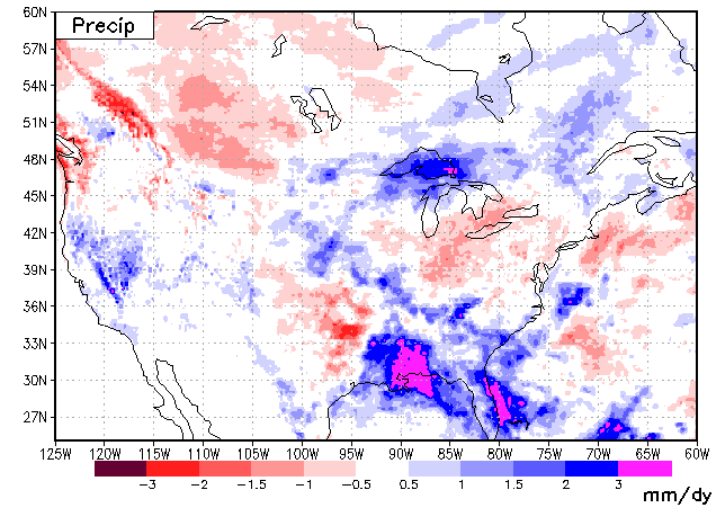
**MAY 2013**

Precipitation

Observations

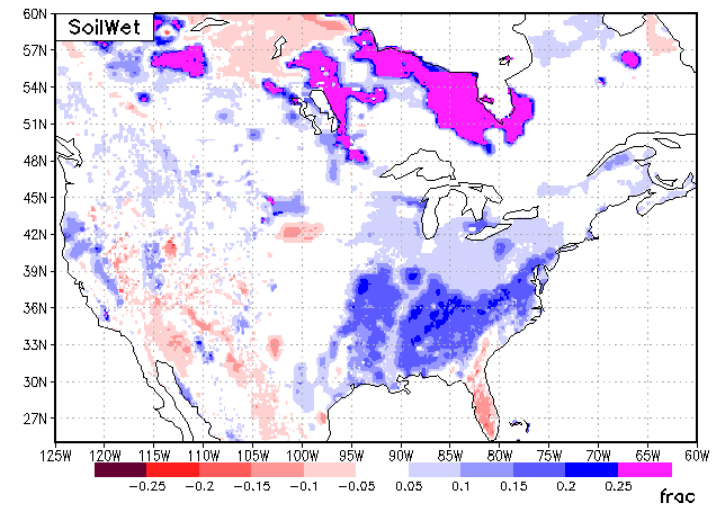
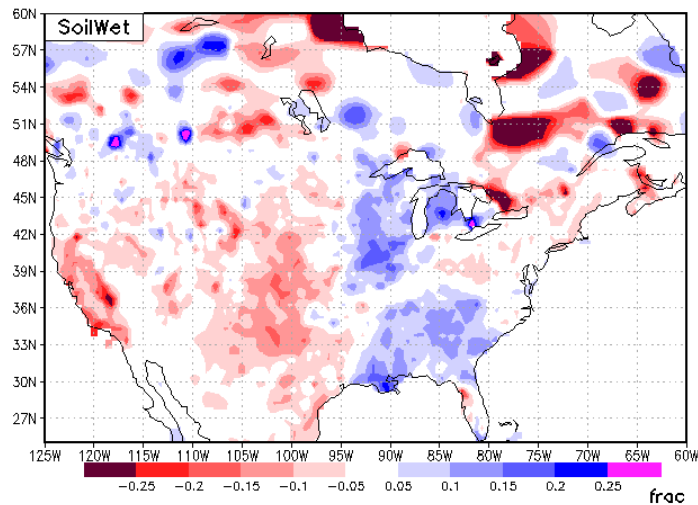


CGCM MINERVA t1279



IC: 01 May 2013  
Ens: 2 members  
May Mean

Soil Wetness





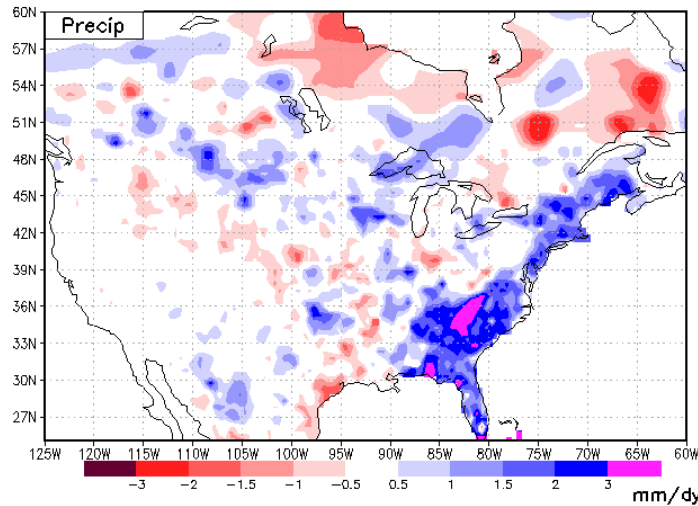


# Individual Forecast Anomalies

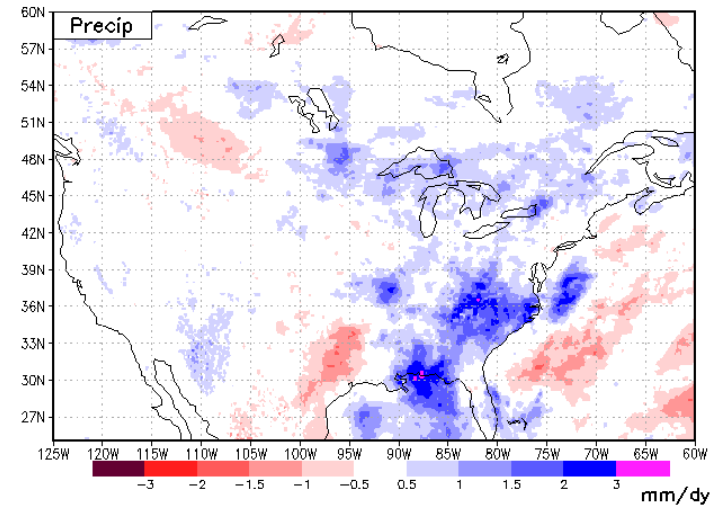
**MAY-JULY 2013**

Precipitation

Observations

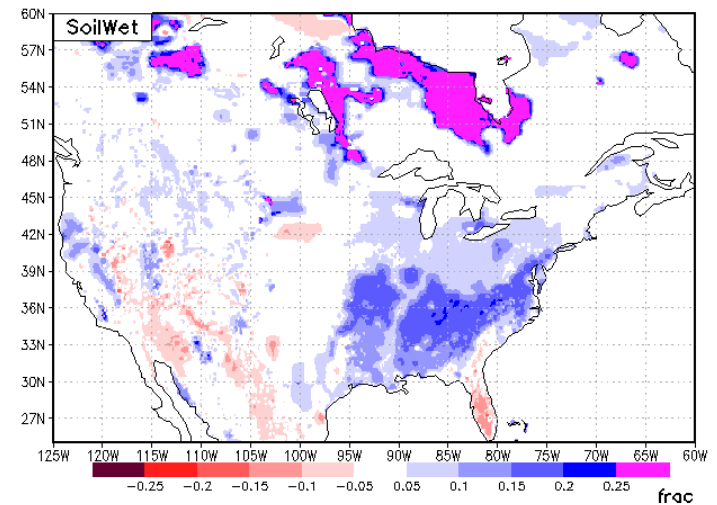
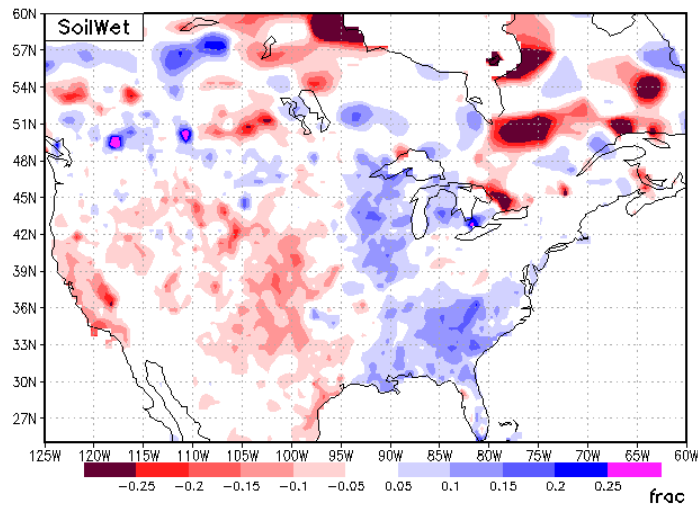


CGCM MINERVA t1279



IC: 01 May 2013  
Ens: 2 members  
May-July Mean

Soil Wetness

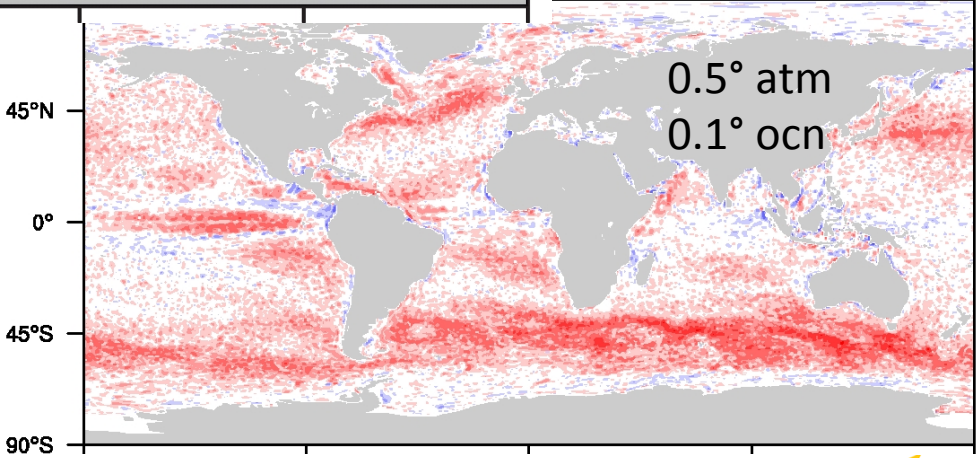
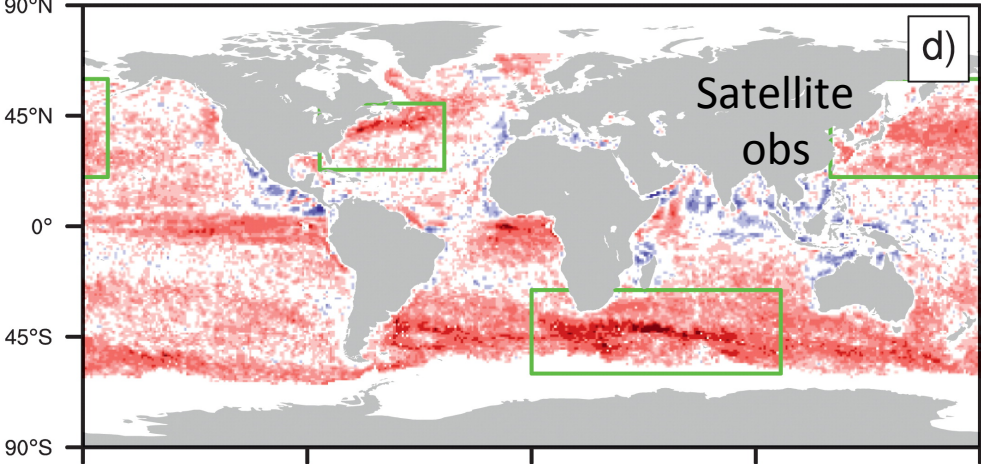
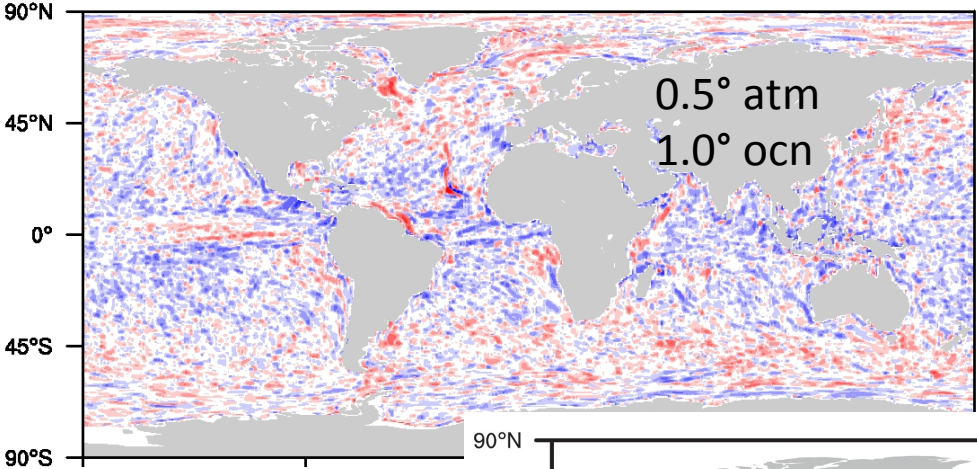


# Peta-Apps Team (2010-2012)

- Kinter, COLA (PI)
- Collins, UC Berkeley (co-PI)
- Kirtman, U. Miami (co-PI)
- Loft, NCAR (co-PI)
- Yelick, LBL (co-PI)
- Ahearn, NCAR
- Bitz, U. Washington
- Bryan, NCAR
- Dennis, NCAR
- Min, U. Miami
- Nolan, UC Berkeley
- Siquiera, U. Miami
- Stan, COLA

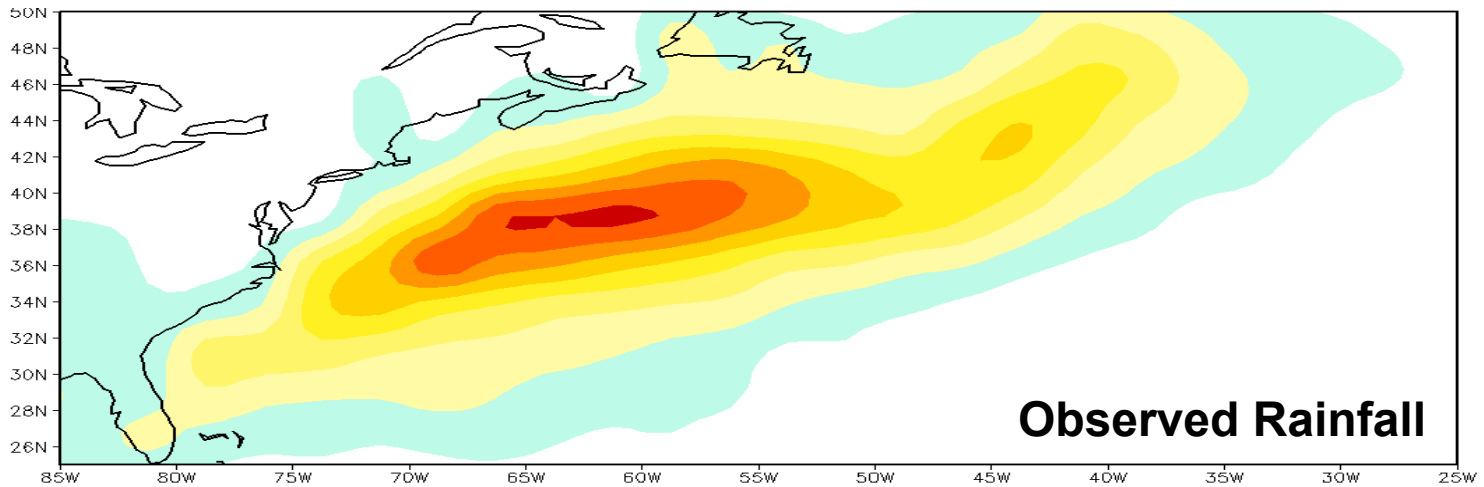
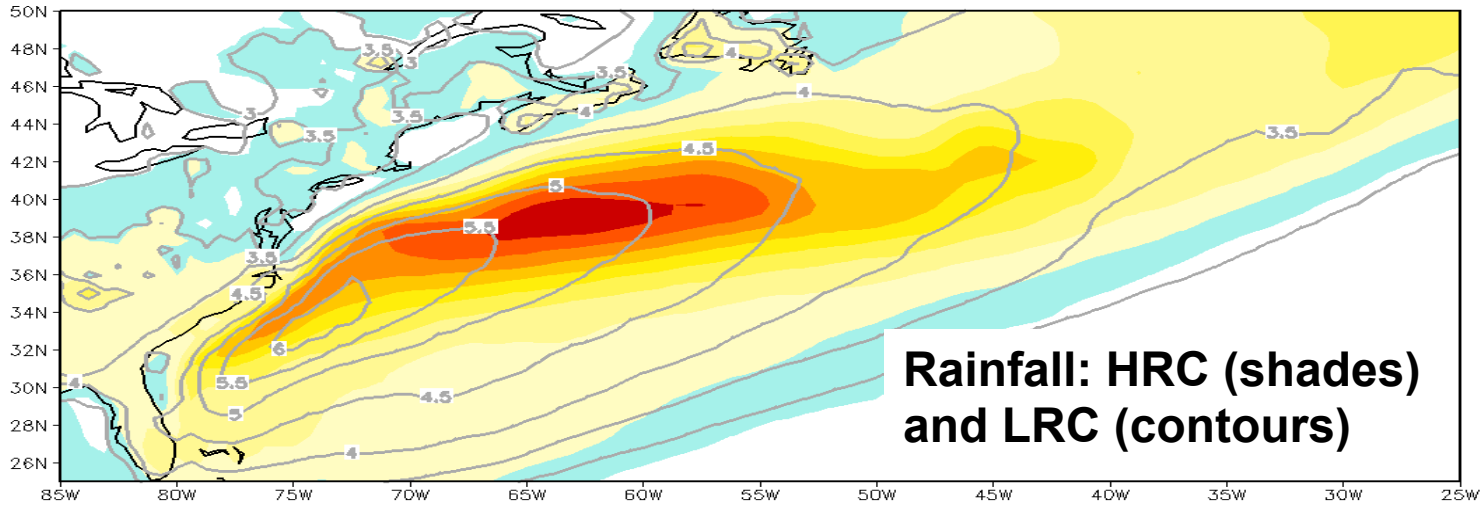
**Many thanks to TACC and  
NICS for resources and  
sustained support!**

# Correlation High-Pass SST vs. $|V_{srf}|$



Bryan et al. 2010 *J. Climate*

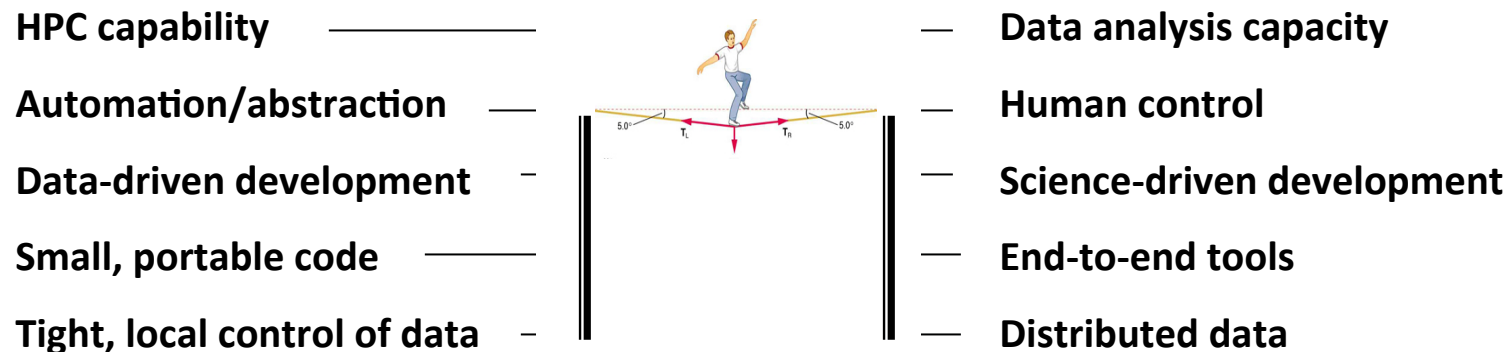
# PetaApps: Rainfall Simulation



# Challenges and Tensions

- Making effective use of large allocations – takes a village
- **Exaflood of data**
- Resolution vs. parameterization
- Sampling (e.g. extreme events)
- **Climate scientists are being forced to think about data & code issues**

## TENSIONS



**“Having more data won’t substitute for thinking hard, recognizing anomalies, and exploring deep truths.”**

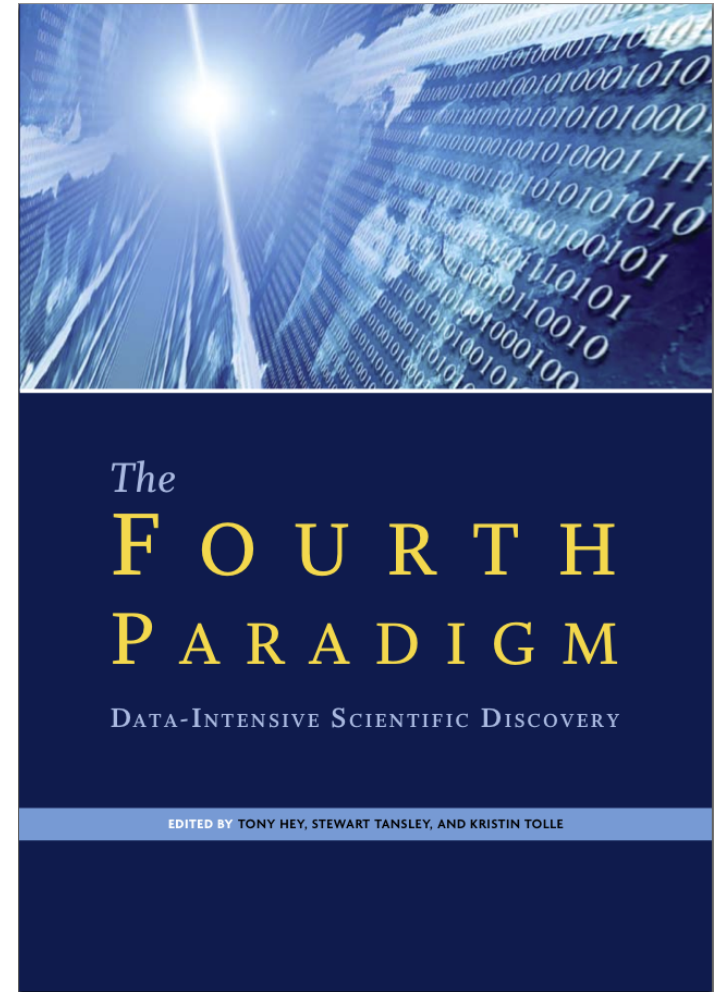
Samuel Arbeson, *Wash. Post* (18 Aug. 2013)

# Exaflood: Challenge and Opportunity

- In January 2007, Bret Swanson of the Discovery Institute coined the term **exaflood** for the **impending flood of exabytes** that would cause the Internet's congestive collapse.



- Hay et al., 2010: *The Fourth Paradigm* →



# Data Volumes

- *Project Athena*: Total data volume  
Spinning disk  
**1.2 PB (~500 TB unique)\***  
**40 TB at COLA**  
**0 TB at NICS** (was 340 TB)
- **\* no home after April 2014**
- *Project Minerva*: Total data volume  
Spinning disk  
**0.9 PB (~700 TB unique)**  
**100 TB at COLA**  
**500TB at NCAR** (for now)
- That much data breaks everything: H/W, systems management policies, networks, apps S/W, tools, and shared archive space
- **NB: Generating 700 TB using 28 M core-hours took ~3 months; this would take about a day on a system with 1M cores!**

# Athena and Minerva: Harbingers of the Exaflood

- Familiar diagnostics are hard to do at very high resolution
- Have we wrung all the “science” out of the data sets, given that we can only keep a small percentage of the total data volume on spinning disk? **How can we tell?**
- Must move from ad hoc solutions → systematic, repeatable solutions  
(transform Noah’s Ark → a Shipping Industry)
- **“We need exaflood insurance.”**  
- Jennifer Adams

