

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.*

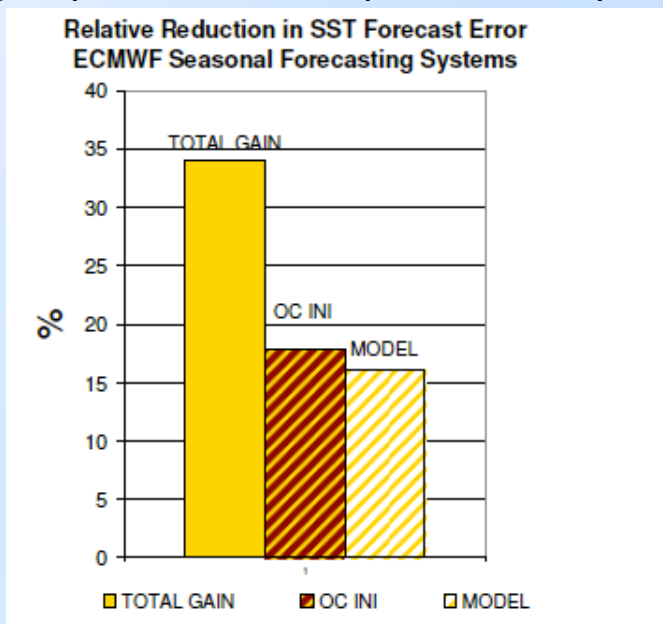
Objective 2: *Work towards better understanding of decadal variability and predictability through metrics that can be used as a strategy to assess and validate decadal climate prediction experiments.*

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



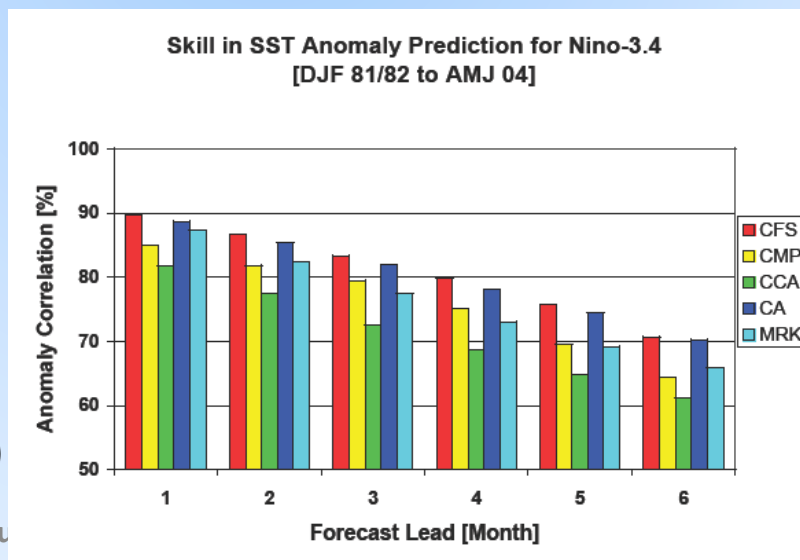
Example from SI:

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

(Saha et al. 2004, J.Clim)

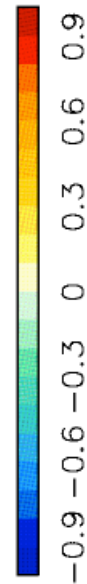
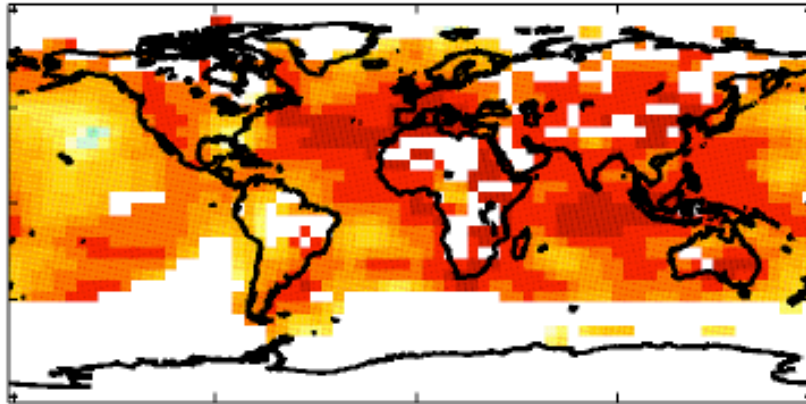


Example from SI:
NCEP-CFS reaches parity with statistical fcsts for ENSO

How “good” are they?: Deterministic Metrics

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

DePreSys anomaly correlation

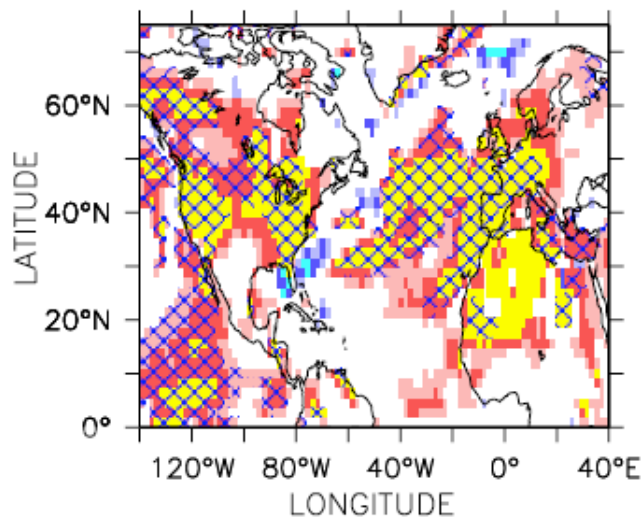


- HadCM3
- 9 member perturbed physics ensemble
- Starting every Nov from 1960 to 2005

(Courtesy: Doug Smith)

Grid Scale; 10-Year Means:

(a) Hindcast



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

(Keenlyside et al. 2008, Nature)

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?

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

$$MSSS = 1 - \frac{MSE_{fcst}}{MSE_{ref}}; \text{ if ref = climatological avg.}$$

$$MSSS(f, \bar{x}, x) = r_{fx}^2 - \left[r_{fx} - \left(\frac{s_x}{s_f} \right) \right]^2 = \text{Correlation}^2 + \text{Cond.Bias}^2$$

$$MSSS = 1 - \frac{MSE_{init}}{MSE_{uninit}}; \text{ here ref = uninitialized hindcasts}$$

$$MSSS(f, r, x) = \frac{MSSS(f, \bar{x}, x) - MSSS(r, \bar{x}, x)}{1 - MSSS(r, \bar{x}, x)}$$

(from Murphy, Mon Wea Rev, 1988)

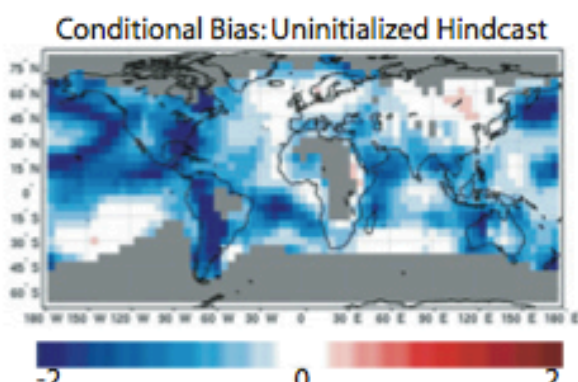
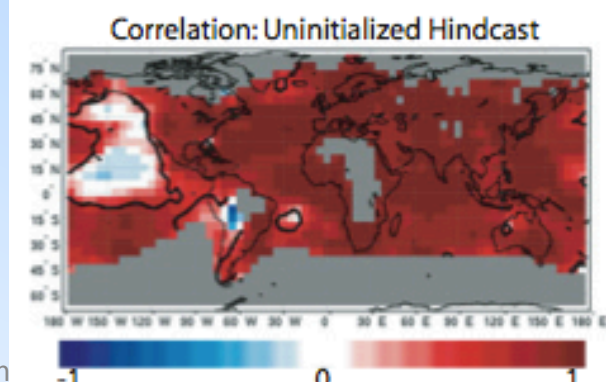
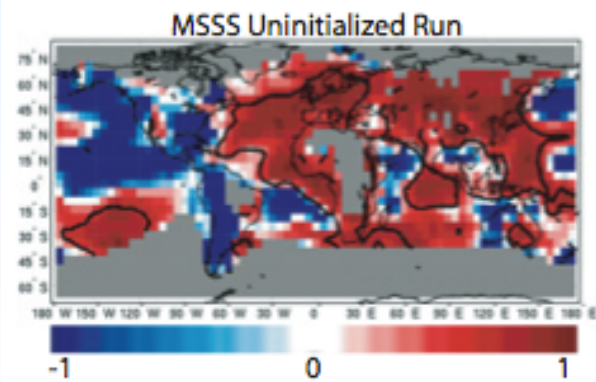
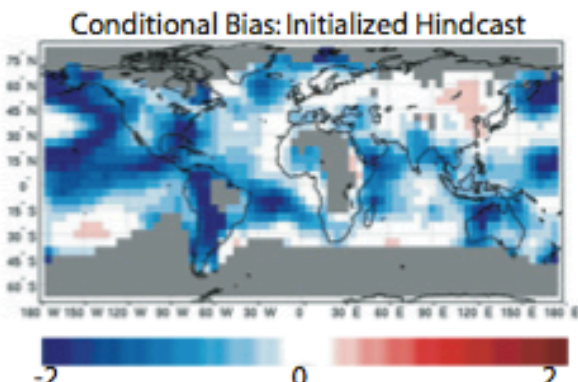
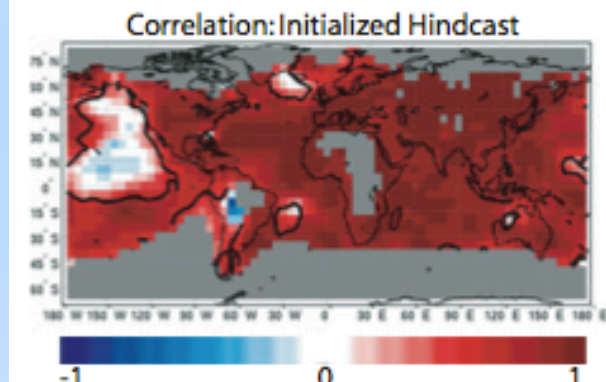
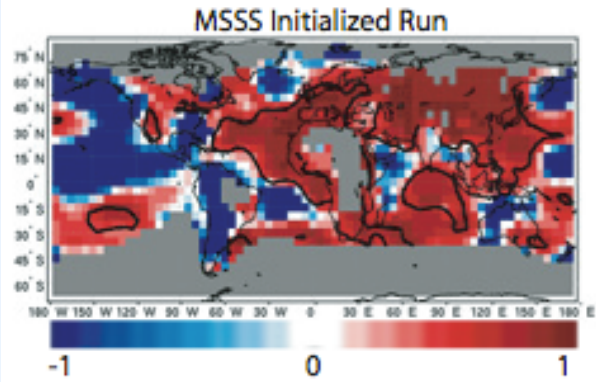
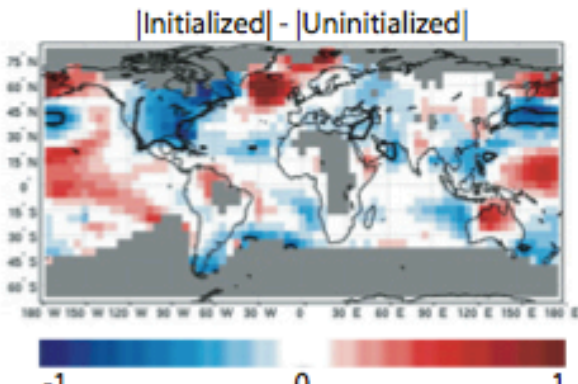
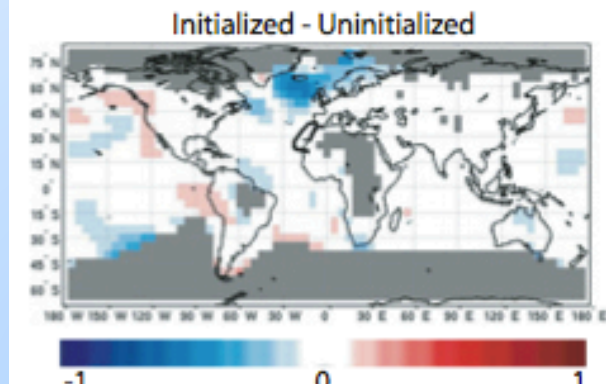
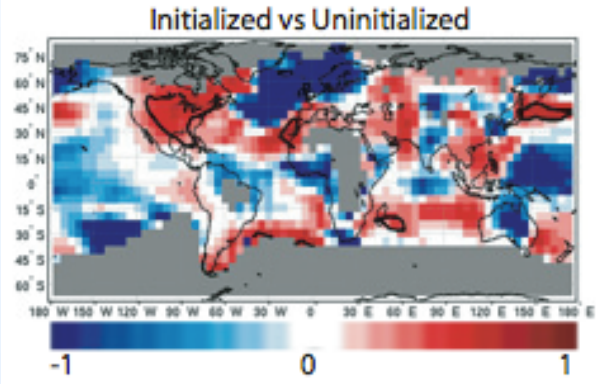
Deterministic Metrics: Mean Squared Skill Score (MSSS)

CanCM4

MSSS

Correlation

Conditional Bias



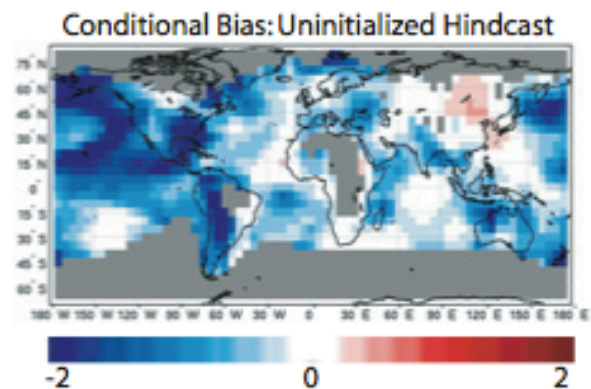
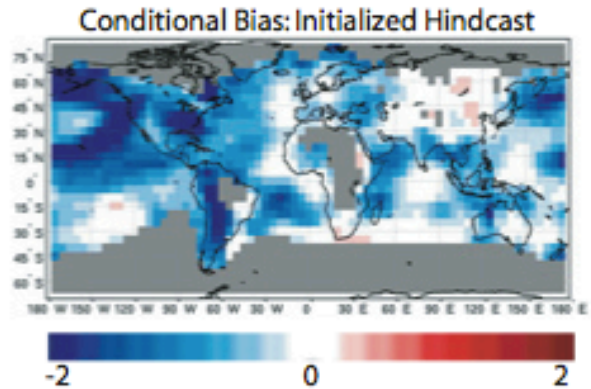
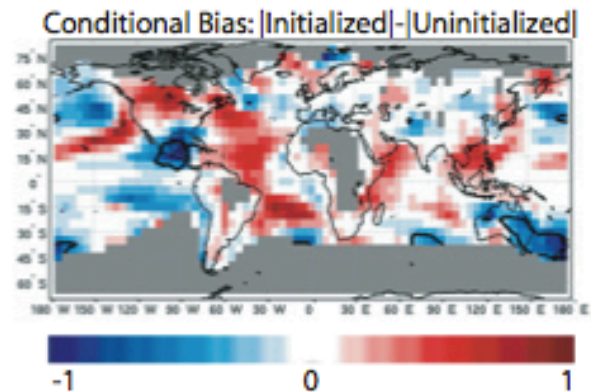
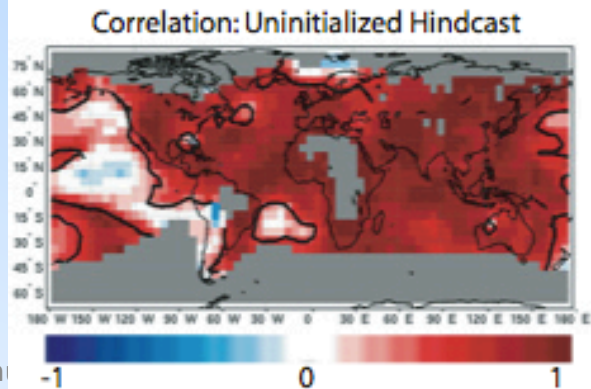
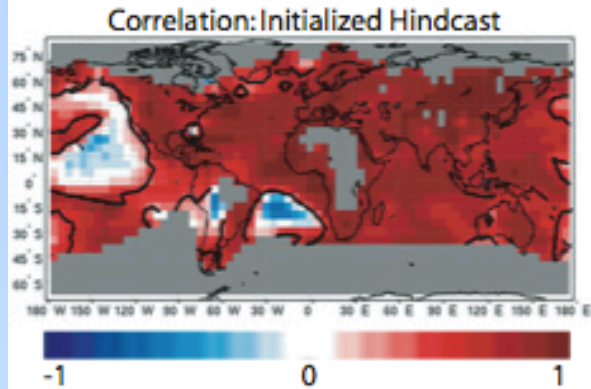
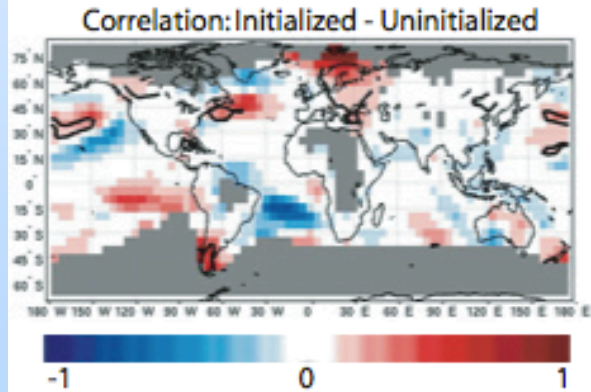
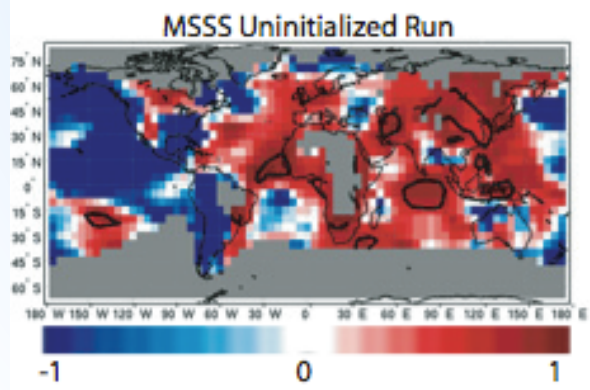
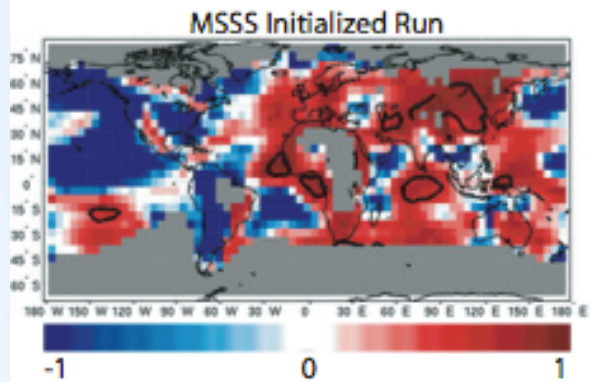
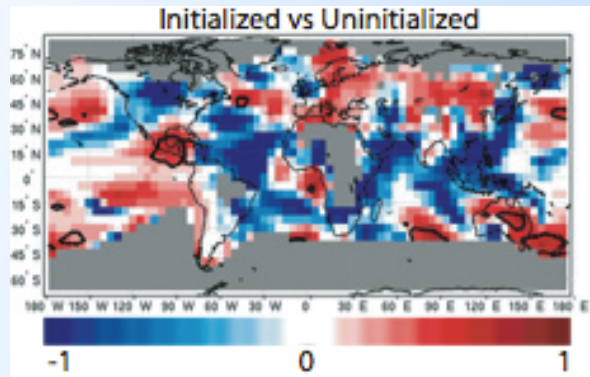
Deterministic Metrics: Mean Squared Skill Score (MSSS)

DePreSys

MSSS

Correlation

Conditional Bias



Deterministic Metrics: Mean Squared Skill Score (MSSS)

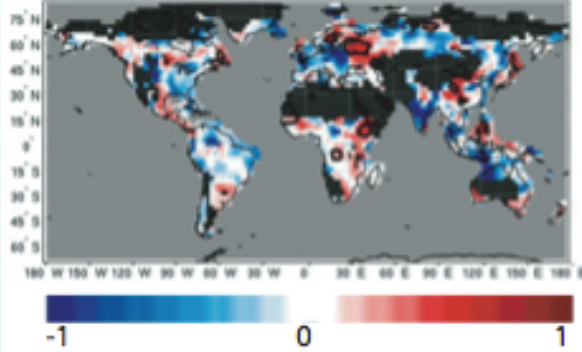
CanCM4

MSSS

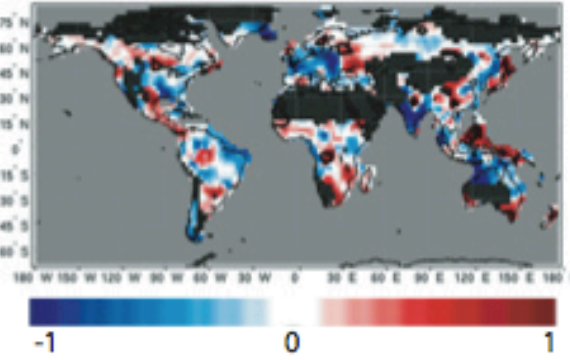
Correlation

Conditional Bias

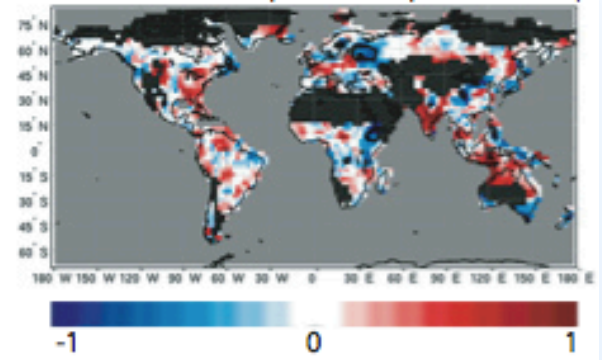
Initialized vs Uninitialized



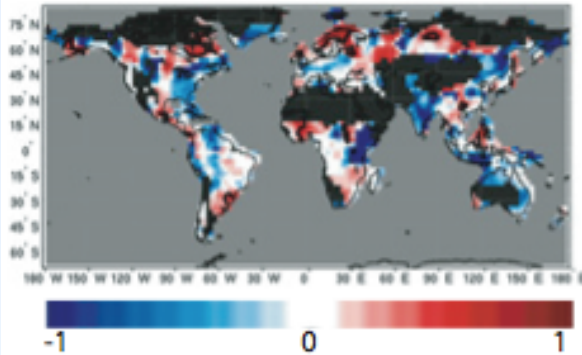
Correlation: Initialized - Uninitialized



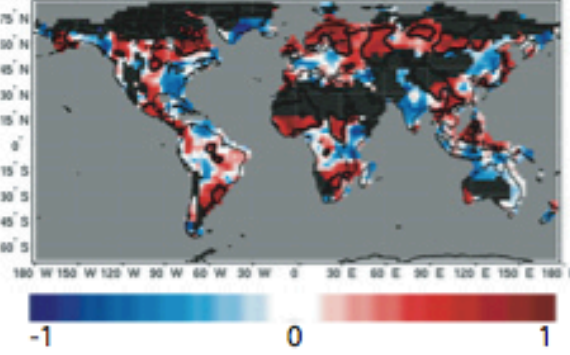
Conditional Bias: |Initialized - Uninitialized|



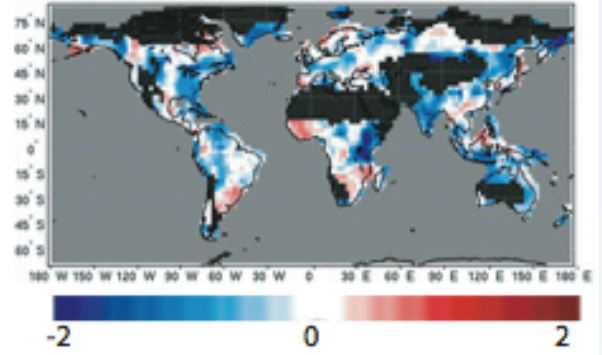
MSSS Initialized Run



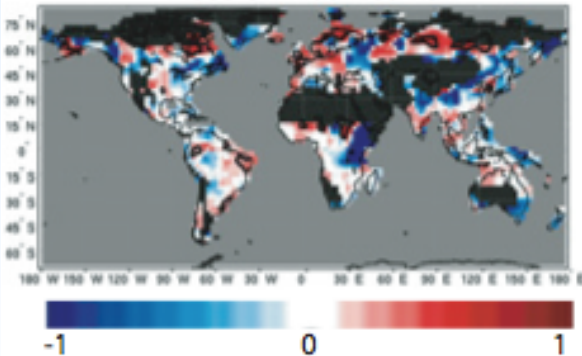
Correlation: Initialized Hindcast



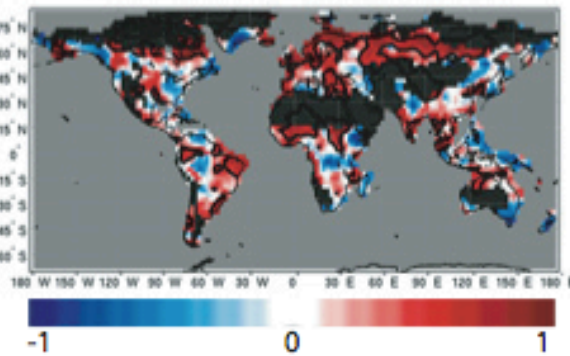
Conditional Bias: Initialized Hindcast



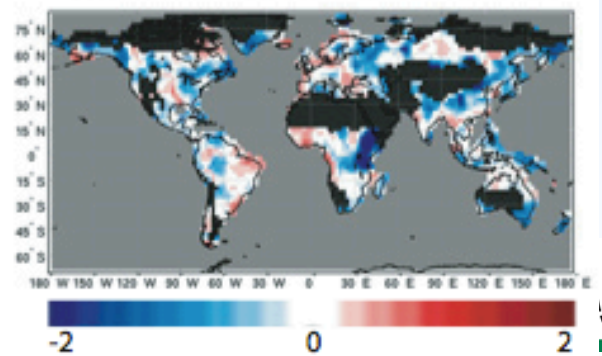
MSSS Uninitialized Run



Correlation: Uninitialized Hindcast



Conditional Bias: Uninitialized Hindcast



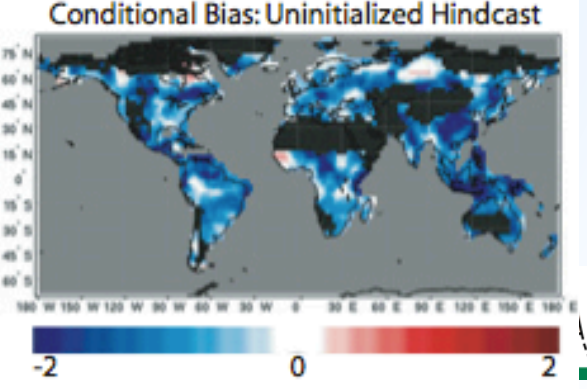
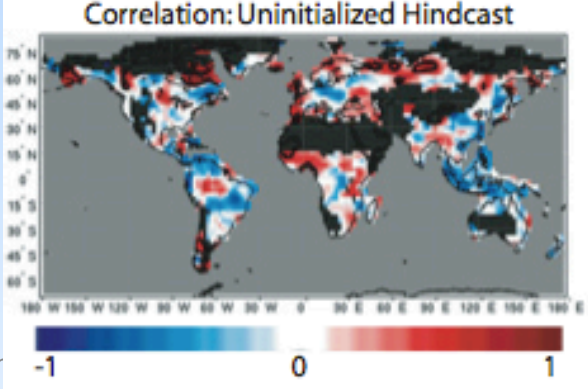
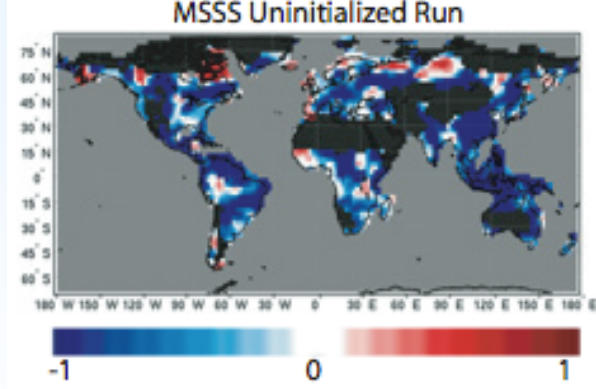
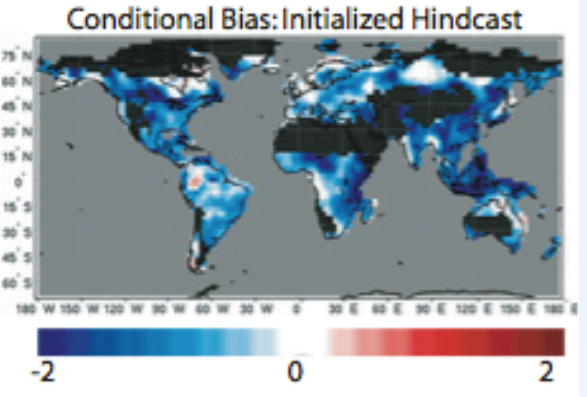
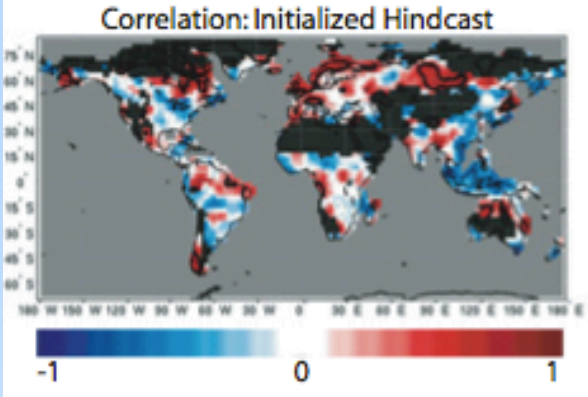
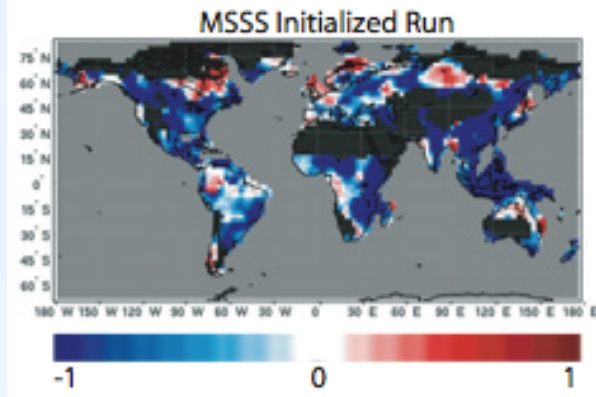
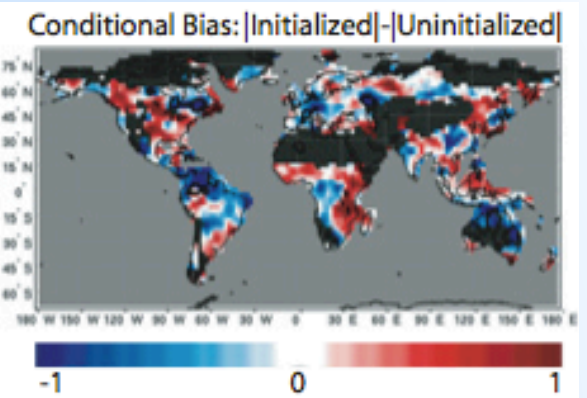
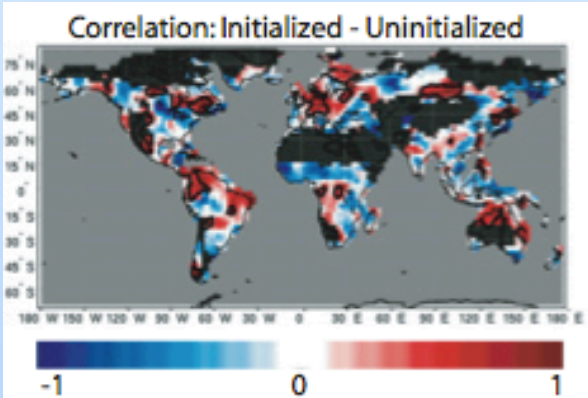
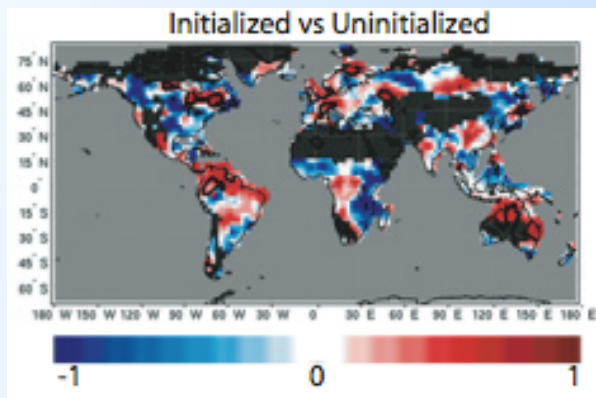
Deterministic Metrics: Mean Squared Skill Score (MSSS)

DePreSys

MSSS

Correlation

Conditional Bias



Asking Questions of the Initialized Hindcasts

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

→ Continuous Ranked Probability Skill Score (CRPSS)

$$\text{CRPSS} = 1 - (\text{CRPS}_{\text{fcst}} / \text{CRPS}_{\text{ref}})$$

Q2: **Fcst** uncertainty = average ensemble spread

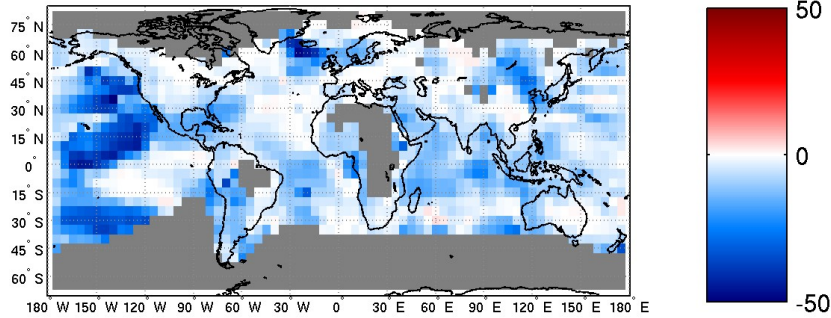
Ref uncertainty = standard error of ensemble mean

$$F = \hat{Y} + \bar{\sigma}_{\text{Ens}}; \quad R = \hat{Y} + \sigma_{\text{StdErr}}$$

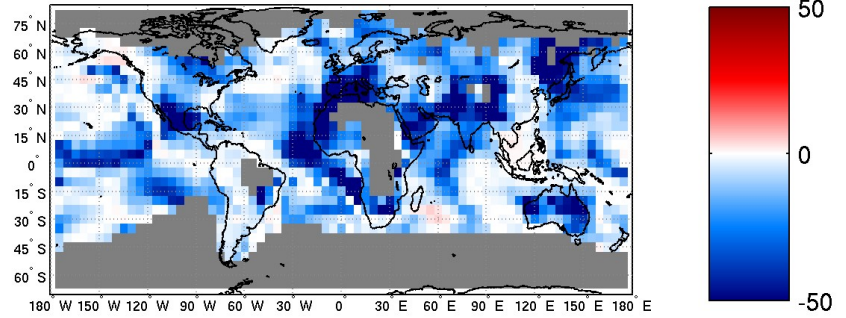
Probabilistic Metrics: CRPSS

$$F = \hat{Y} + \bar{\sigma}_{Ens}; \quad R = \hat{Y} + \sigma_{StdErr}$$

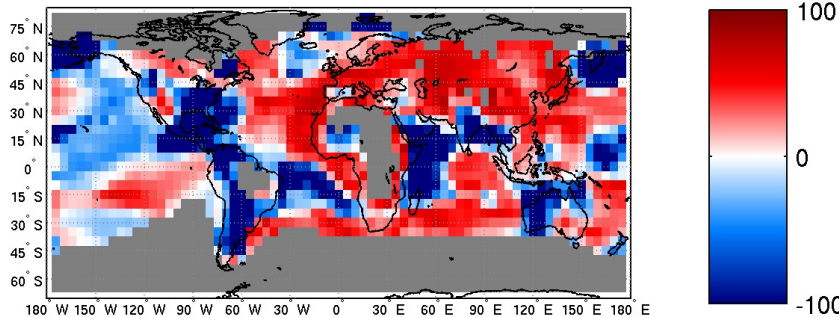
CanCM4 CRPSS (%): Year 2-9 (Obs=HadCRUT3v smooth temp)
Avg Ens Spread vs Standard Error



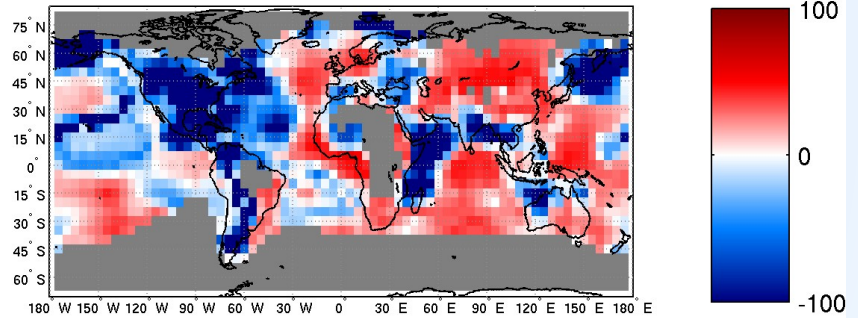
DePreSys CRPSS (%): Year 2-9 (Obs=HadCRUT3v smooth temp)
Avg Ens Spread vs Standard Error



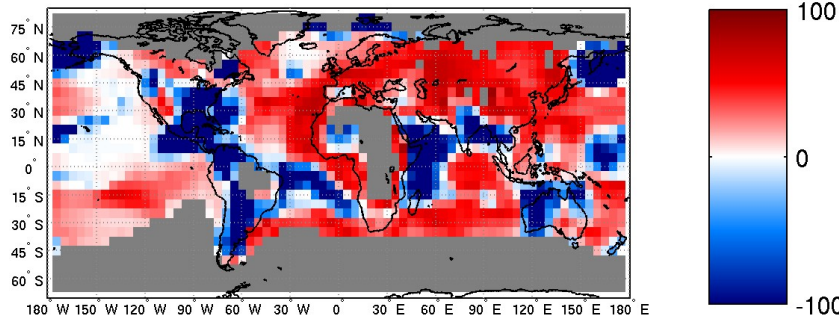
CRPSS (%):Time-Avg Ens Spread vs Climo



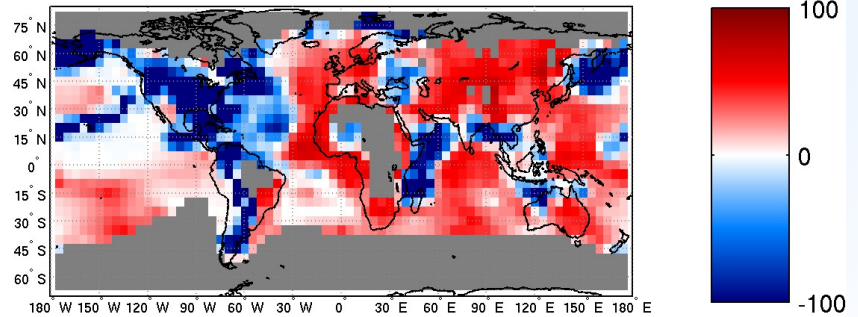
CRPSS (%):Time-Avg Ens Spread vs Climo



CRPSS (%):Standard Error of Ens Mean vs Climo



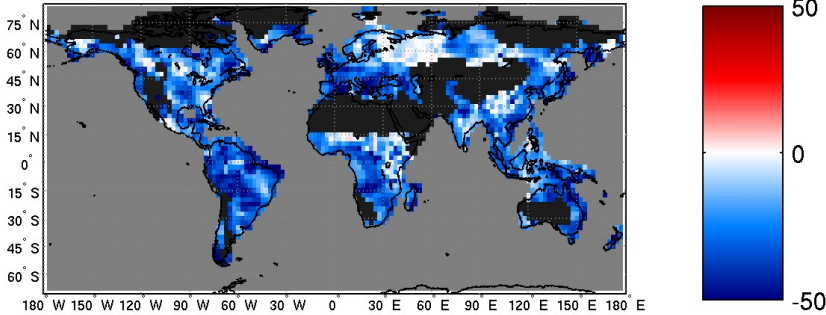
CRPSS (%):Standard Error of Ens Mean vs Climo



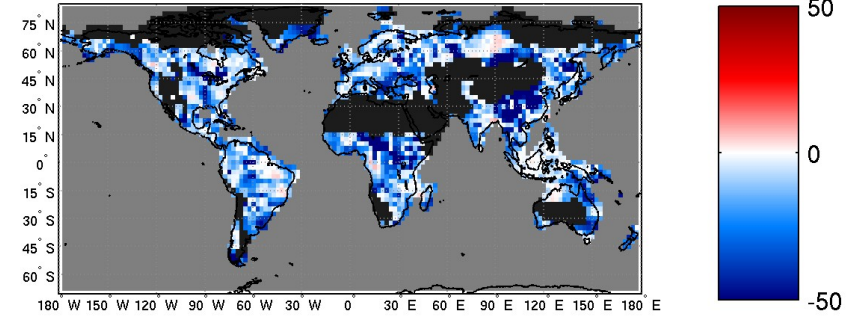
Probabilistic Metrics: CRPSS

$$F = \hat{Y} + \bar{\sigma}_{Ens}; \quad R = \hat{Y} + \sigma_{StdErr}$$

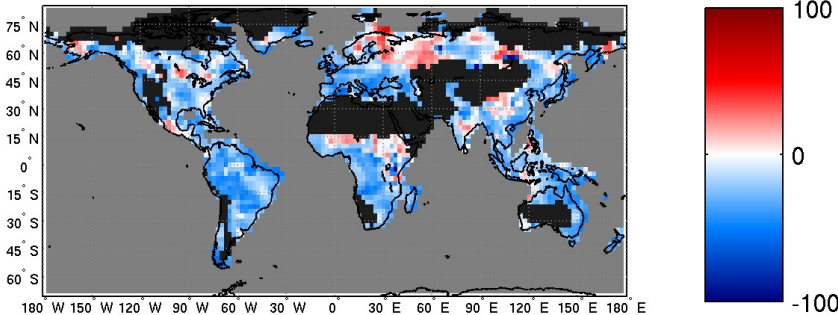
CanCM4 CRPSS (%): Year 2-9 (Obs=GPCC smooth precip)
Avg Ens Spread vs Standard Error



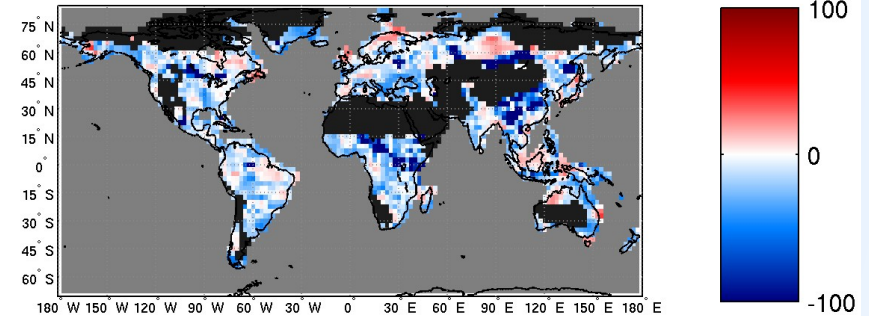
DePreSys CRPSS (%): Year 2-9 (Obs=GPCC smooth precip)
Avg Ens Spread vs Standard Error



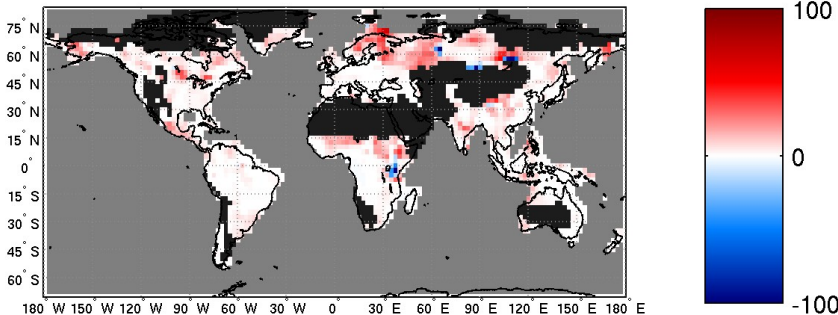
CRPSS (%): Time-Avg Ens Spread vs Climo



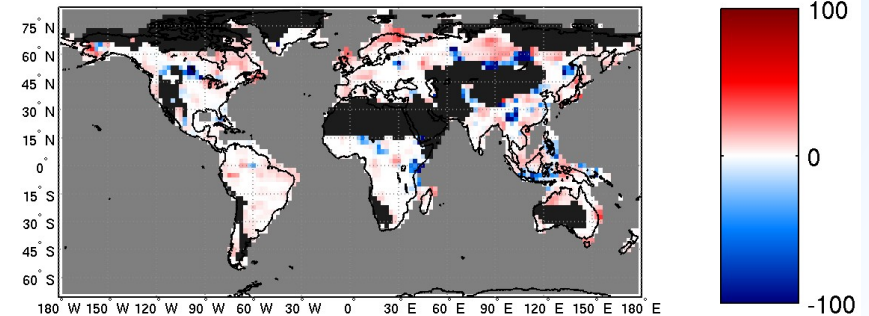
CRPSS (%): Time-Avg Ens Spread vs Climo



CRPSS (%): Standard Error of Ens Mean vs Climo



CRPSS (%): Standard Error of Ens Mean vs Climo



Statistical Significance: Non-parametric bootstrap

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

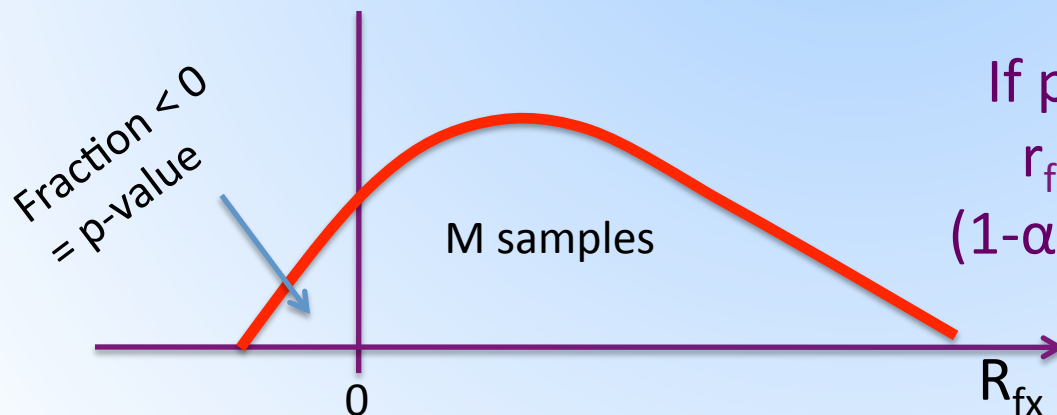
Start out with nominally $n=10$ start times.

Draw random start times as pairs up to n values.

i.e. 1st draw: $i=1 \rightarrow$ e.g. $I(i,k)=5$ (1980), so $i=2 \rightarrow I(i+1,k)=6$ (1985), etc.
up to $i=10$

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

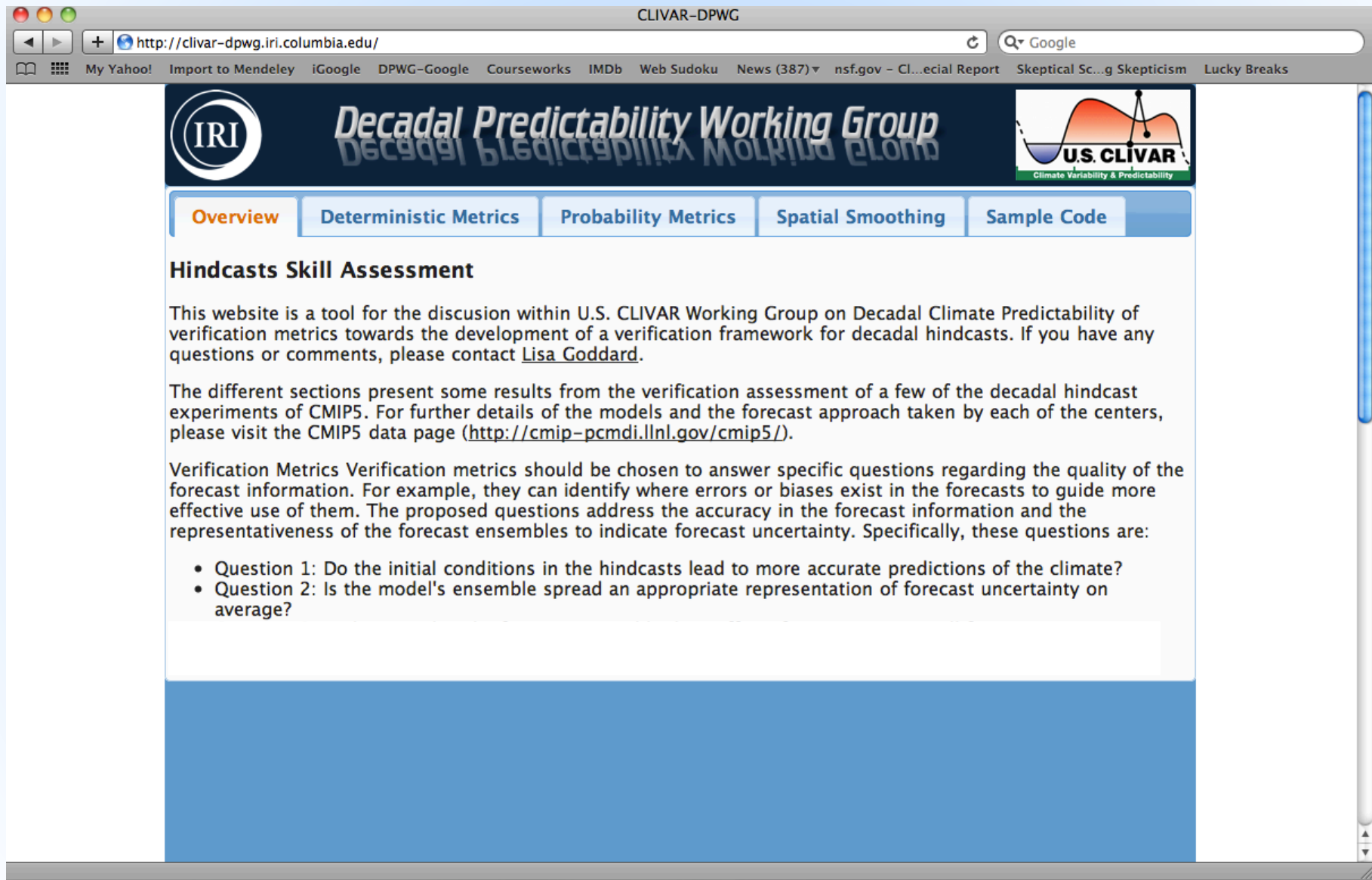
$$\tilde{f}_i^E(k) = f_{I(i,k)}^{E(I)}$$



If p-value $\leq \alpha$, then
 r_{fx} is significant at
 $(1-\alpha) \times 100\%$ confidence

Proto-type Website: *Work in progress*



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



CLIVAR-DPWG

<http://clivar-dpwg.iri.columbia.edu/> Google

My Yahoo! Import to Mendeley iGoogle DPWG-Google Courseworks IMDb Web Sudoku News (387) nsf.gov - Cl...ecial Report Skeptical Sc...g Skepticism Lucky Breaks

 **Decadal Predictability Working Group** 

Overview Deterministic Metrics Probability Metrics Spatial Smoothing Sample Code

Hindcasts Skill Assessment

This website is a tool for the discussion within U.S. CLIVAR Working Group on Decadal Climate Predictability of verification metrics towards the development of a verification framework for decadal hindcasts. If you have any questions or comments, please contact [Lisa Goddard](#).

The different sections present some results from the verification assessment of a few of the decadal hindcast experiments of CMIP5. For further details of the models and the forecast approach taken by each of the centers, please visit the CMIP5 data page (<http://cmip-pcmdi.llnl.gov/cmip5/>).

Verification Metrics Verification metrics should be chosen to answer specific questions regarding the quality of the forecast information. For example, they can identify where errors or biases exist in the forecasts to guide more effective use of them. The proposed questions address the accuracy in the forecast information and the representativeness of the forecast ensembles to indicate forecast uncertainty. Specifically, these questions are:

- 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?

Outline

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

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 addresses specific questions of the hindcast quality and offers 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.

Paper to be submitted to Climate Dynamics.

goddard@iri.columbia.edu