A Metrics Framework for Interannual-to-Decadal Predictions Experiments

L. Goddard, on behalf of the
US CLIVAR Decadal Predictability Working Group & Collaborators:
Lisa Goddard, Arun Kumar, Amy Solomon, James Carton, Clara Deser,
Ichiro Fukumori, Arthur M. Greene, Gabriele Hegerl, Ben Kirtman,
Yochanan Kushnir, Matthew Newman, Doug Smith, Dan Vimont,
Tom Delworth, Jerry Meehl, and Timothy Stockdale
Paula Gonzalez, Simon Mason, Ed Hawkins, Rowan Sutton,
Rob Bergman, Tom Fricker, , Chris Ferro, David Stephenson

US CLIVAR Decadal Predictability Working Group

Formally approved January 2009

Objective 1: *To define a framework to distinguish natural variability from anthropogenically forced variability on decadal time scales for the purpose of assessing predictability of decadal-scale climate variations in coupled climate models.*

Objective 2: *Work towards better understanding of decadal variability and predictability through metrics that can be used as a strategy to assess and validate decadal climate prediction simulations.*

Proposed FRAMEWORK for Verification:

1. **Feasibility** (of particular model/fcst system)
   - Realistic, and relevant, variability?
   - Translation of ICs to realistic and relevant variability?

2. **Prediction skill** – Quality of system; quality of information
   - Where? What space & time scales?
   - Actual anomalies & 'decadal scale trends'
   - Conditional skill?
   - Values of ICs: higher correlations, lower RMSEs

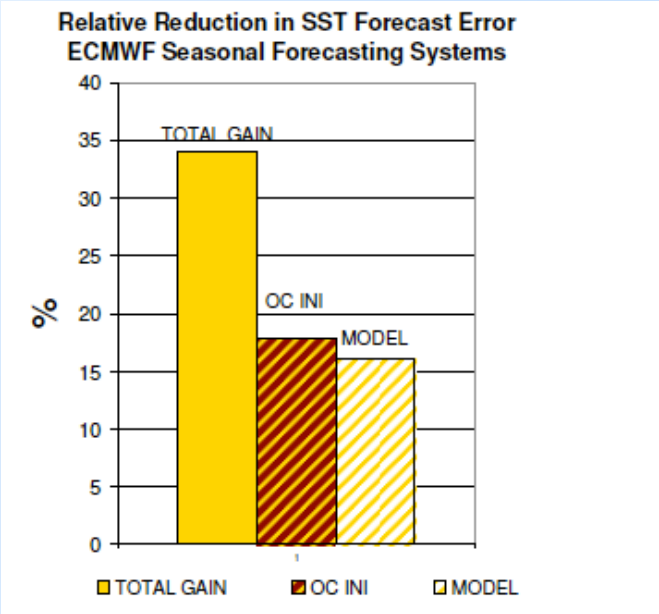
3. **Issues** – for research, for concern
   i.e. limited ability to quantify uncertainty;
   limited understanding of processes, etc.

Outline

- Objective
- Framework
  - Metrics & examples of results
  - Statistical significance
  - Website
- Issues relevant to verification endeavor
  - Bias correction
  - Spatial scale
  - Stationarity/reference period

Motivation: Forecasts need verification

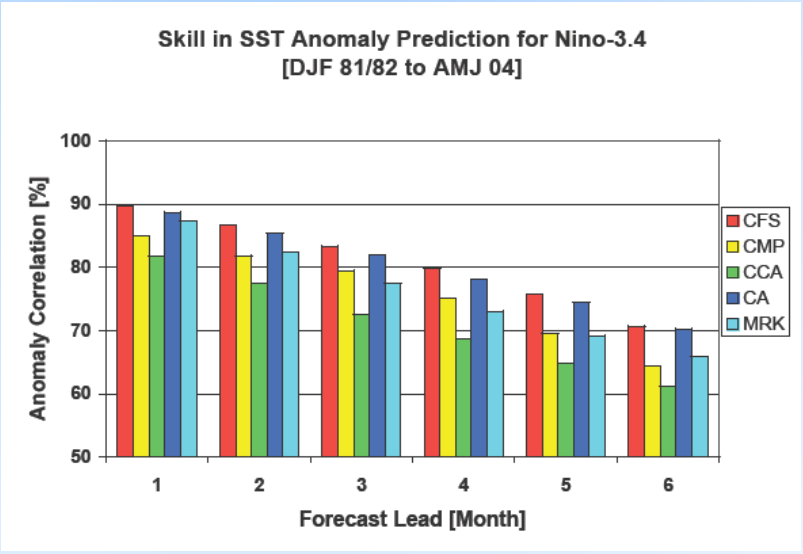
… for tracking improvements in prediction systems



Example from SI:
Recent improvements to ECMWF seasonal forecast system came in almost equal parts from improvements to the model and the ODA

(Balmaseda et al. 2009, OceanObs'09)

… for comparison against other systems and other approaches



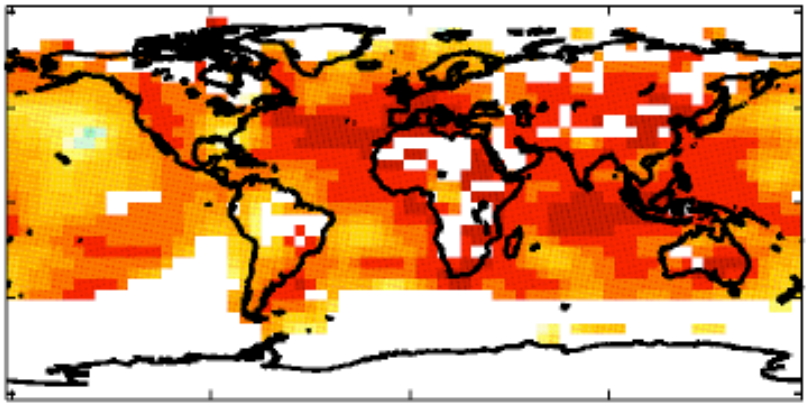
Example from SI:
NCEP-CFS reaches parity with statistical fcsts for ENSO

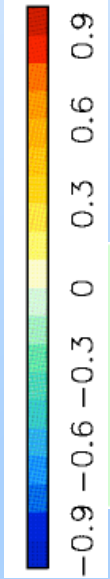
(Saha et al. 2004, J.Clim)

How "good" are they?: Deterministic Metrics

Regional Average (15°x15°); 5-Year Means:

DePreSys anomaly correlation



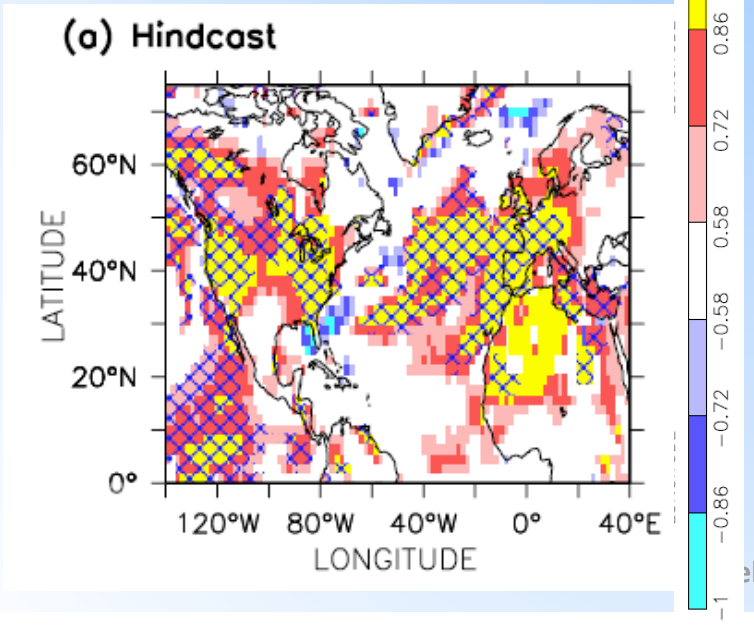


Met Office Hadley Centre

- HadCM3
- 9 member perturbed physics ensemble
- Starting every Nov from 1960 to 2005

(Courtesy: Doug Smith)

Grid Scale; 10-Year Means:



- ECHAM5 + MPI-OM
- 3 member perturbed IC ensemble
- Starting every 5 years Nov from 1955 to 2005

(Keenlyside et al. 2008, Nature)

el decadal prediction experiments from CMIP5

U.S. CLIVAR Climate Variability & Predictability

Outline

- Objective
- Framework
  - Metrics & examples of results
  - Statistical significance
  - Website
- Issues relevant to verification endeavor
  - Bias correction
  - Spatial scale
  - Stationarity/reference period

Asking Questions of the Initialized Hindcasts

Question 1: Do the initial conditions in the hindcasts lead to more accurate predictions of the climate?

Question 2: Is the model's ensemble spread an appropriate representation of forecast uncertainty on average?

Question 3: In the case that the forecast ensemble does offer information on overall forecast uncertainty, does the forecast-to-forecast variability of the ensemble spread carry meaningful information?

Time scale: Year 1, Years 2-5, Years 2-9

Spatial scale: Grid scale, spatially-smoothed

Asking Questions of the Initialized Hindcasts

Question 1: Do the initial conditions in the hindcasts lead to more accurate predictions of the climate?

→ Mean Squared Skill Score and its decomposition

$$MSSS = \frac{MSE_{ref} - MSE_{fcst}}{MSE_{ref}} = 1 - \frac{MSE_{fcst}}{MSE_{ref}} = 1 - \frac{MSE_{init}}{MSE_{uninit}}$$

$$MSSS(f,\bar{x},x) = r_{fx}^2 - [r_{fx} - \left(s_x / s_f\right)]^2$$

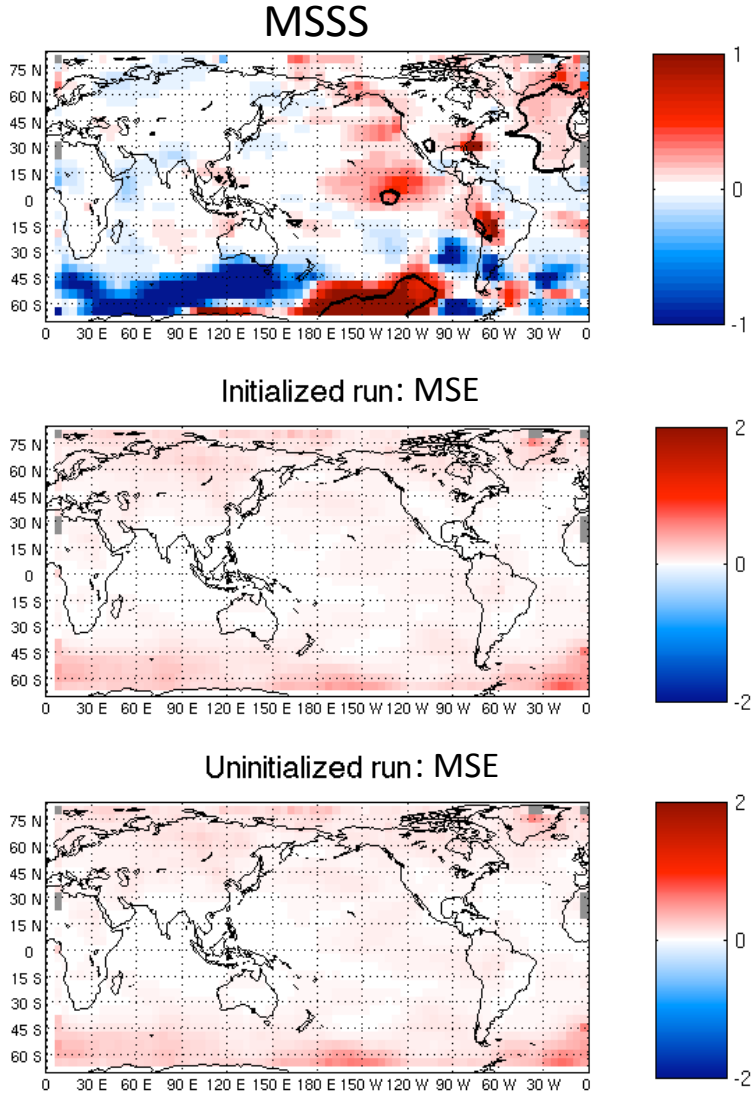
$$MSSS(f,r,x) = r_{fx}^2 - [r_{fx} - \left(s_x / s_f\right)]^2 - r_{rx}^2 - [r_{rx} - \left(s_x / s_r\right)]^2$$

$$MSSS(f,r,x) = MSSS(f,\bar{x},x) - MSSS(r,\bar{x},x)$$

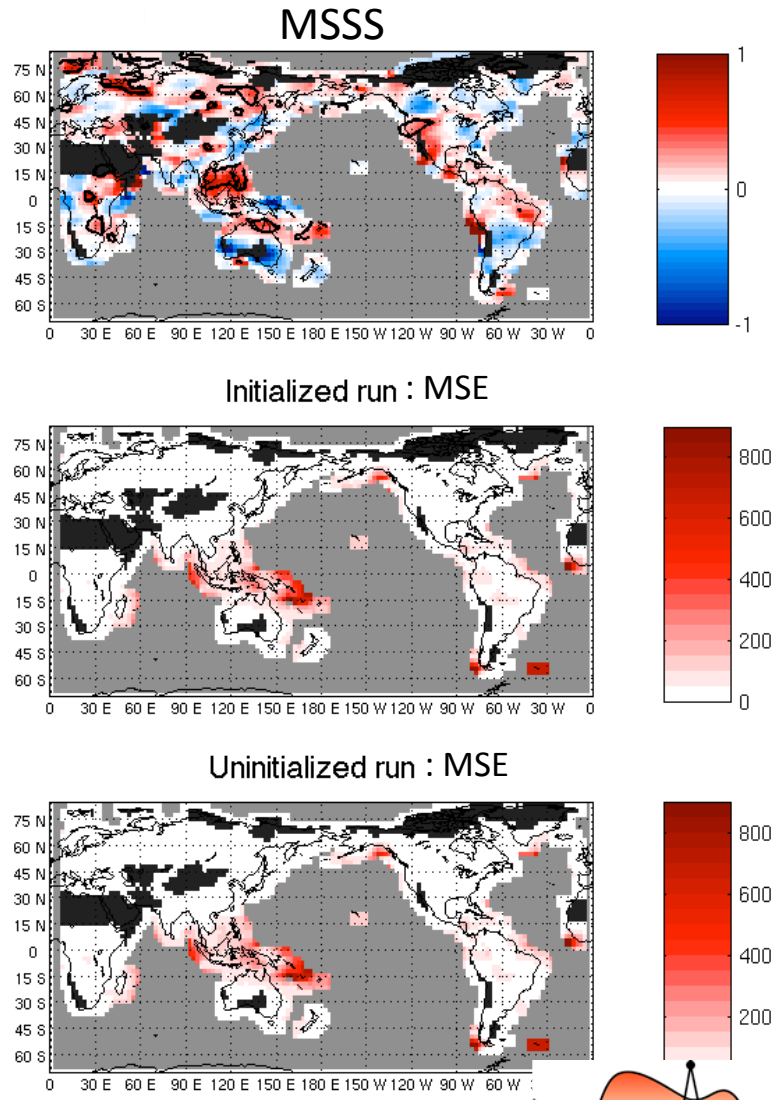
(from Murphy, Mon Wea Rev, 1988)

Deterministic Metrics: Mean Squared Skill Score (MSSS)

Hadley Centre MSSS: Years 2-9 - HadCRUT3v smooth temp. anoms

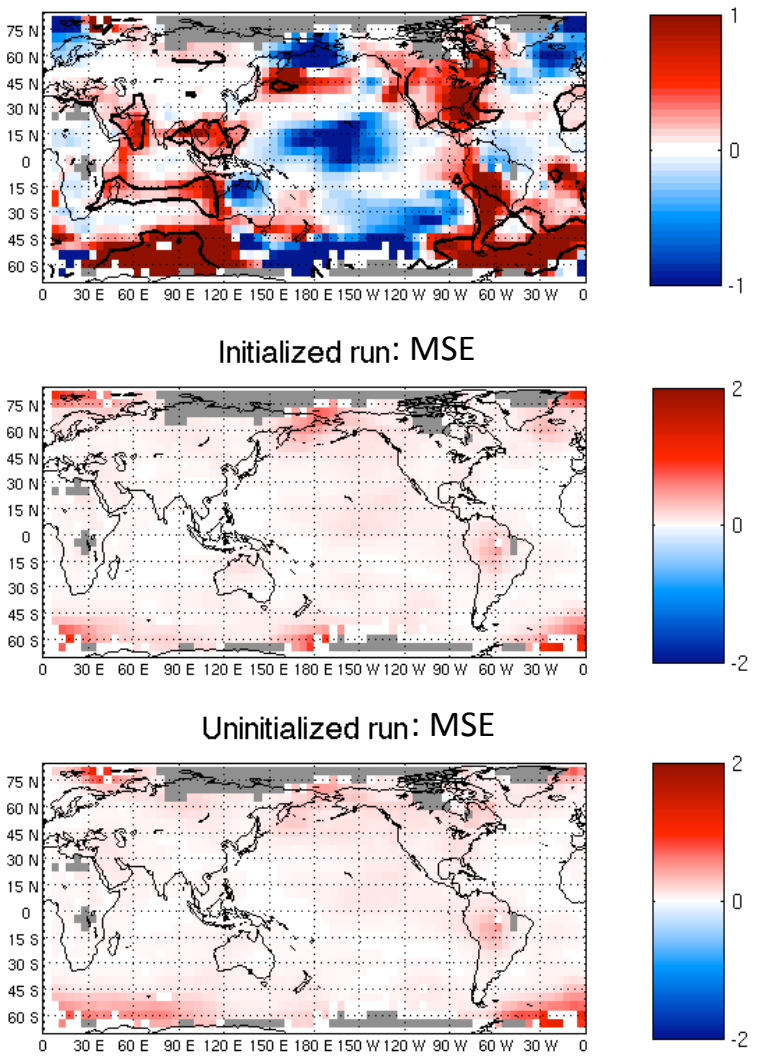


Hadley Centre MSSS: Years 2-9 - GPCC smooth precip. anoms.

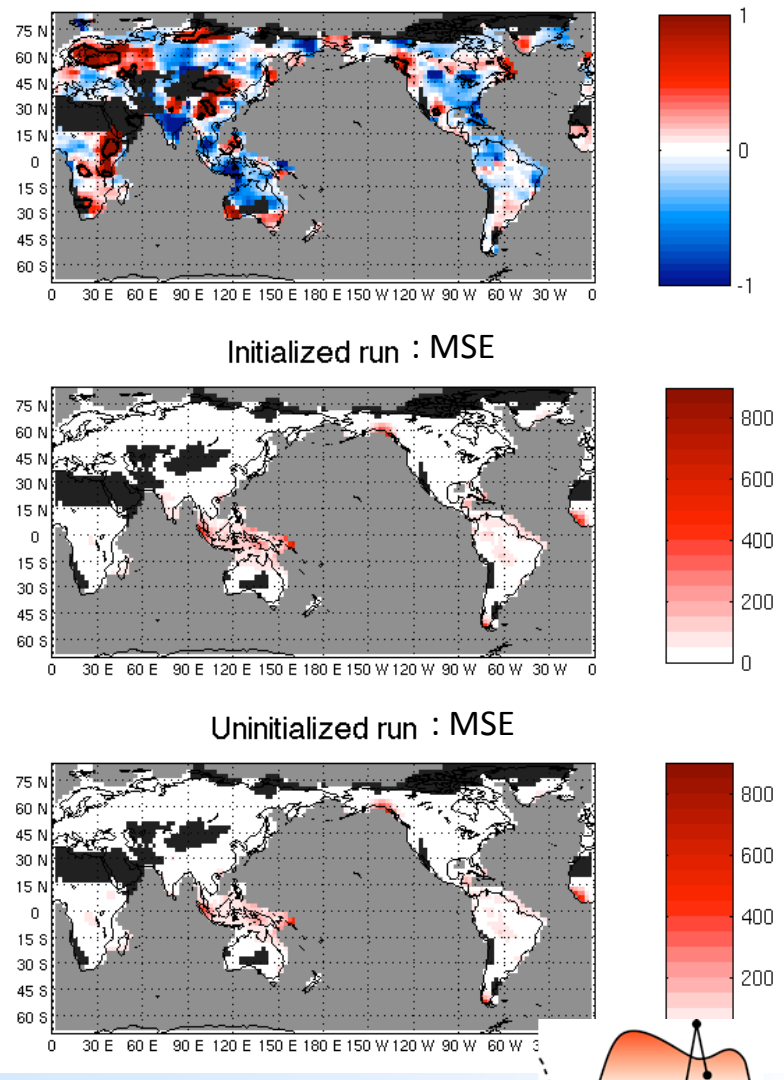


Deterministic Metrics: Mean Squared Skill Score (MSSS)

CCCma MSSS: Years 2-9 - HadCRUT3v smooth temp. anoms.

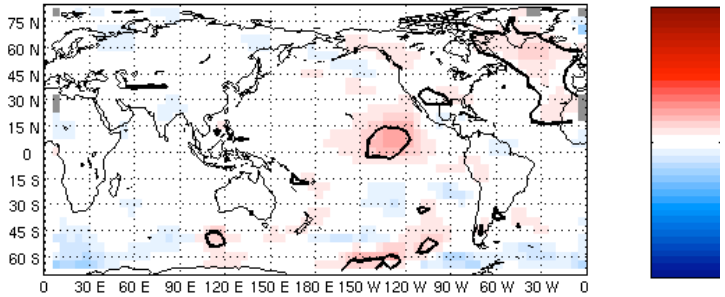


CCCma MSSS: Years 2-9 - GPCC smooth precip. anoms.

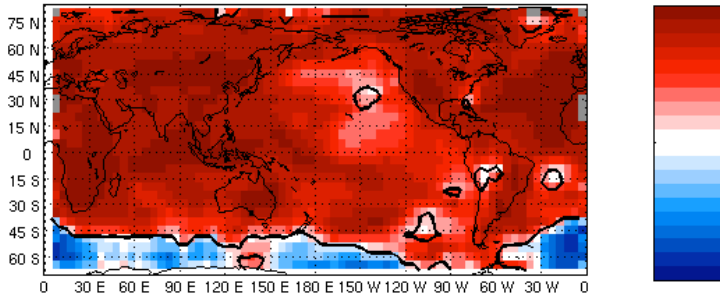


Deterministic Metrics: Anomaly Correlation

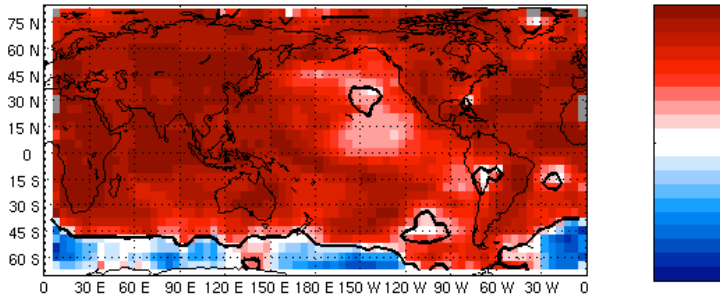
Hadley Centre Correlation: Years 2-9 - HadCRUT3v smooth temp. Diff. Initialized-Uninitialized



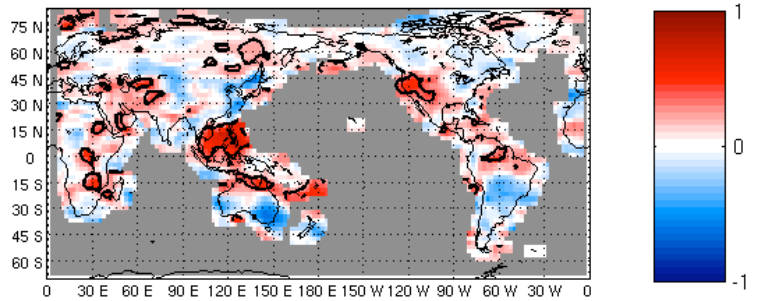
Initialized run



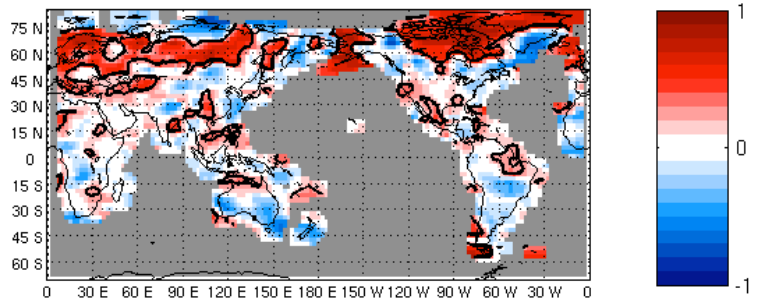
Uninitialized run



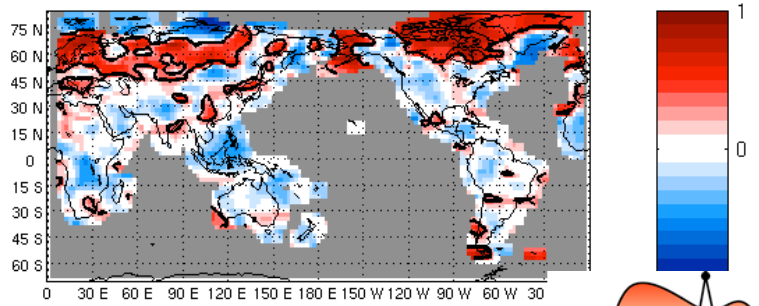
Hadley Centre Correlation: Years 2-9 - GPCC smooth precip. anoms. Diff. Initialized-Uninitialized



Initialized run

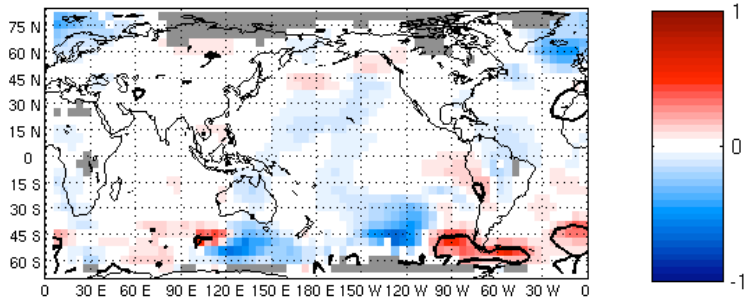


Uninitialized run

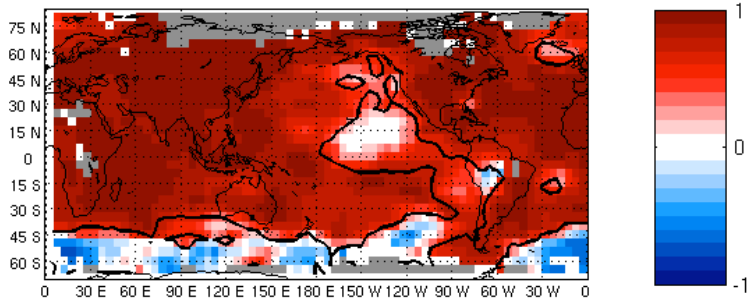


Deterministic Metrics: Anomaly Correlation

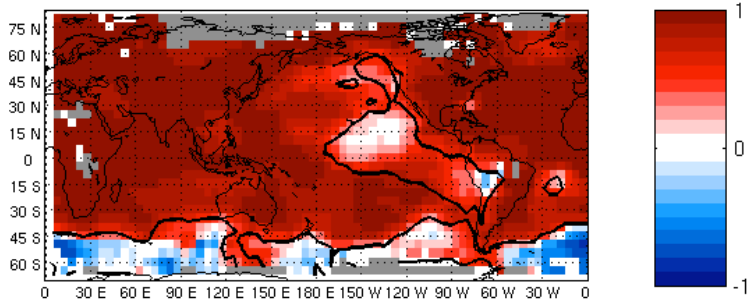
CCCma Correlation: Years 2-9 - HadCRUT3v smooth temp. anoms.
Diff. Initialized-Uninitialized



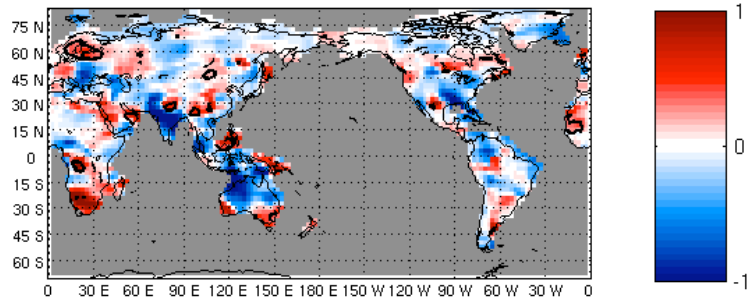
Initialized run



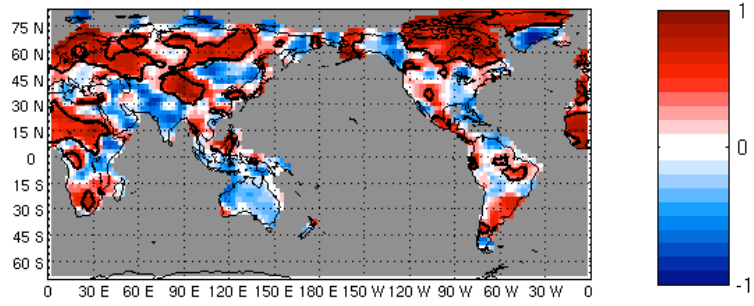
Uninitialized run



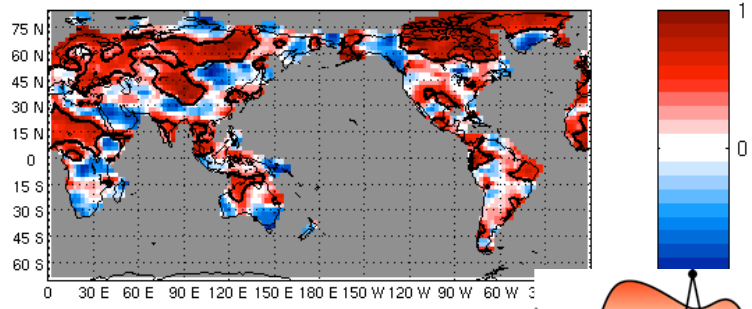
CCCma Correlation: Years 2-9 - GPCC smooth precip. anoms.
Diff. Initialized-Uninitialized



Initialized run

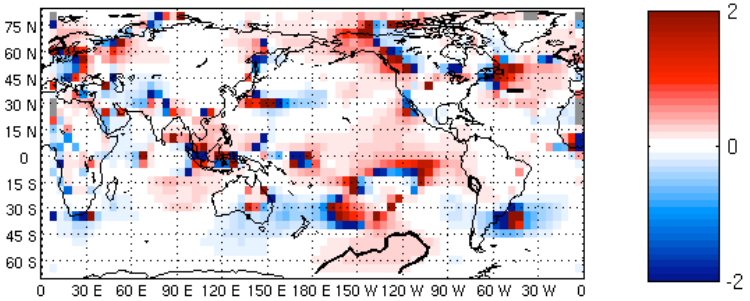


Uninitialized run

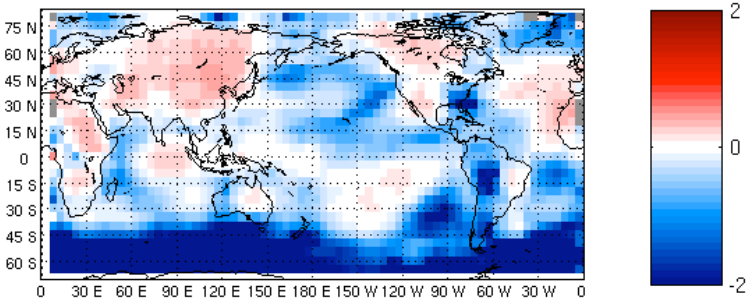


Deterministic Metrics: Conditional Bias

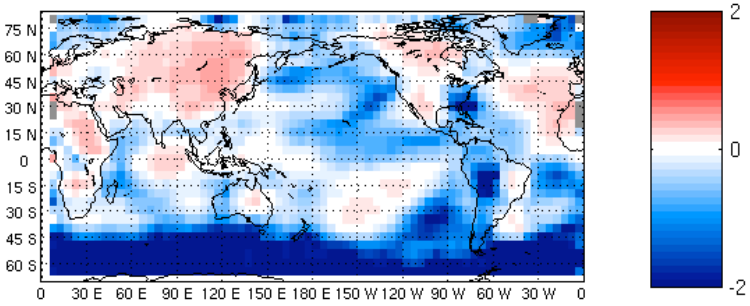
Hadley Centre Conditional Bias: Years 2-9 - HadCRUT3v smooth temp. anoms.
Fractional decrease in conditional bias



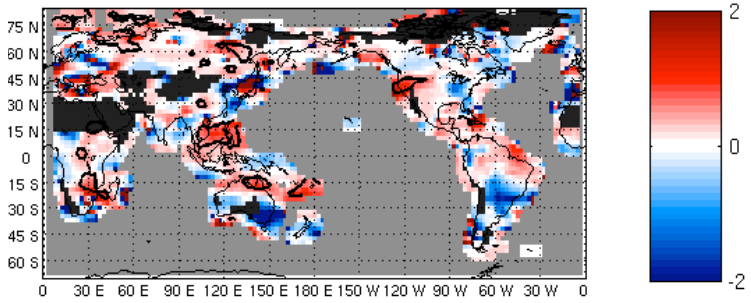
Initialized run



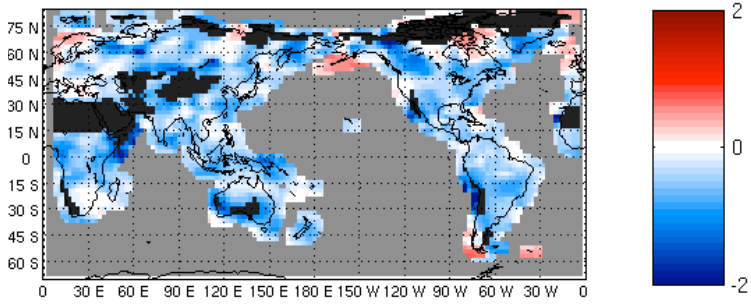
Uninitialized run



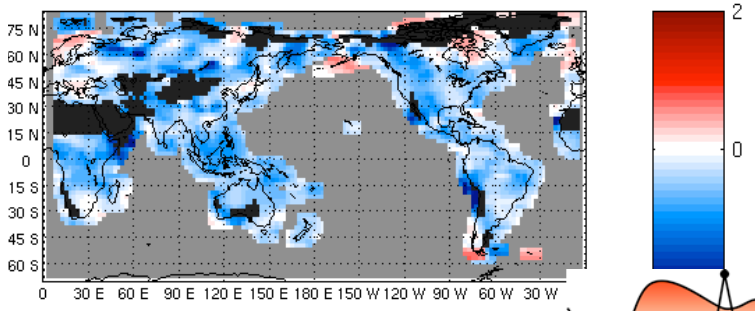
Conditional Bias: Years 2-9 - GPCC smooth precip. anoms.
Fractional decrease in conditional bias



Initialized run

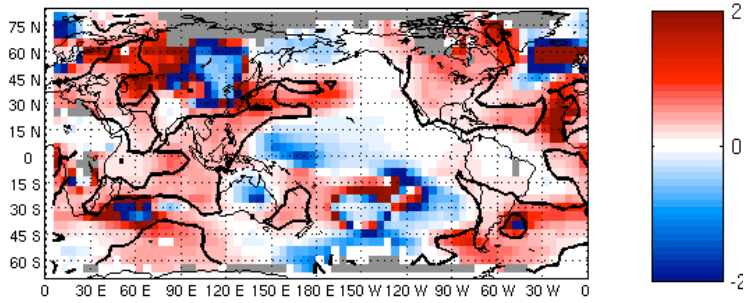


Uninitialized run

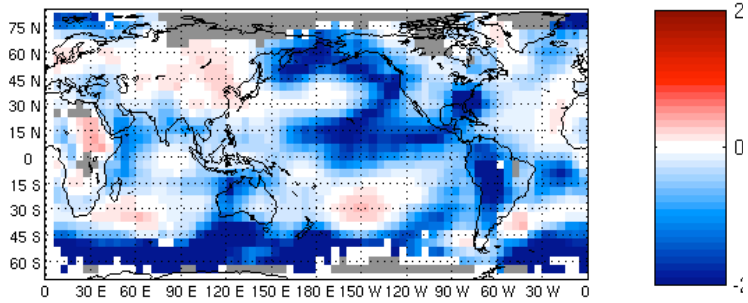


Deterministic Metrics: Conditional Bias

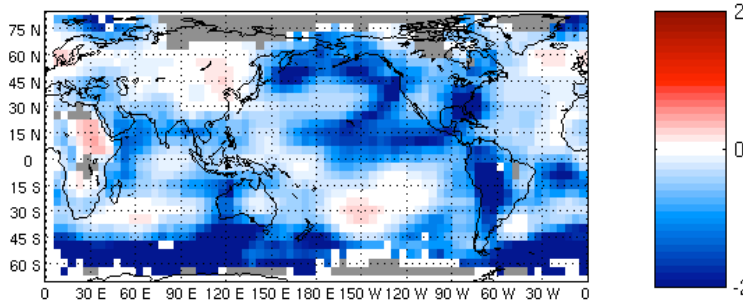
CCCma Conditional Bias: Years 2-9 - HadCRUT3v smooth temp. an
Fractional decrease in conditional bias



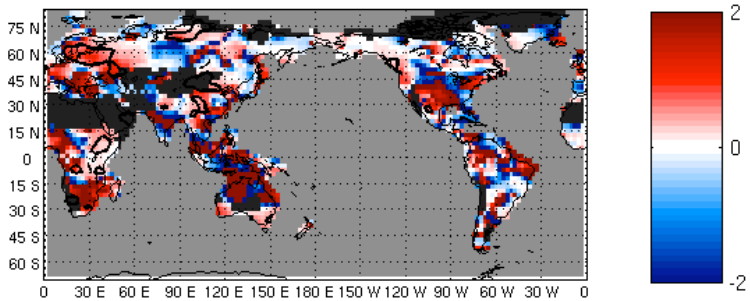
Initialized run



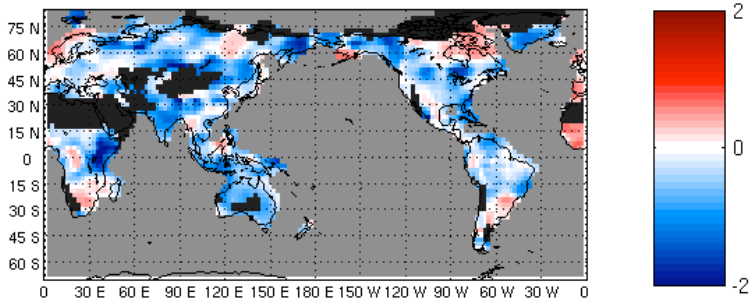
Uninitialized run



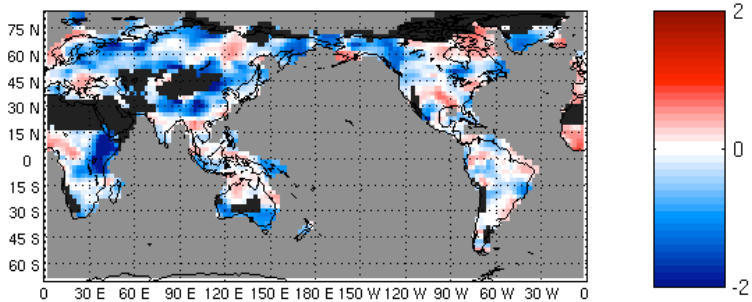
CCCma Conditional Bias: Years 2-9 - GPCC smooth precip. anoms.
Fractional decrease in conditional bias



Initialized run



Uninitialized run



Asking Questions of the Initialized Hindcasts

Question 2: Is the model's ensemble spread an appropriate representation of forecast uncertainty on average?

Question 3: In the case that the forecast ensemble does offer information on overall forecast uncertainty, does the forecast-to-forecast variability of the ensemble spread carry meaningful information?

→ Continuous Ranked Probability Skill Score (CRPSS)

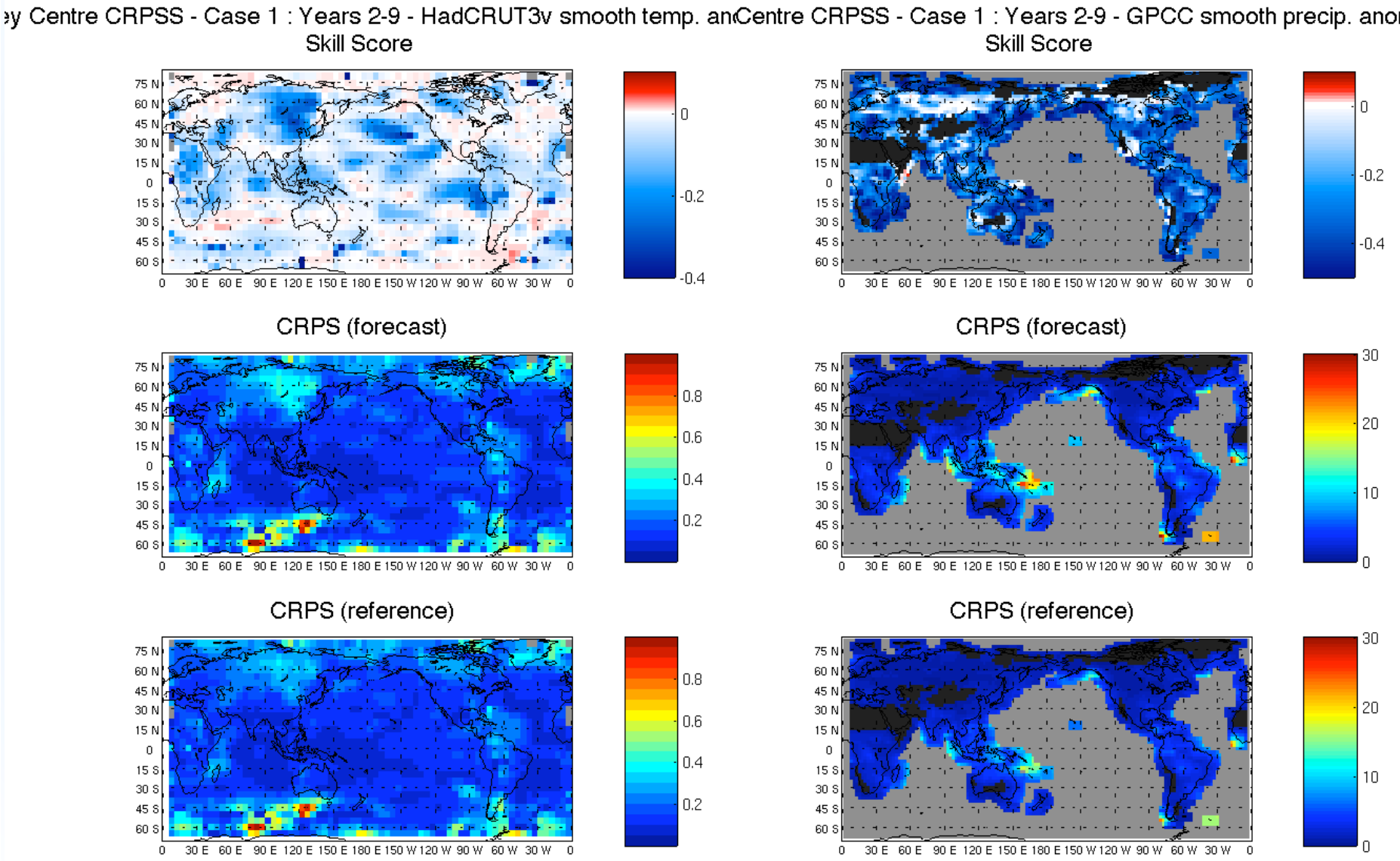
$CRPSS = 1 - (CRPS_{fcst}/CRPS_{ref})$

Q2: **fcst** uncertainty = avg ensemble spread
ref uncertainty = standard error of ensemble mean

Q3: **fcst** uncertainty = time varying ensemble spread
ref uncertainty = avg ensemble spread

Probabilistic Metrics: CRPSS

(Case 1: Ens Spread vs. Std Err)



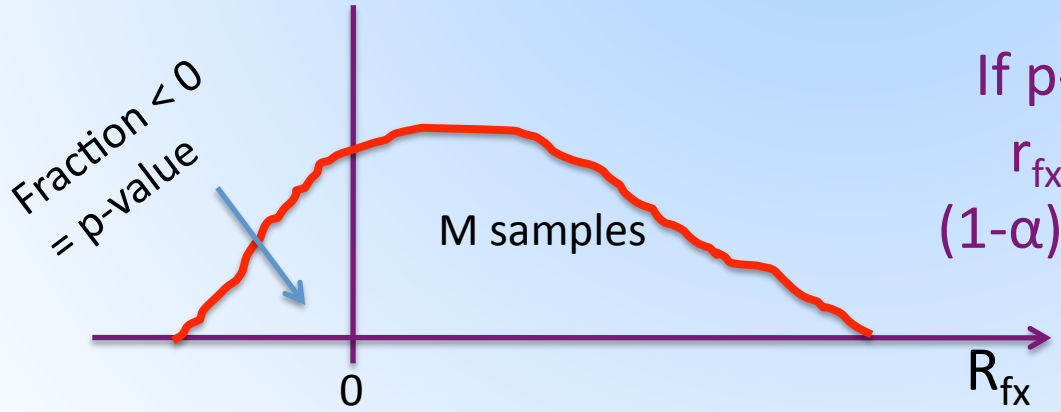
Statistical Significance: Non-parametric bootstrap

Re-sampling, with replacement: k=1,M (~1000) samples

Start out with nominally n=10 start times.
Draw random start times as pairs up to n values.
i.e. 1st draw: i=1 → e.g. I(i,k)=5 (1980), so i=2 → I(i+1,k)=6, etc.
up to i=10

For each I(i,k), draw N random ensemble members, E, with replacement

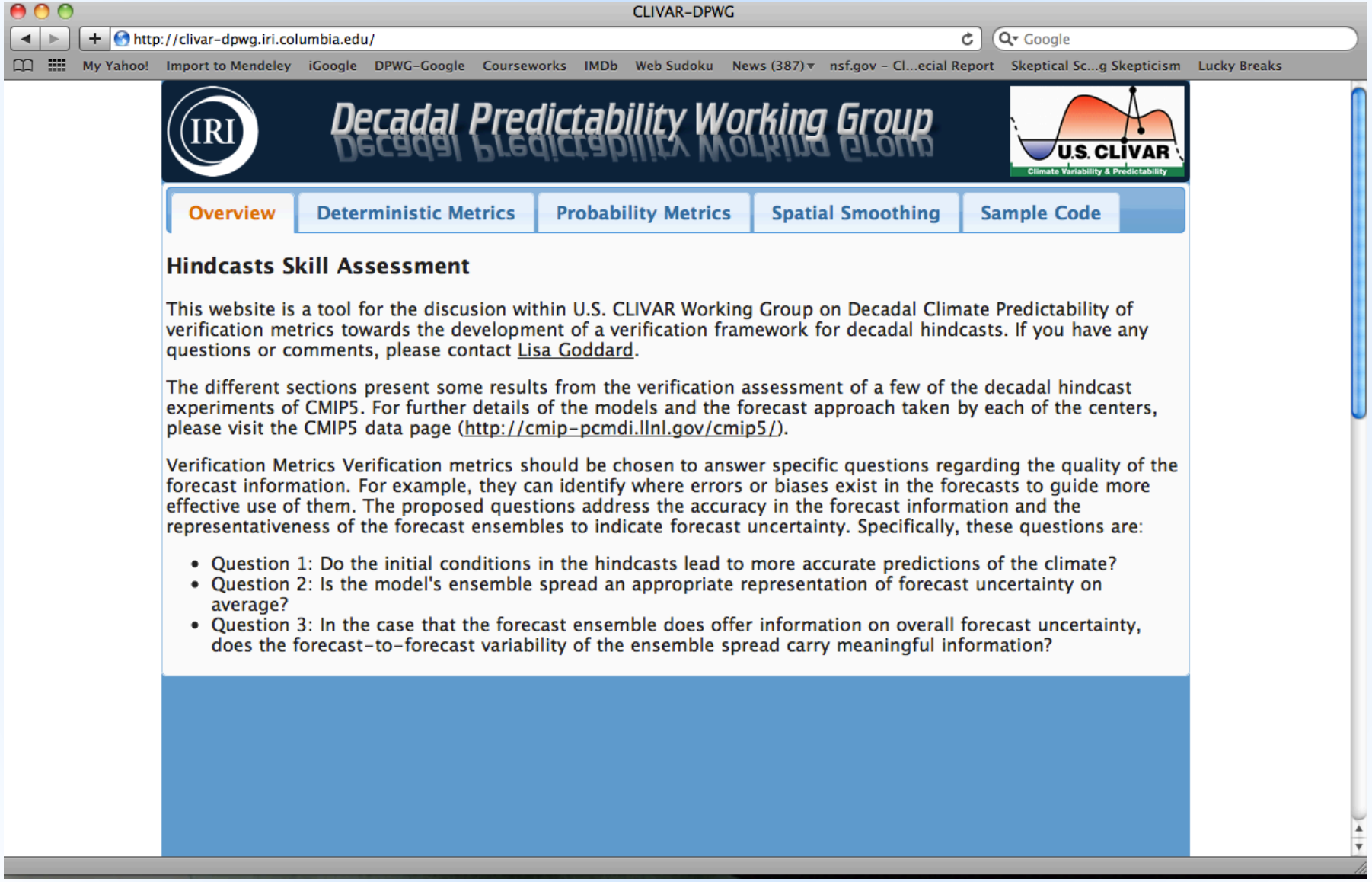
$$\tilde{f}_i^E(k) = f_{I(i,k)}^{E(I)}$$



If p-value <= α, then
r_{fx} is significant at
(1-α)x100% confidence

Proto-type Website: *Work in progress*

http://clivar-dpwg.iri.columbia.edu



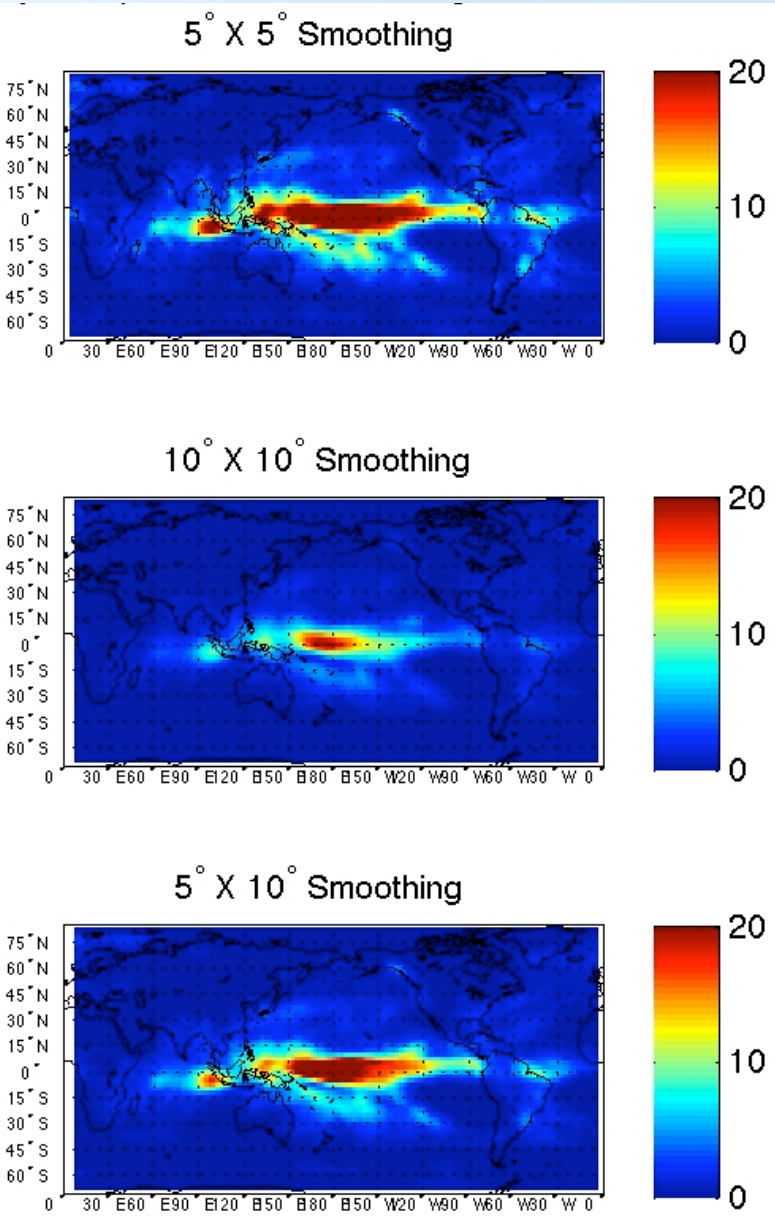
Outline

- Objective
- Framework
  - Metrics & examples of results
  - Statistical significance
  - Website
- Issues relevant to verification endeavor
  - Bias correction
  - Spatial scale
  - Stationarity/reference period

Issues relevant to verification

- Spatial scale for verification
- Bias (mean and conditional)
  - *Mean bias MUST be removed prior to use or verification of forecasts* (WCRP, 2011)
- Forecast uncertainty
  - *Conditional bias MUST be removed prior to assigning forecast intervals*
- Stationarity / reference period

Spatial Scale: Signal-to-noise



GPCP Precipitation Anomalies

Based on de-correlation scales, and S2N considerations, advocating

<u>Temperature smoothing</u>:
15 longitude x 10 latitude

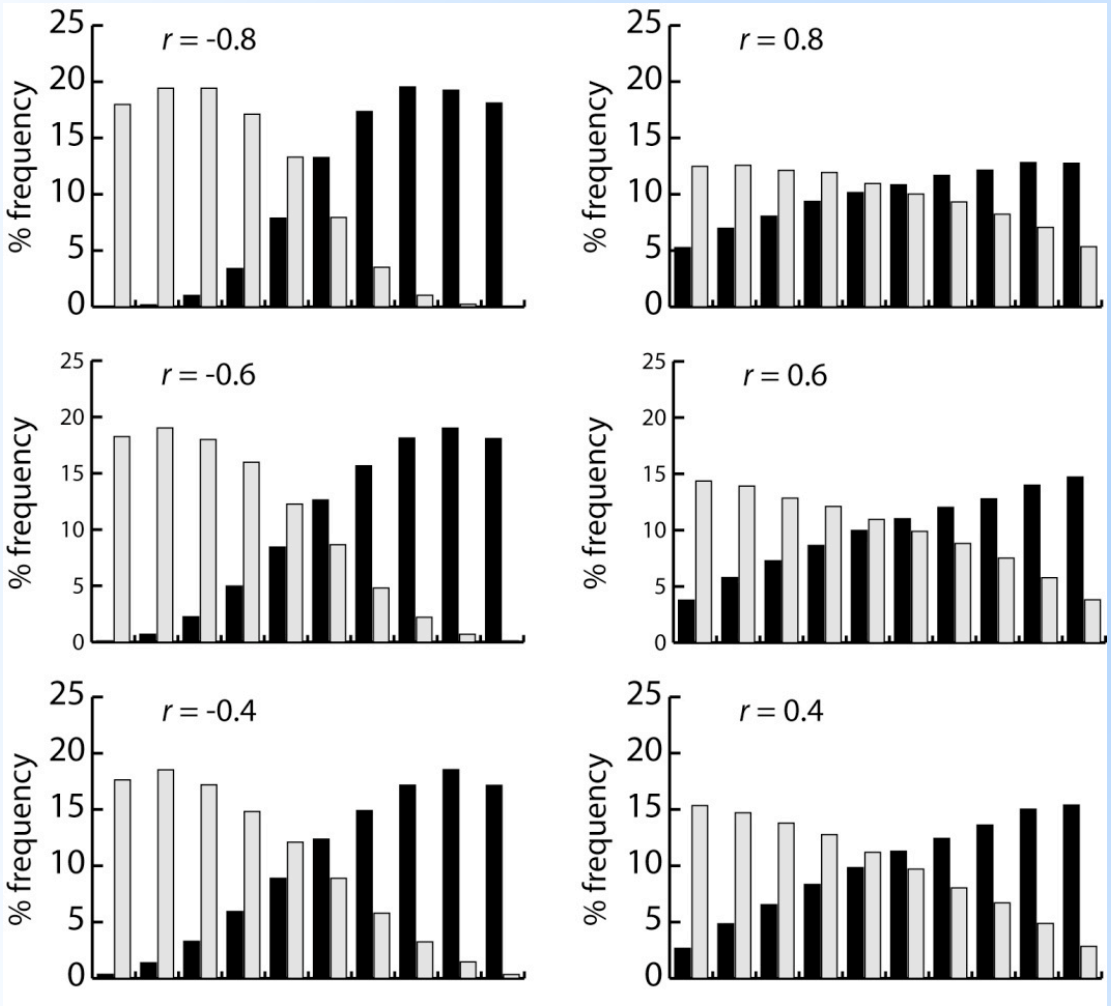
<u>Precipitation smoothing:</u>
10 longitude x 5 latitude

(Goddard, Gonzalez, & Jensen, *in prep*)

diction experiments from CMIP5

U.S. CLIVAR
Climate Variability & Predictability

Effect of Conditional Bias on Reliability



Conditional rank histograms

- 9-member ensemble fcsts
- Normally distributed variable
- Ensemble-mean variance = Observed variance
- Ensemble spread = MSE

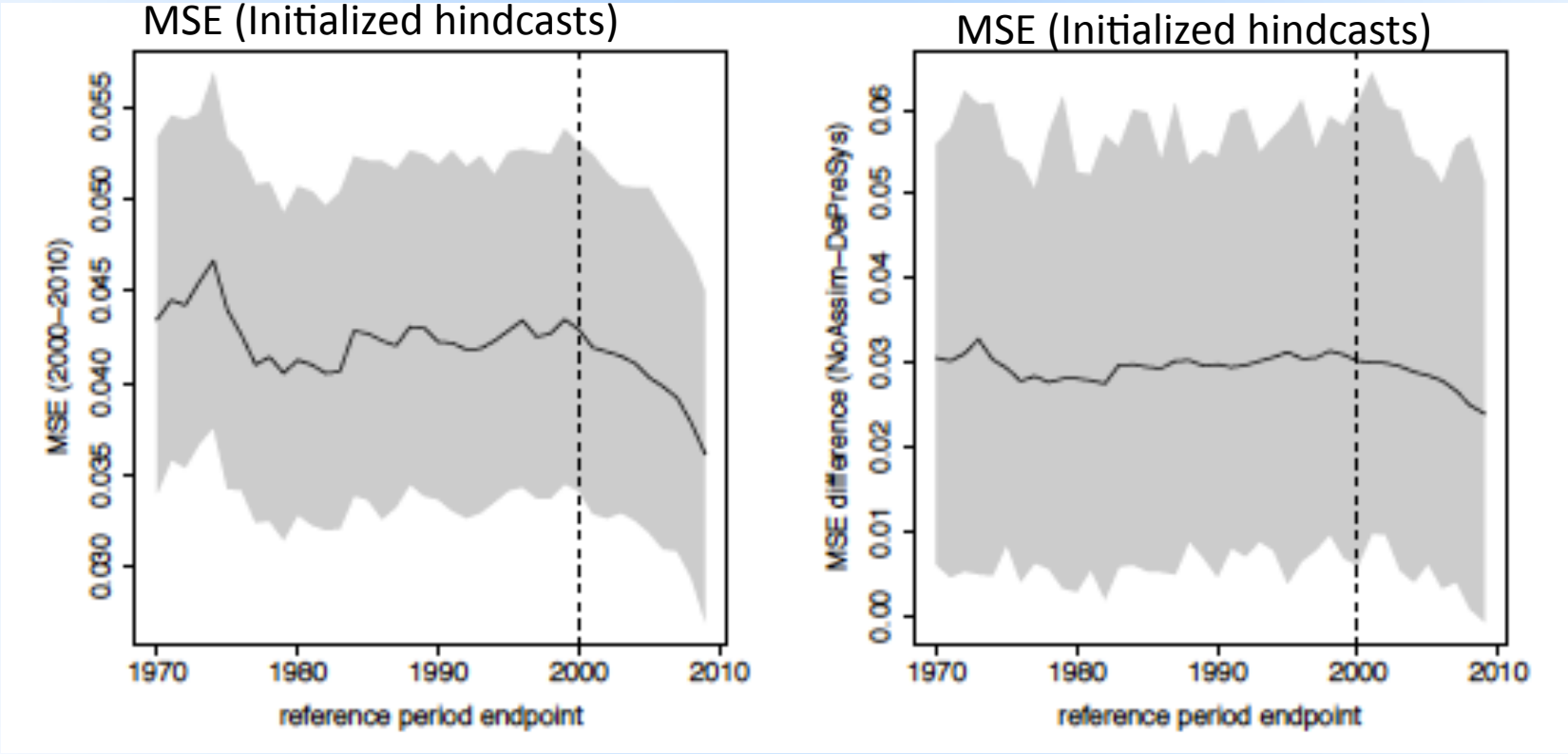
Grey bars are positive anomalies
Black bars are negative anomalies

(Mason, Goddard, and Gonzalez, *in prep.*)

Issues: Non-stationarity

Effect of out-of-sample reference period (pre-2000) vs in-sample (post-2000)

MSE of Global Mean Temperatures for 2001-2010)|
Reference Period = 1950 - endpoint



(Fricker, Ferro, Stephenson, *in prep.*)

Summary

US CLIVAR Working Group on Decadal Predictability has developed a framework for verification of decadal hindcasts that allows for common observational data, metrics, temporal structure, spatial scale, and presentation

The framework is oriented towards addressing specific questions of the hindcast quality and suggestions for how they might be used.

Considerable complementary research has aided this effort in areas of bias and forecast uncertainty, spatial scale of the information, and stationarity impacts on reference period.