Climate Informatics

Recent advances and challenge problems for Machine Learning in Climate Science

Claire Monteleoni

Computer Science

George Washington University

Motivation for Climate Informatics

The threat of climate change:
one of the greatest challenges currently facing society.

We face an explosion in data!
- Climate model outputs
- Satellite measurements
- Environmental sensors
- …

Machine Learning has made profound impacts on:
websearch, internet advertising, Bioinformatics, etc.

Challenge: accelerate discovery in Climate Science with ML

Climate Data is Big Data

GCMs/ESMs (CMIP3/5) (Tb/day)

Satellite retrievals (Tb/day)

Next-gen reanalysis products (Tb/day)

In-situ data

Paleo-data

Regional models

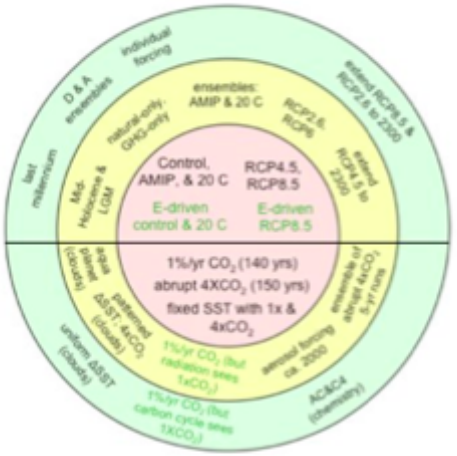
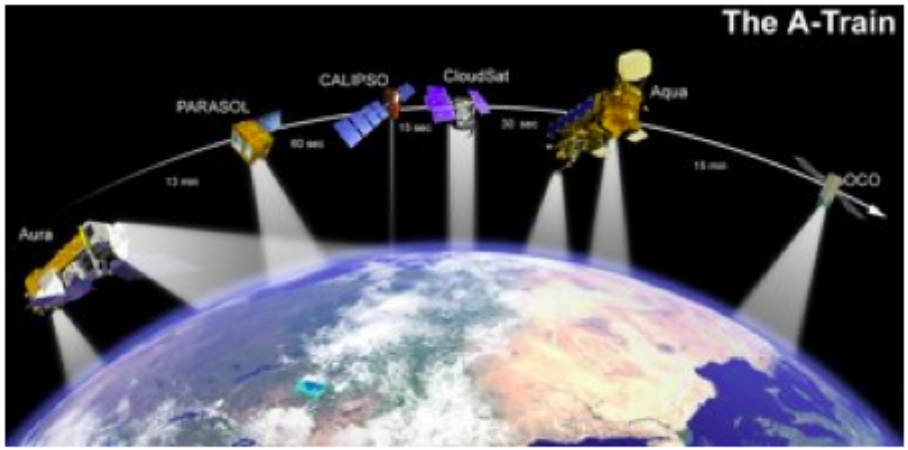
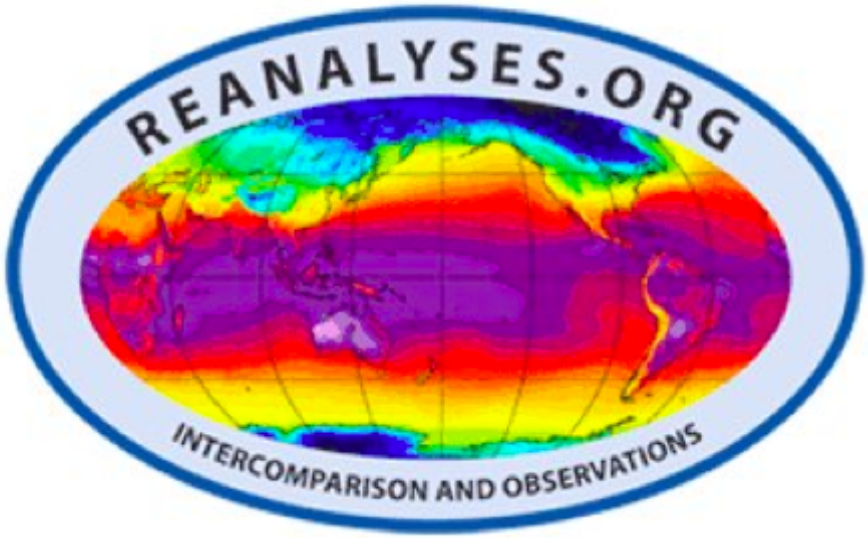


Figure 3: Schematic summary of CMIP5 long-term experiments. Green font indicates simulations that will be performed only by models with carbon cycle representation.



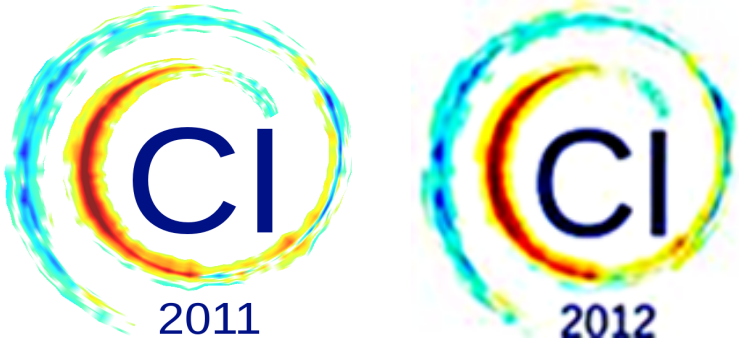


with credit to G. Schmidt

Climate Informatics: related work

- ML and data mining collaborations with climate science
  - Atmospheric chemistry, e.g. Musicant et al. '07 ('05)
  - Meteorology, e.g. Fox-Rabinovitz et al. '06
  - Seismology, e.g. Kohler et al. '08
  - Oceanography, e.g. Lima et al. '09
  - Mining/modeling climate data, e.g. Steinbach et al. '03, Steinhaeuser et al. '10, Kumar '10
- ML and climate modeling
  - Data-driven climate models, Lozano et al. '09
  - Machine learning techniques inside a climate model, or for calibration, e.g. Braverman et al. '06, Krasnopolsky et al. '10
  - ML techniques with ensembles of climate models:
    - Regional models: Sain et al. '10

Climate Informatics



- The First International Workshop on Climate Informatics, 2011
  - New York Academy of Sciences, New York, NY, August 2011
- The Second International Workshop on Climate Informatics, 2012
  - National Center for Atmospheric Research, Boulder, CO, September 2012
- Climate Informatics wiki:

  http://sites.google.com/site/1stclimateinformatics/materials
  - Data sources and descriptions
  - Challenge problems
  - Links to tutorials and materials
- Climate Informatics book chapter [M, Schmidt et al. 2012]
  - Challenge problems
  - Data descriptions
  - Success stories and related work

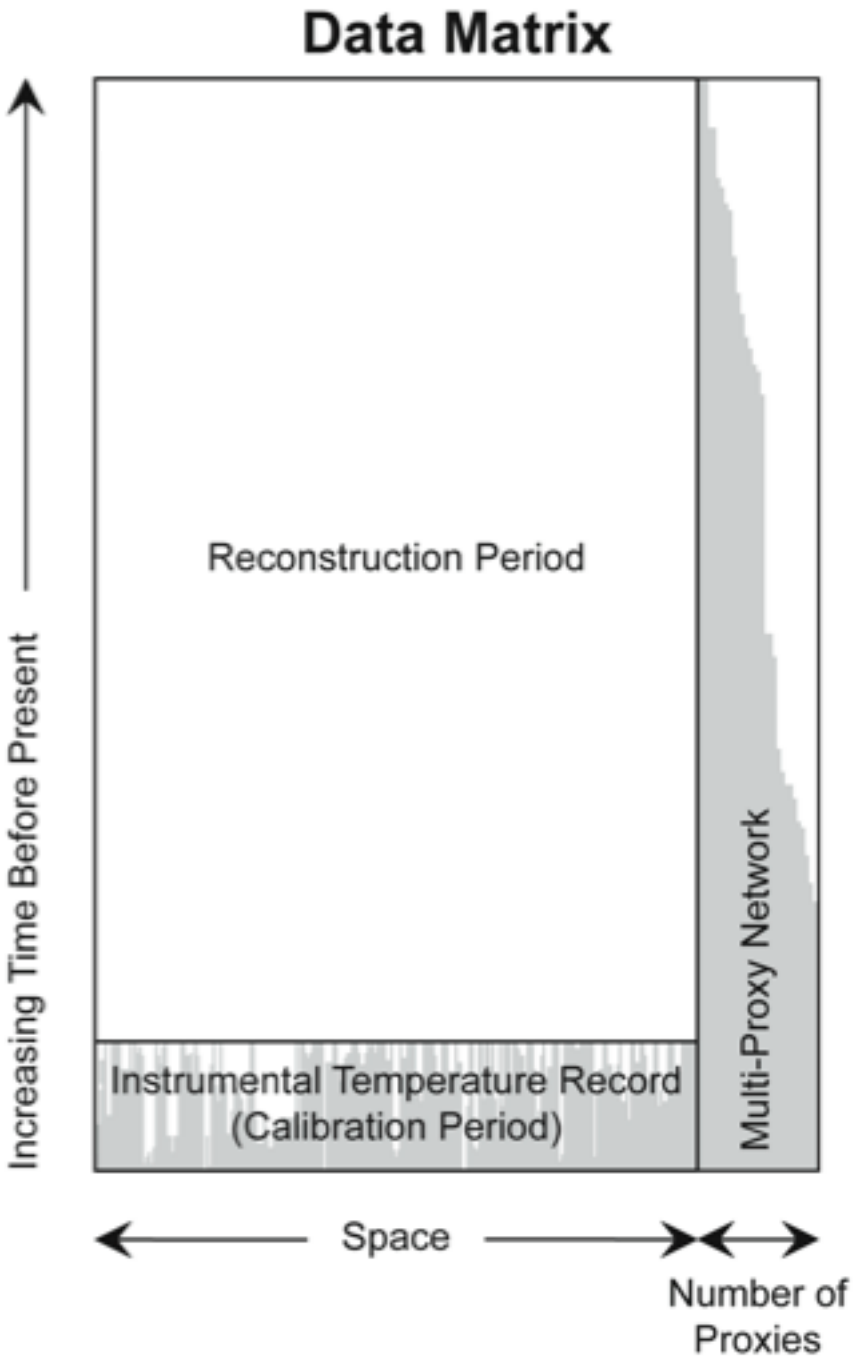
Paleo-temperature

Problem:

- To understand climate change we need to understand past climates.
- NOTE: climate has fluctuated at much greater scales in the past than in the 20th Century.
- However the variance on measurements is higher in the past.
  - We did not have a global grid of measurements
  - Measurements corrupted or lost

Challenge: use paleo-proxies to reconstruct temps

e.g. tree rings, ice cores, water isotopes.



Credit: J. Smerdon

Climate extremes

There is evidence that even with a warming trend, variance is predicted to increase.

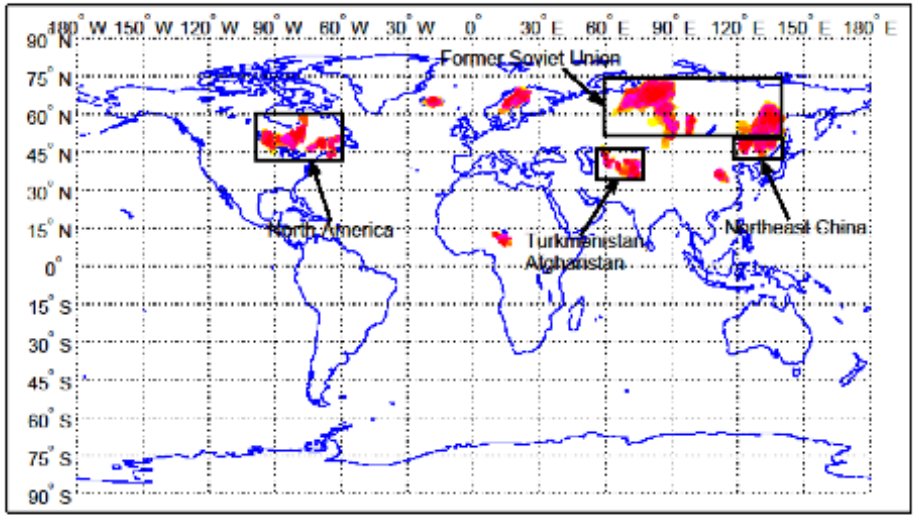
→Increasingly extreme events e.g.

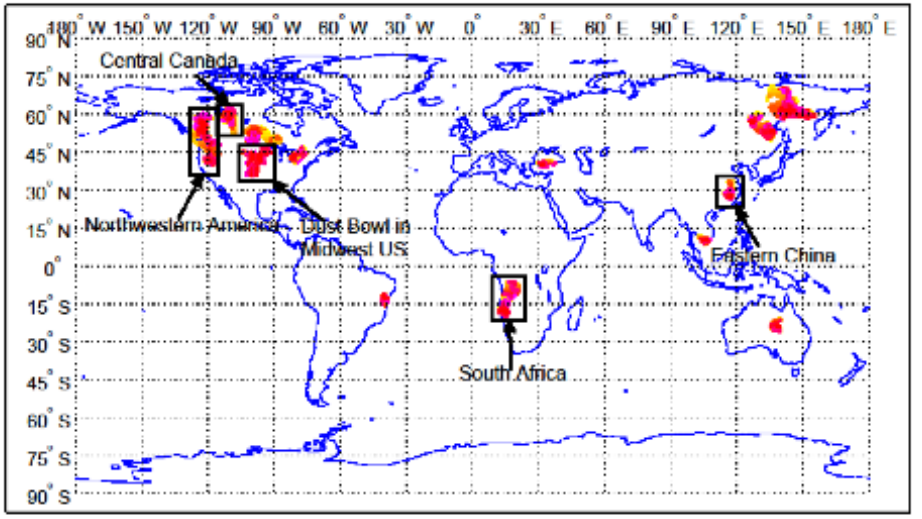
- cold spells
- droughts

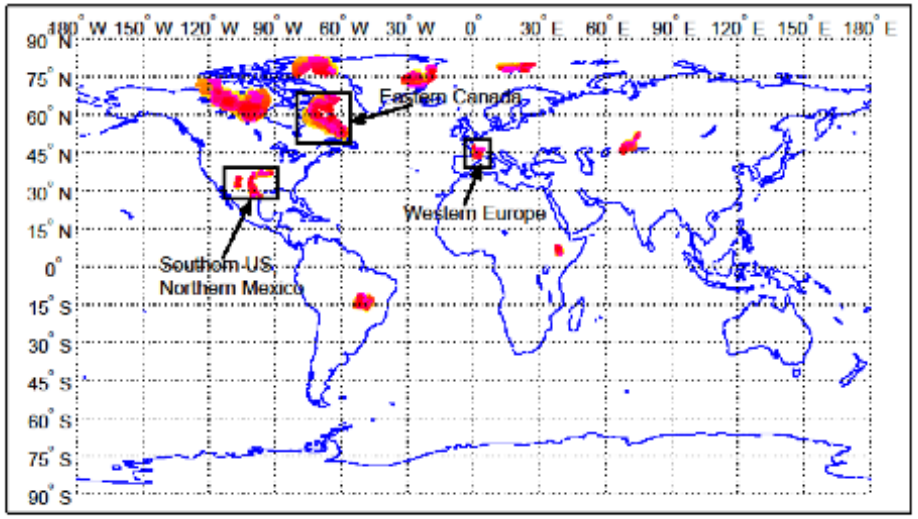
Challenge: detecting, predicting climate extremes.

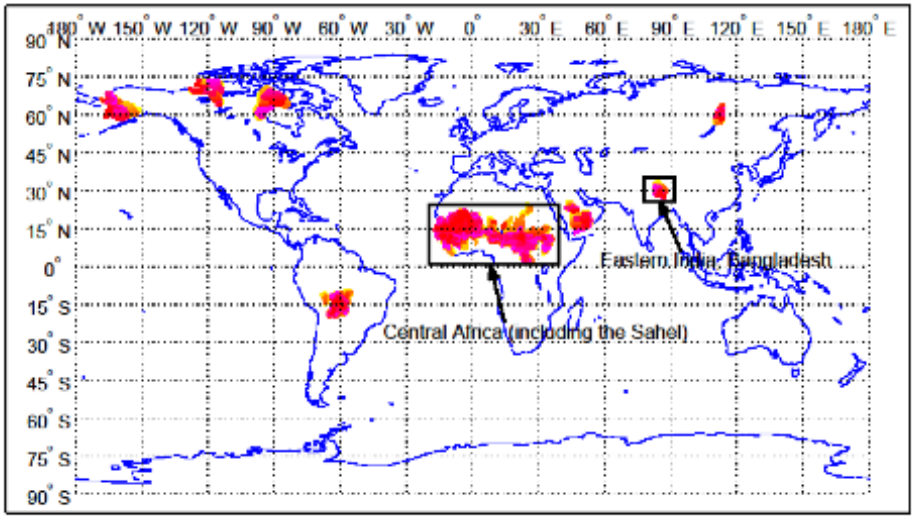
Drought detection over the past century with MRFs

[Q. Fu, A. Banerjee, S. Liess, and P. K. Snyder, SDM 2012]









Each panel shows the drought starting from a particular decade: 1905-1920 (top left), 1921-1930 (top right), 1941-1950 (bottom left), and 1961-1970 (bottom right). The regions in black rectangles indicate the common droughts found by previous work in climate science.

credit: A. Banerjee

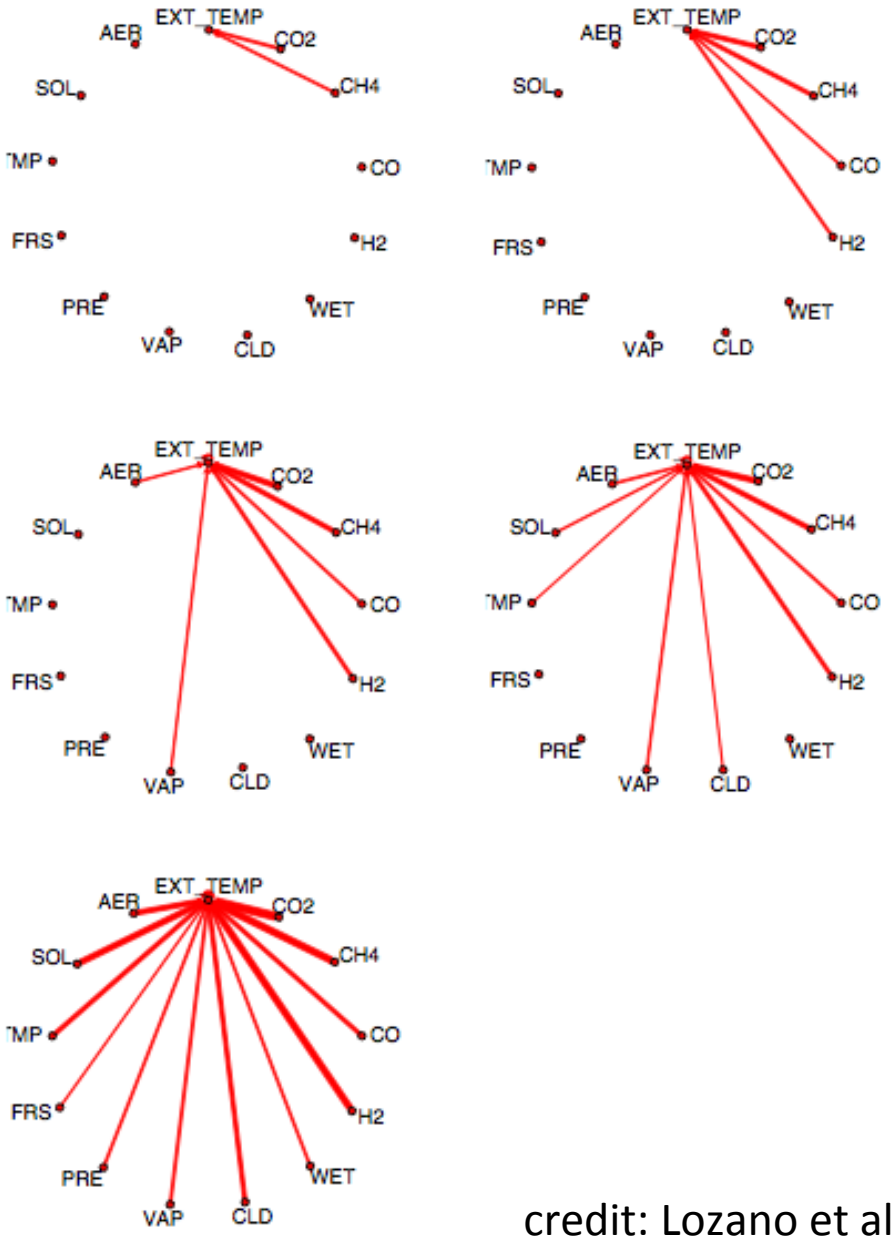
Climate change attribution

Challenge: which factors have contributed to temperature changes and to what extent?

[Lozano et al. KDD 2009]:

Which factors granger-caused extreme temperatures?

Approach using group elastic nets



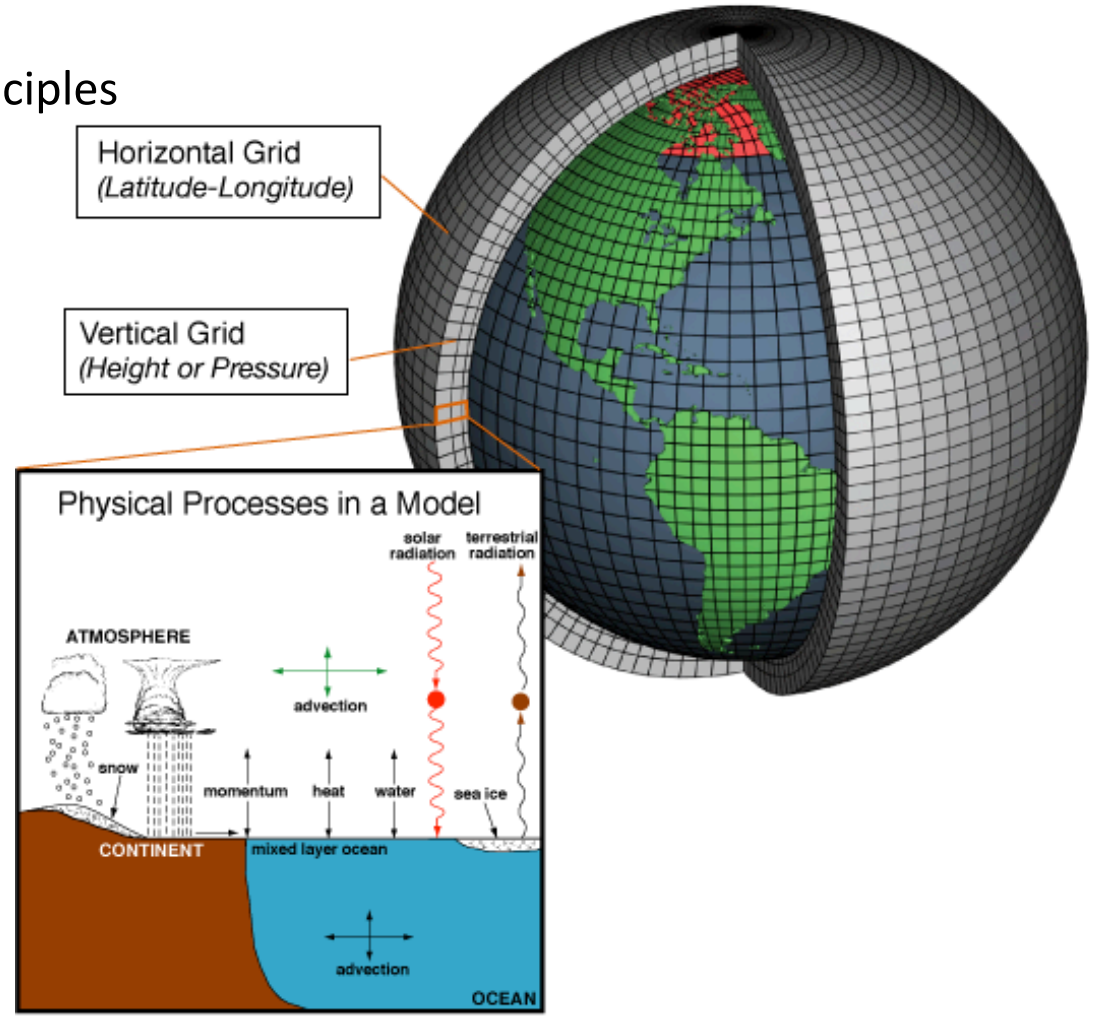
credit: Lozano et al.

Figure 4: Attributing the change in 1-year return level for temperature extremes using annual data. Output causal structures for decreasing degrees of sparsity. Edge thickness represents the causality strength.

Climate modeling

Climate model: a complex system of interacting mathematical models

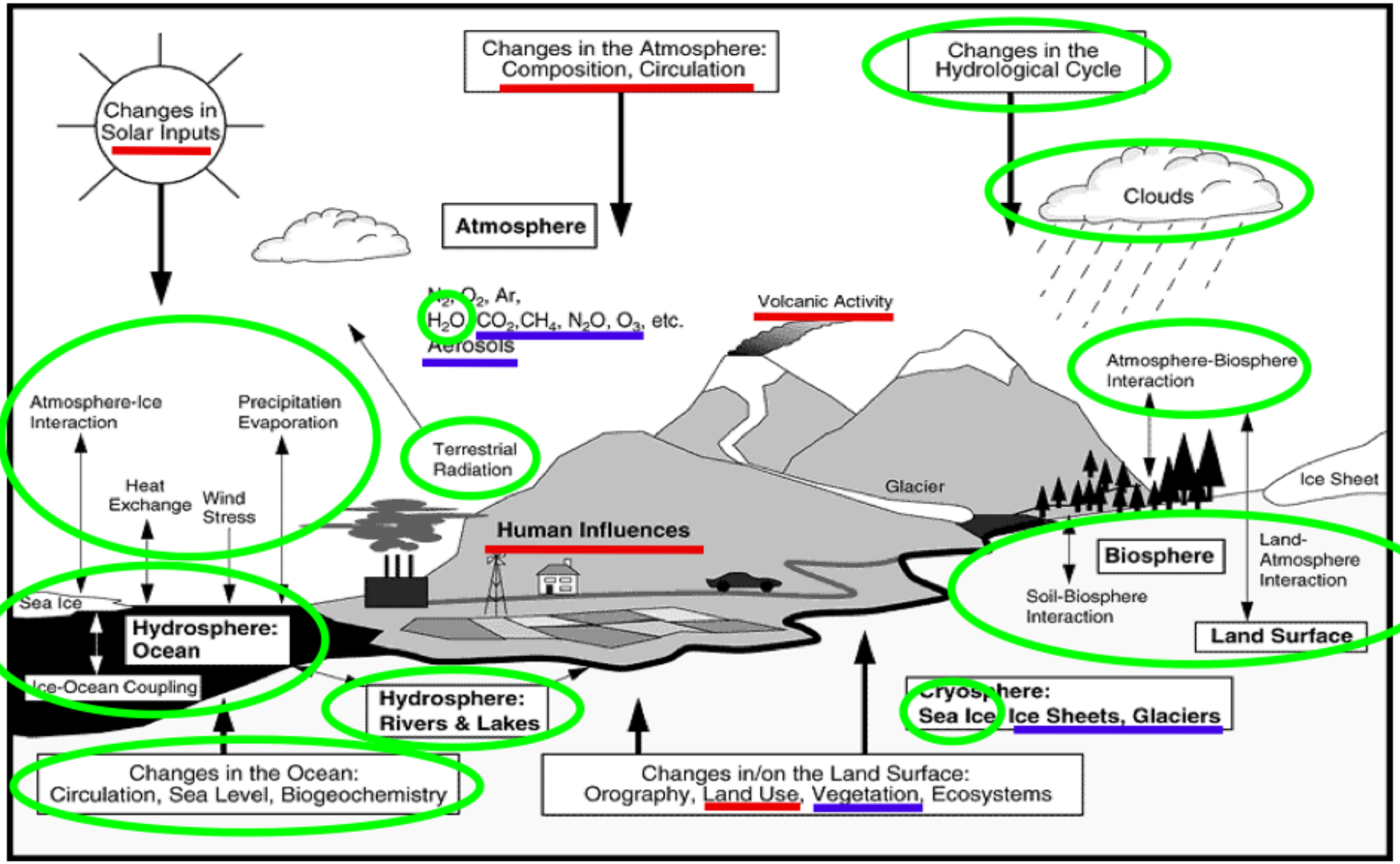
- Not data-driven
- Based on scientific first principles
  - Meteorology
  - Oceanography
  - Geophysics
  - …



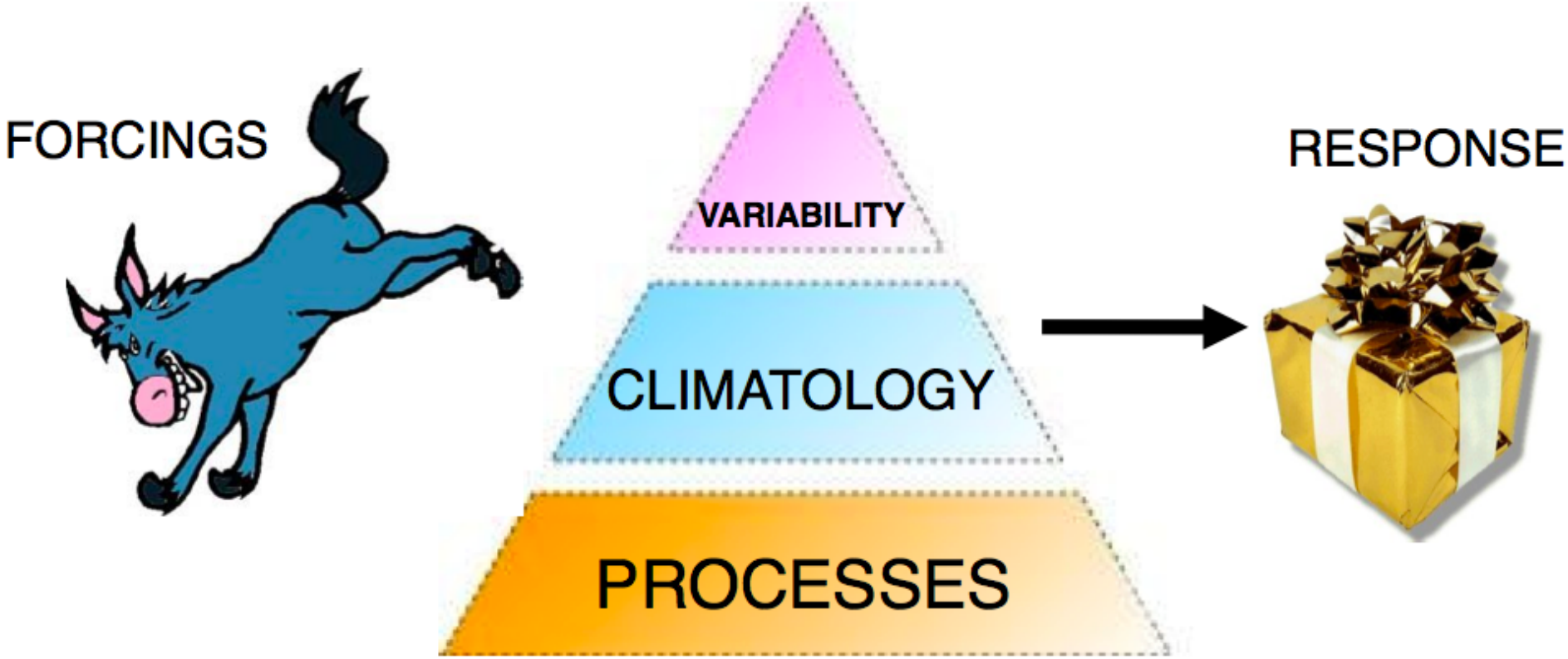
Climate model differences

- Assumptions
- Discretizations
- Scale interactions
  - Micro: rain drop
  - Macro: ocean

A climate model (image credit: G. Schmidt)

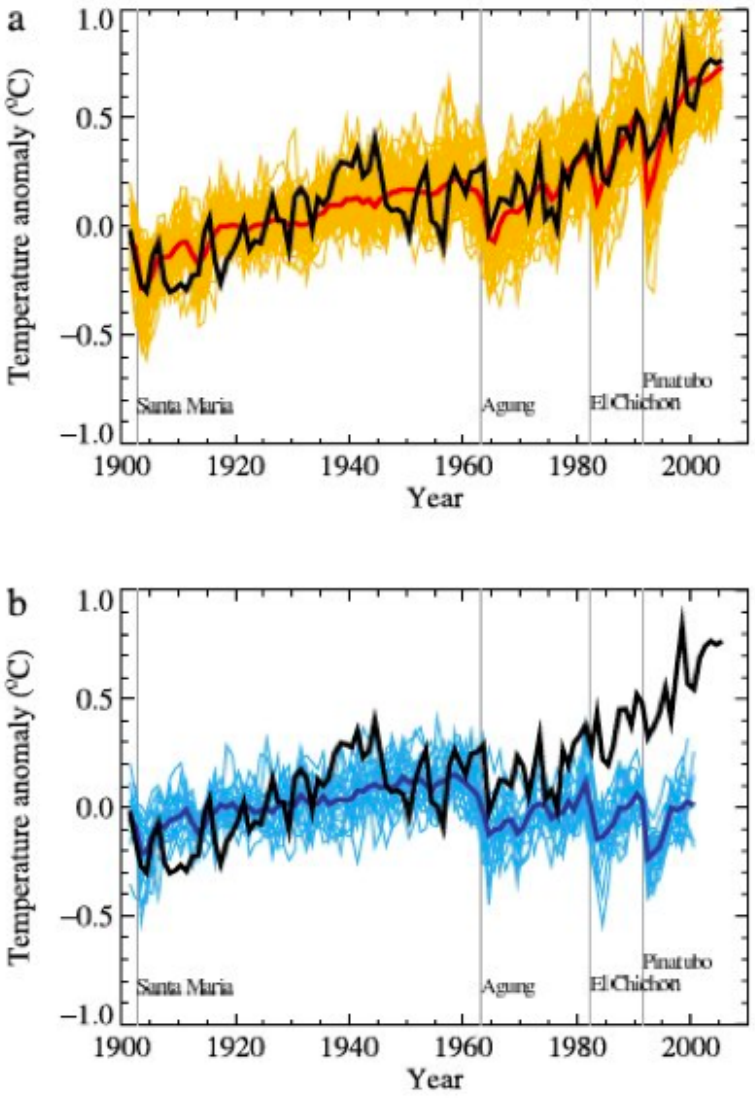


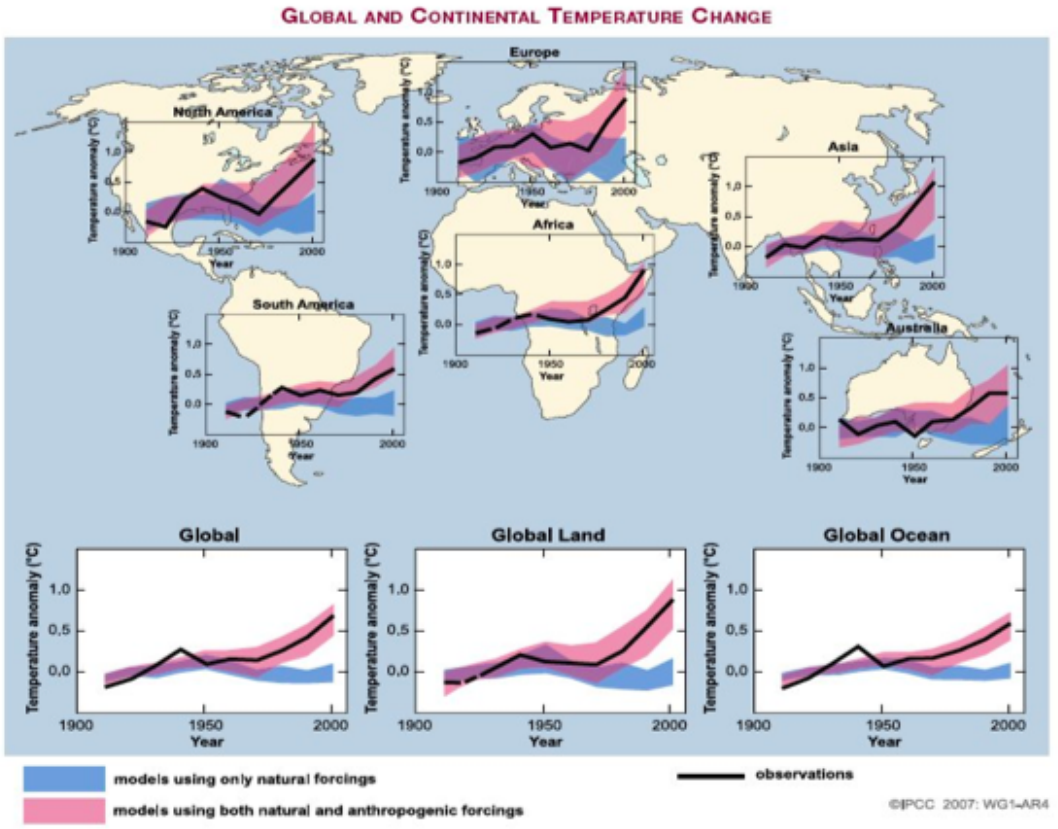
Climate forcings



Slide credit: G. Schmidt

Surface temperature anomalies





Slide credit: G. Schmidt

Climate models

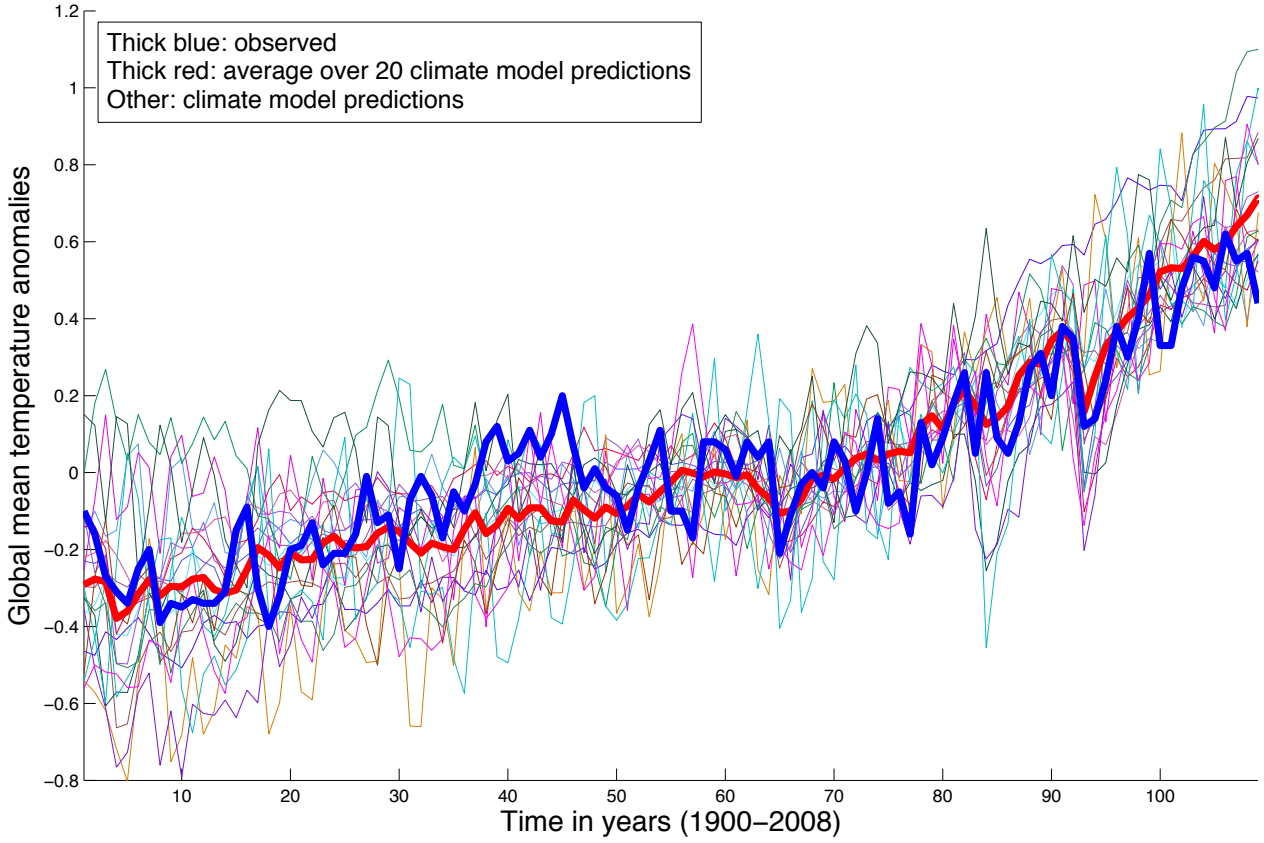
- IPCC: Intergovernmental Panel on Climate Change
  - Nobel Peace Prize 2007 (shared with Al Gore).
  - Interdisciplinary scientific body, formed by UN in 1988.
  - Fourth Assessment Report 2007, on global climate change
    450 lead authors from 130 countries, 800 contributing authors, over 2,500 reviewers.
  - Next Assessment Report is due in 2013.

- Climate models contributing to IPCC reports include:

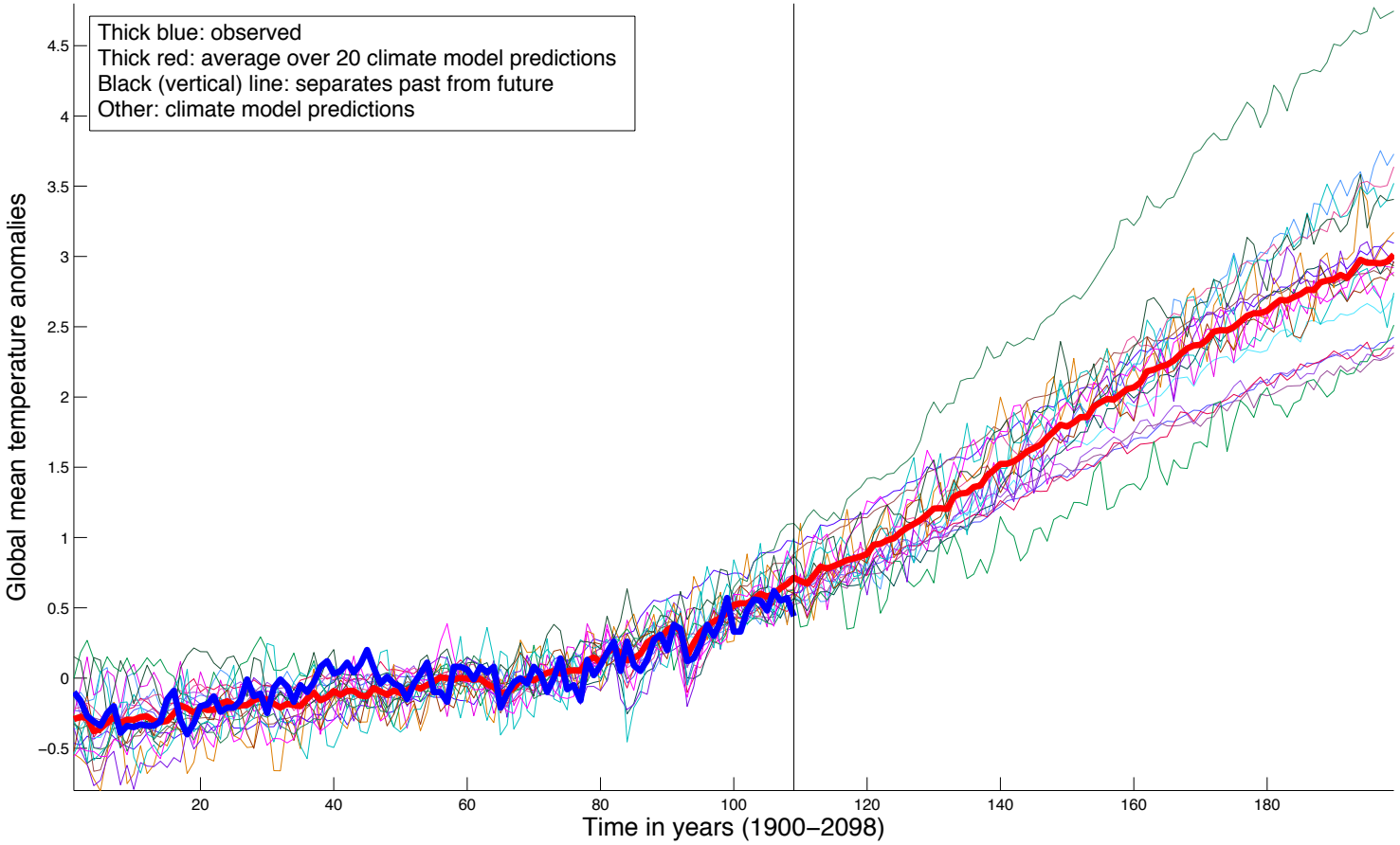
  Bjerknes Center for Climate Research (Norway), Canadian Centre for Climate Modelling and Analysis, Centre National de Recherches Météorologiques (France), Commonwealth Scientific and Industrial Research Organisation (Australia), Geophysical Fluid Dynamics Laboratory (Princeton University), Goddard Institute for Space Studies (NASA), Hadley Centre for Climate Change (United Kingdom Meteorology Office), Institute of Atmospheric Physics (Chinese Academy of Sciences), Institute of Numerical Mathematics Climate Model (Russian Academy of Sciences), Istituto Nazionale di Geofisica e Vulcanologia (Italy), Max Planck Institute (Germany), Meteorological Institute at the University of Bonn (Germany), Meteorological Research Institute (Japan), Model for Interdisciplinary Research on Climate (Japan), National Center for Atmospheric Research (Colorado), among others.

Climate model predictions

Global mean temperature anomalies. Temperature anomaly: difference w.r.t. the temperature at a benchmark time. Magnitude of temperature change. Averaged over many geographical locations, per year.



Climate model predictions



Future fan-out.

Improving predictions of Multi-Model Ensemble of GCMs

- No one model predicts best all the time.
- Average prediction over all models is best predictor over time. [Reichler & Kim, Bull. AMS '08], [Reifen & Toumi, GRL '09]
- IPCC held 2010 Expert Meeting on how to better combine model predictions.

Can we do better, using Machine Learning?

Challenge: How should we predict future climates?

- While taking into account the 20 climate models' predictions

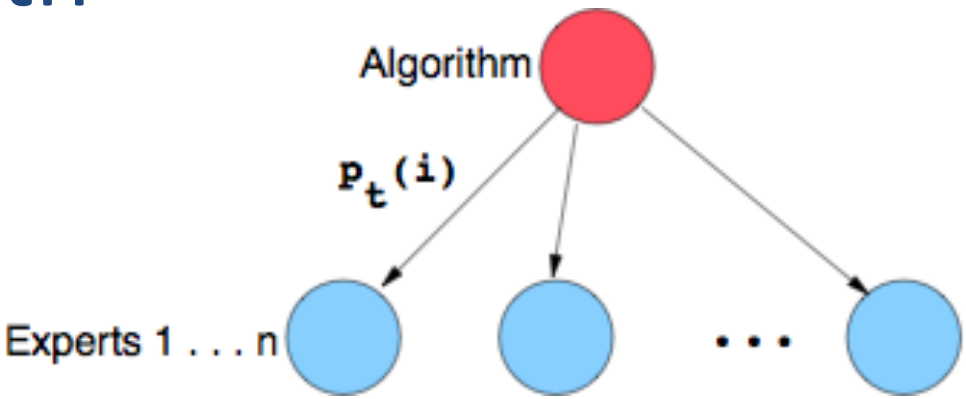
Tracking climate models

[M, Schmidt, Saroha, & Asplund, SAM 2011 (CIDU 2010)]:

- Application of Learn-α algorithm [M & Jaakkola, NIPS '03]
  - Online learning to track a set of "expert" predictors under changing observations.
- Tracking global climate models, on mean temperature anomaly predictions.
  - Experiments at global and regional scales, annual and monthly time-scales.
- Experiments on historical data
  - Valid, since climate models are *not* data-driven.
- Future simulations using "perfect model" assumption from climate science.

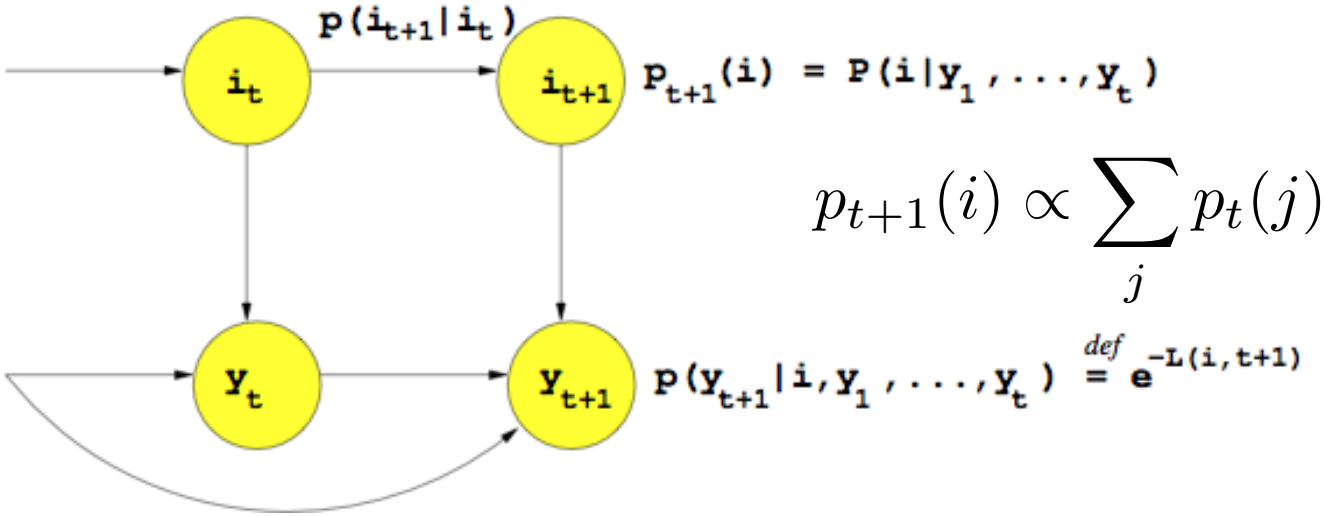
Online learning with expert advice

Learner maintains distribution over *n* "experts."



- Experts are black boxes: need not be good predictors, can vary with time, and depend on one another. We use GCM temp. predictions.
- Learner maintains/updates probability distribution $p_t(i)$ over experts, *i*, representing how well each expert has predicted recently.
  - Used to inform learner's prediction.
- *L(i, t)* is prediction loss of expert *i* at time *t*. We use squared loss.
- Family of Multiplicative Updates algorithms (*cf*. "Hedge," "Weighted Majority"), descended from "Winnow," [Littlestone 1988],[Littlestone & Warmuth'89].

Online learning: time-varying data



$$p_{t+1}(i) \propto \sum_j p_t(j) e^{-L(j,t)} p(i|j)$$

- For a family of these algorithms, [M & Jaakkola, 2003] derived $p_t(i)$ as Bayesian updates of a generalized Hidden Markov Model
- Hidden variable: identity of current "best expert"
- Transition dynamics, $p(i \mid j)$, model non-stationarity

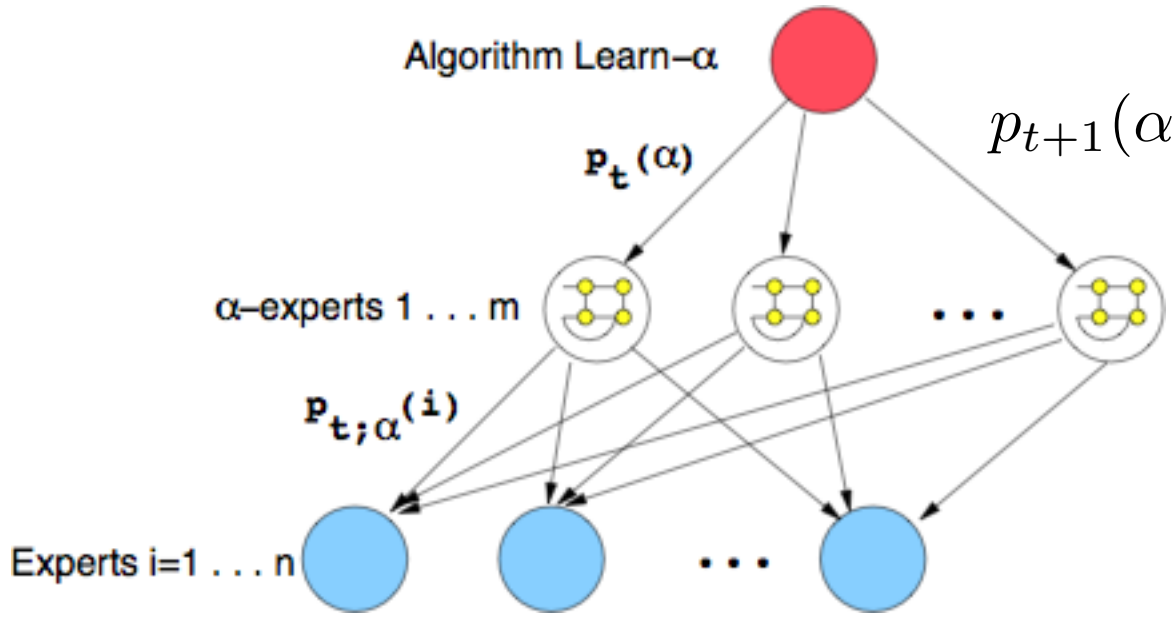
Online learning: time-varying data

Fixed-Share Algorithm [Herbster & Warmuth, 1998]:

- Assumes there is a probability α that the hidden "best expert" switches at each time step

$$P(i|j;\alpha) = \begin{cases} (1-\alpha) & i = j \\ \frac{\alpha}{n-1} & i \neq j \end{cases}$$

Online learning: time-varying data



$$p_{t+1}(\alpha) \propto p_t(\alpha) e^{-L(\alpha,t)}$$

Learn-α Algorithm [M & Jaakkola, 2003]:

- Learns the α parameter by tracking a set of meta-experts, Fixed-Share algorithms, each with a different α value

Performance guarantees

[M & Jaakkola, NIPS 2003]: Bounds on "regret" for using wrong value of α for the observed sequence of length T:

Theorem. O(T) upper bound for Fixed-Share(α) algorithms.

Theorem. Ω(T) sequence dependent lower bound for Fixed-Share(α) algorithms.

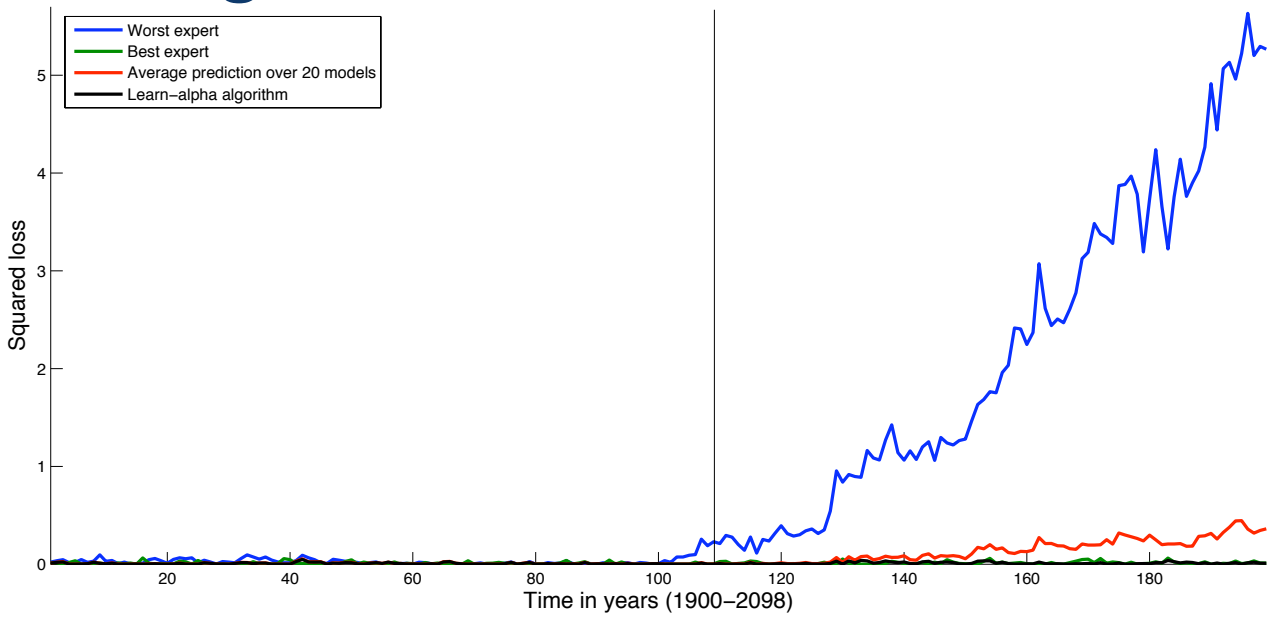
Theorem. O(log T) upper bound for Learn-α algorithm.

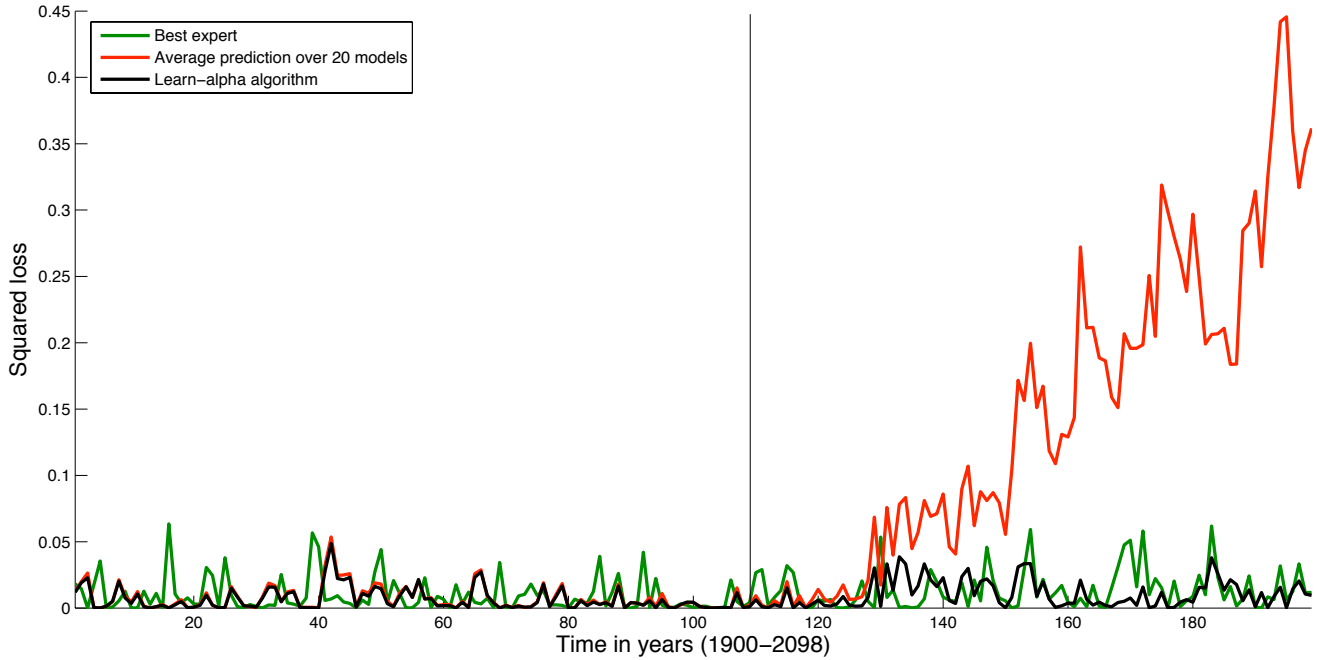
- Regret-optimal discretization of α for fixed sequence length, T.
- Using previous algorithms with wrong α can also lead to poor empirical performance.

Tracking climate models: experiments

- Model predictions from 20 climate models
  - Mean temperature anomaly predictions (1900-2098) – CMIP3 archive
- Historical experiments with NASA temperature data – GISTEMP
- Future simulations with "perfect model" assumption.
  - Ran 10 such global simulations to observe general trends
  - Collected detailed statistics on 4 representative ones: best and worst model on historical data, and 2 in between.
- Regional experiments: data from KNMI Climate Explorer
  - Africa (-15 – 55E, -40 – 40N)
  - Europe (0 – 30E, 40 – 70N)
  - North America (-60 – -180E, 15 – 70N)
  - Annual and monthly time-scales; historical & 2 future simulations/region.

Learning curves





Global results

Algorithm:	Historical	Future Sim. 1	Future Sim. 2	Future Sim. 3	Future Sim. 4
Learn-α Algorithm	0.0119 $\sigma^2 = 0.0002$	0.0085 $\sigma^2 = 0.0001$	**0.0125** $\sigma^2 = 0.0004$	**0.0252** $\sigma^2 = 0.0010$	**0.0401** $\sigma^2 = 0.0024$
Linear Regression*	0.0158 $\sigma^2 = 0.0005$	**0.0051** $\sigma^2 = 0.0001$	0.0144 $\sigma^2 = 0.0004$	0.0264 $\sigma^2 = 0.0125$	0.0498 $\sigma^2 = 0.0054$
Best Climate Model (for the observations)	**0.0112** $\sigma^2 = 0.0002$	0.0115 $\sigma^2 = 0.0002$	0.0286 $\sigma^2 = 0.0014$	0.0301 $\sigma^2 = 0.0018$	0.0559 $\sigma^2 = 0.0053$
Average Prediction (over climate models)	0.0132 $\sigma^2 = 0.0003$	0.0700 $\sigma^2 = 0.0110$	0.0306 $\sigma^2 = 0.0016$	0.0623 $\sigma^2 = 0.0055$	0.0497 $\sigma^2 = 0.0036$
Median Prediction (over climate models)	0.0136 $\sigma^2 = 0.0003$	0.0689 $\sigma^2 = 0.0111$	0.0308 $\sigma^2 = 0.0017$	0.0677 $\sigma^2 = 0.0070$	0.0527 $\sigma^2 = 0.0038$
Worst Climate Model (for the observations)	0.0726 $\sigma^2 = 0.0068$	1.0153 $\sigma^2 = 2.3587$	0.8109 $\sigma^2 = 1.4109$	0.3958 $\sigma^2 = 0.5612$	0.5004 $\sigma^2 = 0.5988$

TABLE 1. Mean and variance of annual losses. The best score per experiment is in bold. The Average Prediction over climate models is the benchmark technique. *Linear Regression cannot form predictions for the first 20 years (19 in the future simulations), so its mean is over fewer years than all the other algorithms, starting from the 21st (20th in future simulations) year.

On 10 future simulations (including 1-4 above), Learn-α suffers less loss than the mean prediction (over remaining models) on 75-90% of the years.

Regional results: historical

Algorithm:	Africa	Europe	North America	Africa	Europe	North America
Learn-α Algorithm	0.0283 $\sigma^2 = 0.0020$	**0.1794** $\sigma^2 = 0.0520$	**0.0407** $\sigma^2 = 0.0036$	**0.0598** $\sigma^2 = 0.0085$	**0.3048** $\sigma^2 = 0.3006$	**0.0959** $\sigma^2 = 0.0311$
Linear Regression*	0.0391 $\sigma^2 = 0.0039$	38.9724** $\sigma^2 = 134700.0$	0.0704 $\sigma^2 = 0.0156$	0.0741 $\sigma^2 = 0.0301$	1.7442 $\sigma^2 = 43.9616$	0.1119 $\sigma^2 = 0.0432$
Best Climate Model (for the observations)	**0.0254** $\sigma^2 = 0.0015$	0.2752 $\sigma^2 = 0.1207$	0.0450 $\sigma^2 = 0.0035$	0.1144 $\sigma^2 = 0.0285$	2.2498 $\sigma^2 = 15.4041$	0.1629 $\sigma^2 = 0.0935$
Average Prediction (over climate models)	0.0331 $\sigma^2 = 0.0025$	0.2383 $\sigma^2 = 0.0868$	0.0493 $\sigma^2 = 0.0058$	0.0752 $\sigma^2 = 0.0106$	1.4781 $\sigma^2 = 7.5964$	0.1101 $\sigma^2 = 0.0417$
Median Prediction (over climate models)	0.0291 $\sigma^2 = 0.0021$	0.2391 $\sigma^2 = 0.0964$	0.0502 $\sigma^2 = 0.0066$	0.0777 $\sigma^2 = 0.0117$	1.5001 $\sigma^2 = 8.1498$	0.1116 $\sigma^2 = 0.0456$
Worst Climate Model (for the observations)	0.1430 $\sigma^2 = 0.0368$	1.0180 $\sigma^2 = 2.4702$	0.1593 $\sigma^2 = 0.0372$	0.2333 $\sigma^2 = 0.1020$	4.2104 $\sigma^2 = 71.2737$	1.1698 $\sigma^2 = 6.3192$

Annual

Monthly

Regional results: future simulations

Algorithm:	Africa 1	Africa 2	Europe 1	Europe 2	N. Amer. 1	N. Amer. 2
Learn-α Algorithm	**0.0890** $\sigma^2 = 0.0167$	**0.1053** $\sigma^2 = 0.0249$	**0.2812** $\sigma^2 = 0.4134$	**0.6624** $\sigma^2 = 3.6678$	0.0968 $\sigma^2 = 0.0272$	**0.6061** $\sigma^2 = 1.6429$
Linear Regression*	0.0985 $\sigma^2 = 0.2680$	0.1384 $\sigma^2 = 0.0455$	1.1487 $\sigma^2 = 4.2672$	3.0836 $\sigma^2 = 44.1931$	**0.0923** $\sigma^2 = 0.0365$	1.0458 $\sigma^2 = 4.4447$
Best Expert (for the observations)	0.1912 $\sigma^2 = 0.0757$	0.1967 $\sigma^2 = 0.0754$	2.1210 $\sigma^2 = 12.6767$	3.7893 $\sigma^2 = 39.2087$	0.1713 $\sigma^2 = 0.0903$	1.0478 $\sigma^2 = 3.9090$
Average Prediction (over climate models)	0.1388 $\sigma^2 = 0.0410$	0.1806 $\sigma^2 = 0.0716$	1.1106 $\sigma^2 = 4.4023$	2.9353 $\sigma^2 = 29.9128$	0.1432 $\sigma^2 = 0.0478$	1.0745 $\sigma^2 = 4.1346$
Median Prediction	0.1266 $\sigma^2 = 0.0352$	0.1711 $\sigma^2 = 0.0637$	1.1385 $\sigma^2 = 4.5734$	2.9093 $\sigma^2 = 30.3332$	0.1835 $\sigma^2 = 0.0827$	1.1075 $\sigma^2 = 4.2544$
Worst Expert (for the observations)	0.5236 $\sigma^2 = 0.5782$	0.5625 $\sigma^2 = 0.7018$	3.8266 $\sigma^2 = 47.7359$	5.0029 $\sigma^2 = 76.7785$	1.2311 $\sigma^2 = 3.3160$	2.2641 $\sigma^2 = 12.0301$

TABLE 4. Regional results on two future simulations per region. Mean and variance of monthly losses. The best score per experiment is in bold. The Average Prediction over climate models is the benchmark technique. *Linear Regression cannot form predictions for the first 18 months, so its mean is over fewer months than all the other algorithms, starting from the 19th month.

Contributions

- Tracking climate models (TCM)

  [M, Schmidt, Saroha, & Asplund, SAM 2011 (CIDU 2010)]:

  – Applied online learning with expert advice to track GCMs

  – Considered each geospatial region as a separate problem

- Neighborhood-Augmented TCM (NTCM)

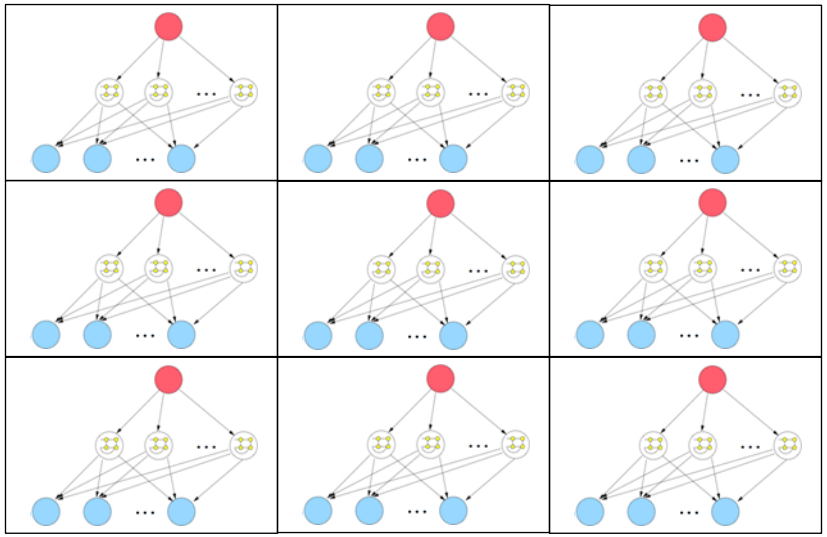
  [McQuade & M, AAAI 2012]:

  – Build a rich modeling framework in which the climate predictions are made at higher geospatial resolutions

  – Model neighborhood influences among geospatial regions

Neighborhood-Augmented TCM (NTCM)

- Run instances of Learn-α (variant) on multiple sub-regions that partition the globe
- Global temperature anomaly computed as mean of sub-region algorithm predictions
- Experiments conducted using several different region sizes



Neighborhood-Augmented TCM (NTCM)

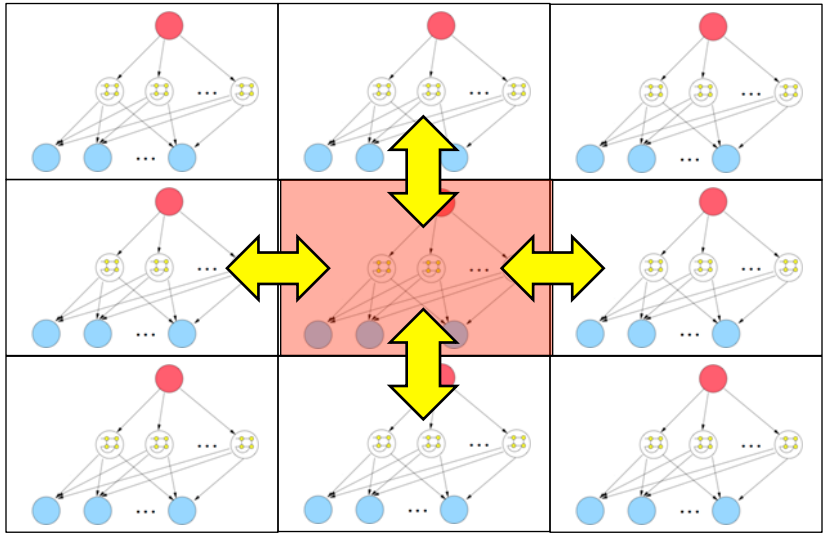
Non-homogenous HMM transition matrix:

$$P(i \mid k;\alpha) = \begin{cases} (1-\alpha) & \text{if i=k} \\ \frac{\alpha}{Z}\left[(1-\beta) + \beta \frac{1}{|S(r)|} \sum_{s \in S(r)} P_{t,s}(i)\right] & \text{if i}\neq\text{k} \end{cases}$$

- *S(r)* - neighborhood scheme: set of "neighbors" of region *r*
- $P_{t,s}(i)$ - probability of expert *i* in region *s*
- β - regulates geospatial influence
- *Z* - normalization factor

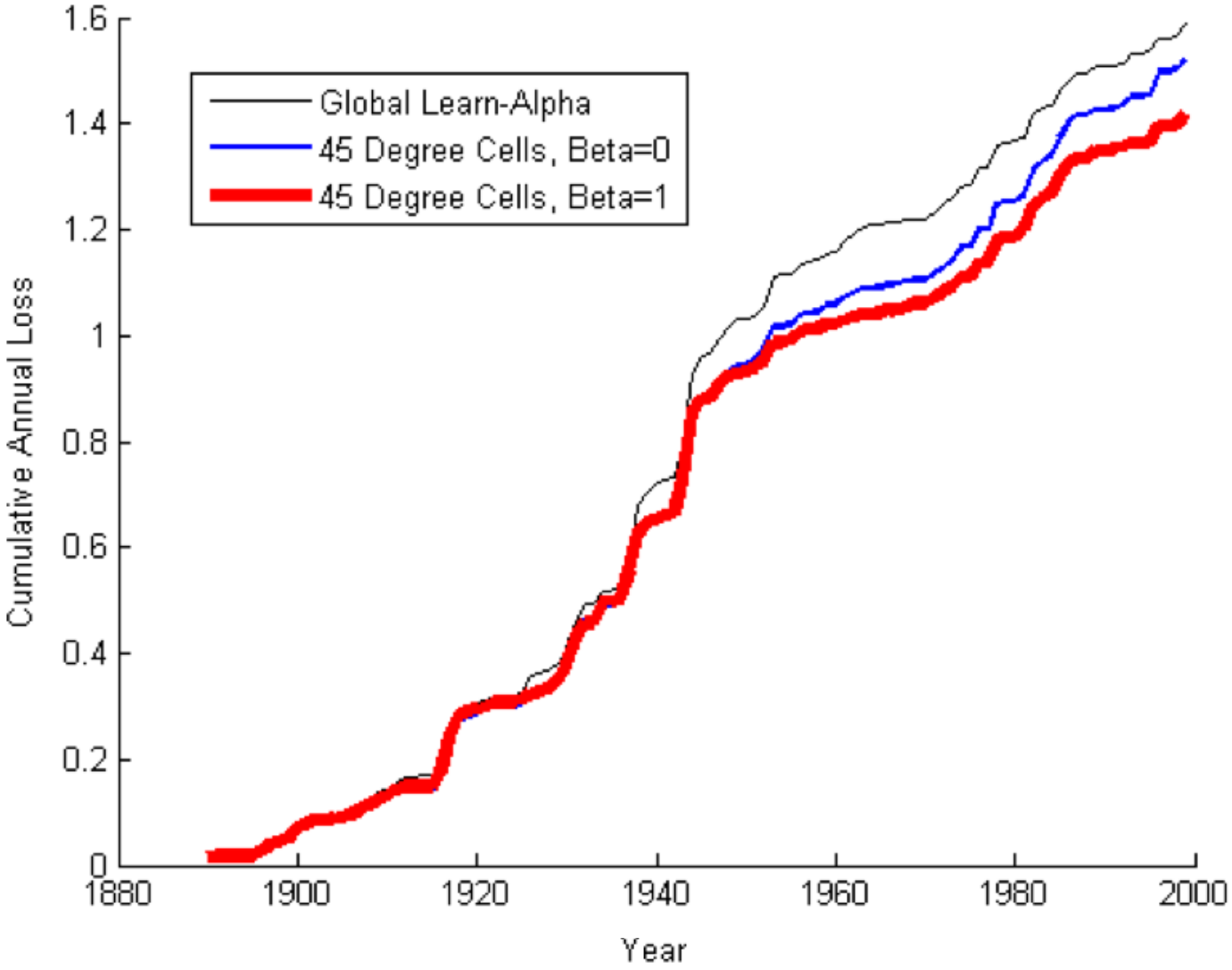
Neighborhood schemes

- Update is modular with respect to neighborhood scheme S(r).
- A possible neighborhood scheme below
  - Could also be continuous set of neighboring regions, e.g. using a Gaussian.



Experimental Setup

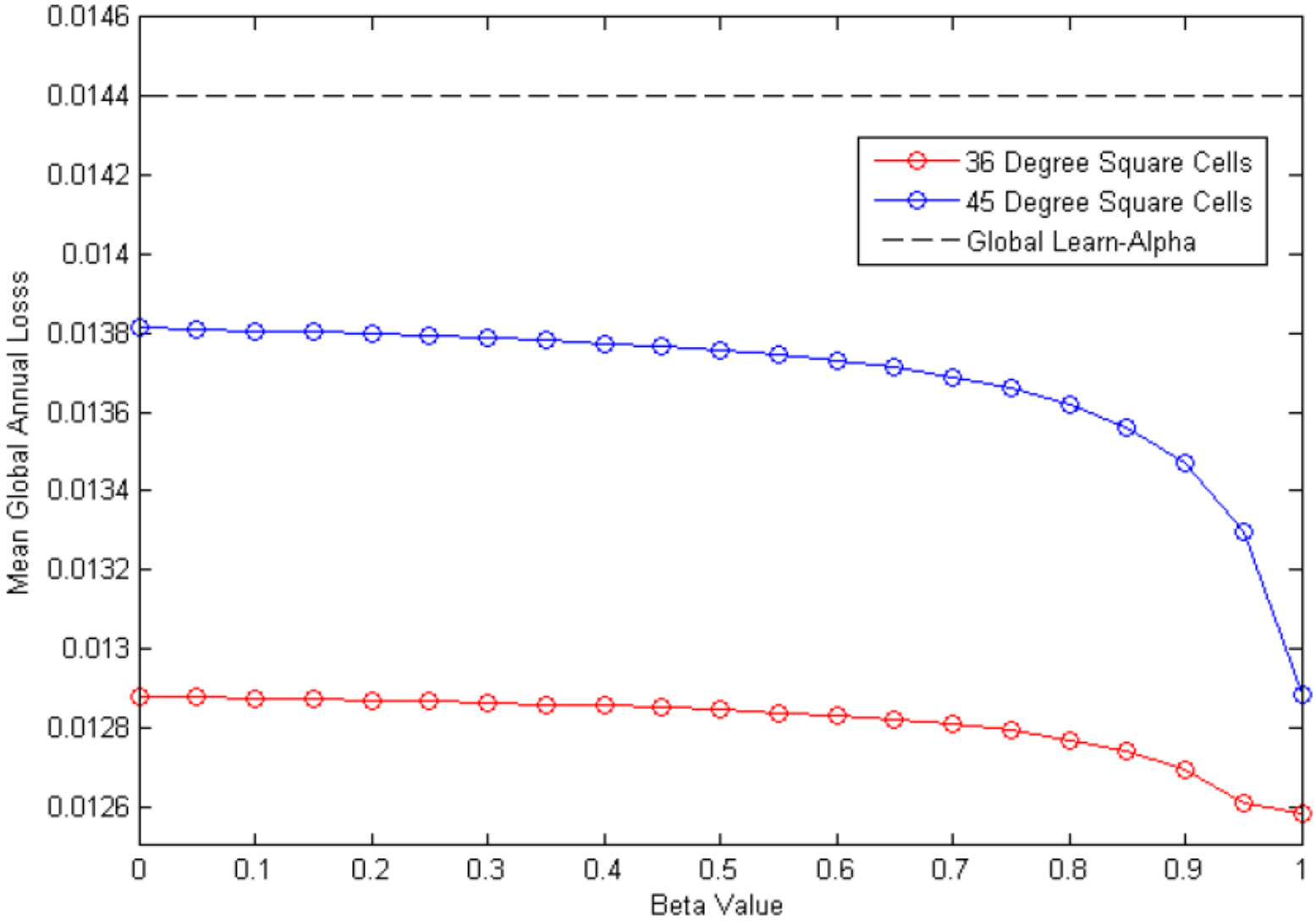
- GCM hindcasts from the years 1890-2000
  - IPCC Phase 3 Coupled Model Intercomparison Project (CMIP3)
  - Climate of the 20th Century Experiment (20C3M)
  - One run from each contributing institution arbitrarily selected
- Observed temperature anomaly data from NASA GISTEMP
- All data converted to temperature anomalies
  - Benchmark period 1951-1980



	Mean Annual Loss	Variance	Cumulative Annual Loss (1890-2000)
Global Learn-α	0.0144	0.0003	1.5879
45 Degree Squares $\beta = 0$	0.0138	0.0004	1.5194
45 Degree Squares $\beta = 1$	0.0129	0.0003	1.4173

Table 1: Cumulative Annual Losses for 45 degree square cells and Global Learn-α.

Results


0.0146
0.0144
0.0142
0.014
0.0138
0.0136
0.0134
0.0132
0.013
0.0128
0.0126
Mean Global Annual Losss
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Beta Value
36 Degree Square Cells
45 Degree Square Cells
Global Learn-Alpha

Challenges in climate modeling

Challenge: Improve the predictions of the multi-model ensemble

- Extensions to Tracking Climate Models
  - Different experts per location; spatial (in addition to temporal) transition dynamics
  - Tracking other climate benchmarks, e.g. carbon dioxide concentrations
- {Semi,un}-supervised learning with experts. Largely open in ML.
- Challenge: try other ML approaches! (e.g. batch, transductive regression?)

Challenge: Improve the predictions of a climate model

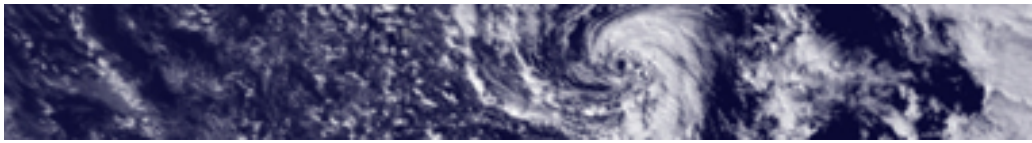
- Challenge: resolve scale interactions ("climate model parameterization")
- Challenge: harness both physics and data!
  - Hybrid methods
  - Data assimilation
- Challenge: Calibrating and comparing climate models in a principled manner

More challenges in Climate Informatics

Challenge: Clustering / detecting spatiotemporal patterns

e.g. droughts, cyclones

- Algorithms for streaming and online clustering
- Graphical model approaches, e.g. from topic modeling



Challenge: Building theoretical foundations for Climate Informatics

- Coordinating on reasonable assumptions in practice, that allow for the design of theoretically justified learning algorithms

Challenge: tracking polar ice melt from satellite image data

Challenge: short term climate prediction

Challenge: regional climate prediction

:

Thank You!

And thanks to my coauthors:

"Tracking Climate Models"

Gavin Schmidt, *NASA GISS & Columbia University*
Shailesh Saroha, *Amazon*
Eva Asplund, *Columbia University*

"Global Climate Model Tracking using Geospatial Neighborhoods"

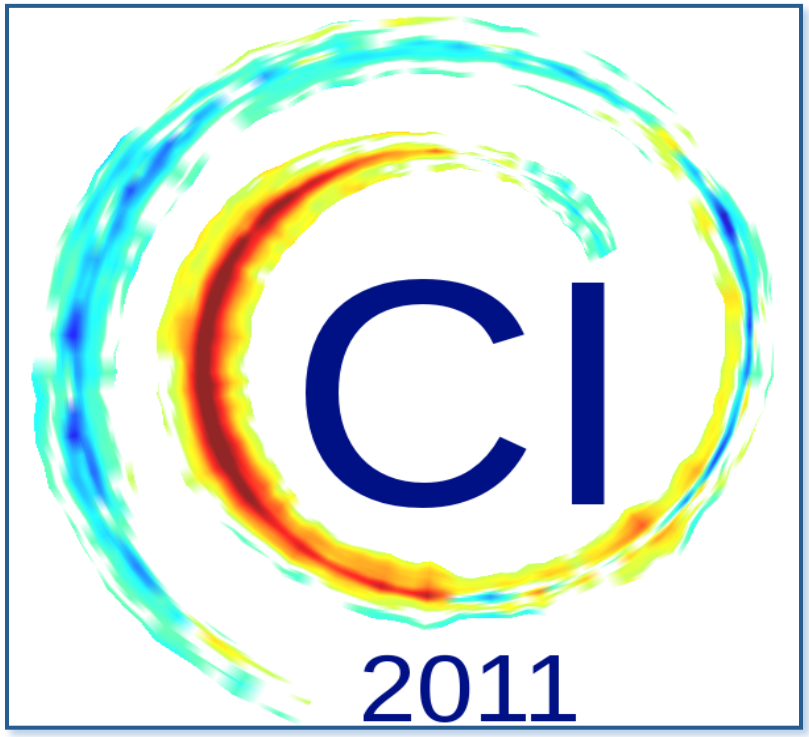
Scott McQuade, *George Washington University*

"Climate Informatics"

Gavin Schmidt, *NASA GISS & Columbia University*
Frank Alexander, *Los Alamos National Laboratory*
Alex Niculescu-Mizil, *NEC Laboratories America*
Karsten Steinhaeuser, *University of Minnesota*
Michael Tippett, *The International Research Institute for Climate and Society, Columbia*
Arindam Banerjee, *University of Minnesota*
M. Benno Blumenthal, *International Research Institute for Climate and Society, Columbia*
Auroop R. Ganguly, *Civil and Environmental Engineering, Northeastern University*
Jason E. Smerdon, *Lamont-Doherty Earth Observatory, Columbia University*
Marco Tedesco, *CUNY City College and Graduate Center*

Climate Informatics wiki: http://sites.google.com/site/1stclimateinformatics

data, links, and challenge problems



LDEO/GISS CLIMATE CENTER

The New York Academy of Sciences

Information Science and Technology Center at Los Alamos National Laboratory

NEC Laboratories America
Relentless passion for innovation

Columbia University Department of Statistics

YAHOO! LABS

COLUMBIA CCLS

Climate Informatics 2011 *August 2011, New York Academy of Sciences*

Co-Chairs: Claire Monteleoni & Gavin Schmidt

Organizing Committee:

- Frank Alexander
- Alex Niculescu-Mizil
- Karsten Steinhaeuser
- Michael Tippett

Climate Informatics 2012

National Science Foundation
WHERE DISCOVERIES BEGIN

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

Los Alamos
NATIONAL LABORATORY

THE CLIMATE CORPORATION

CRAY
THE SUPERCOMPUTER COMPANY

Organizing Committee:

Co-Chairs:
Claire Monteleoni
Karsten Steinhaeuser
Advisor:
Gavin Schmidt
Local Chairs:
Doug Nychka
Steve Sain
Program Chairs:
Arindam Banerjee
Jason Smerdon
Breakouts Chair:
Jim Gattiker
Communications Chair:
Evan Kodra

Program Committee:

Kevin Anchukaitis
Nitesh Chawla
Julien Emile-Geay
Bo Li
Yan Liu
Aurelie Lozano
Bala Rajaratnam
Padhraic Smyth
Martin Tingley
Wim Wiegerinch