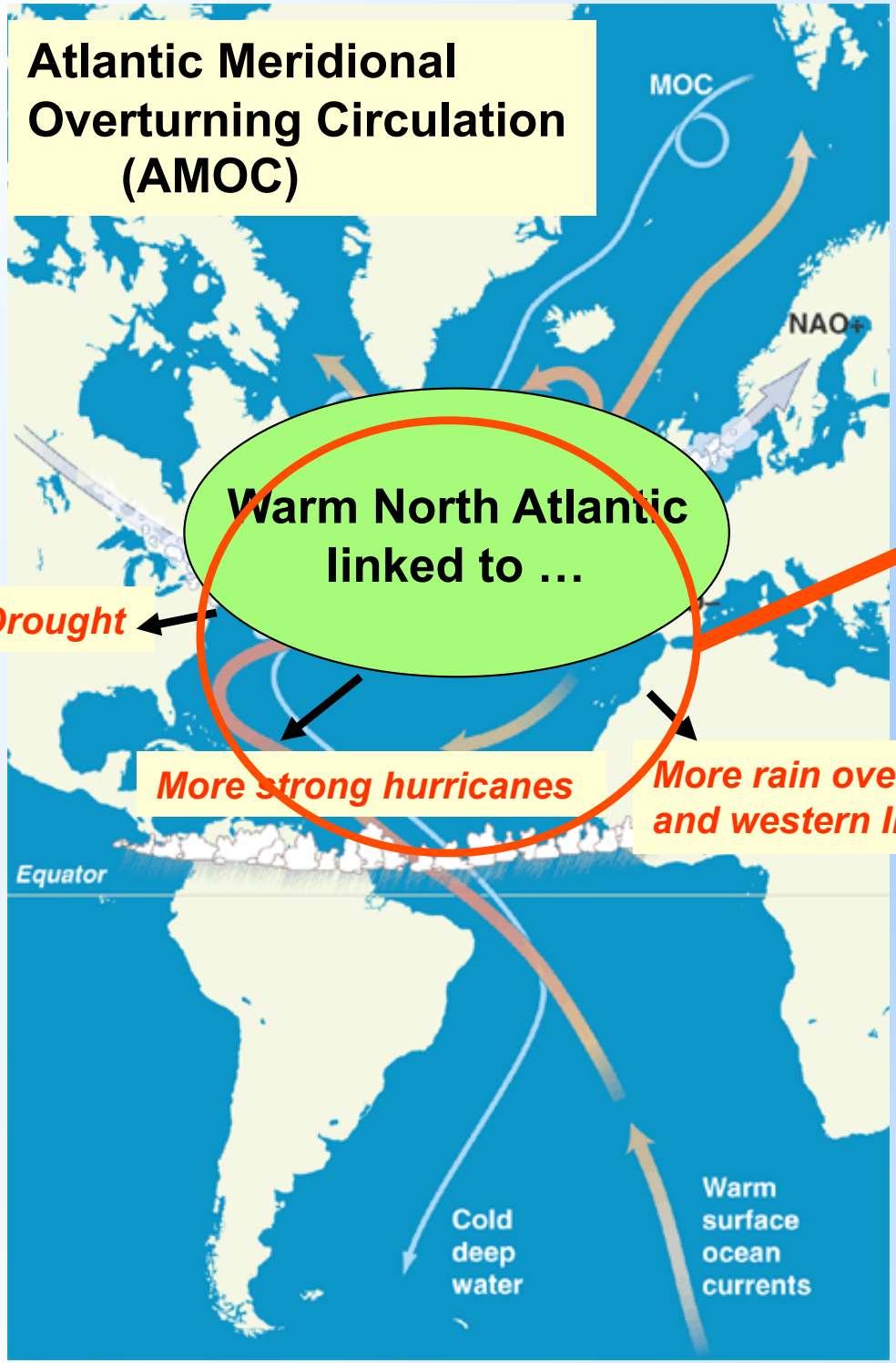
A Metrics Framework for Interannual-to-Decadal Predictions Experiments

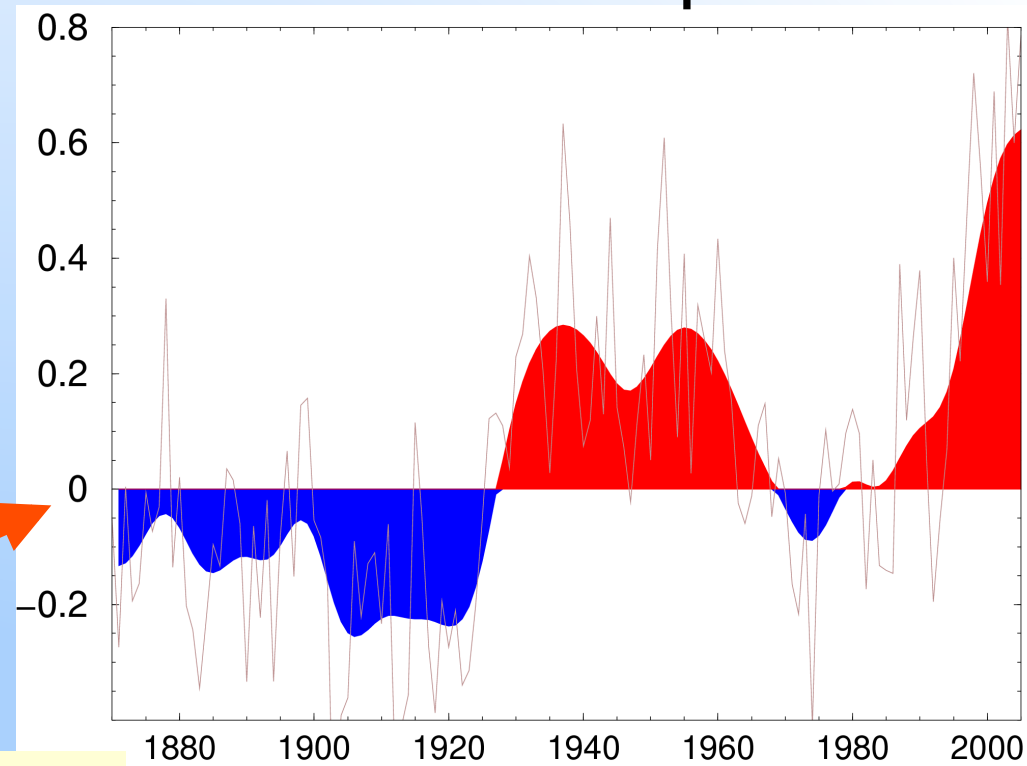
L. Goddard*, on behalf of the
US CLIVAR Decadal Predictability Working Group & Collaborators:
Lisa Goddard, Arun Kumar, Amy Solomon, James Carton, Clara Deser, Ichiro Fukumori, Arthur M. Greene, Gabriele Hegerl, Ben Kirtman, Yochanan Kushnir, Matthew Newman, Doug Smith, Dan Vimont, Tom Delworth, Jerry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

* email: goddard@iri.columbia.edu

Atlantic Meridional Overturning Circulation (AMOC)



North Atlantic Temperature

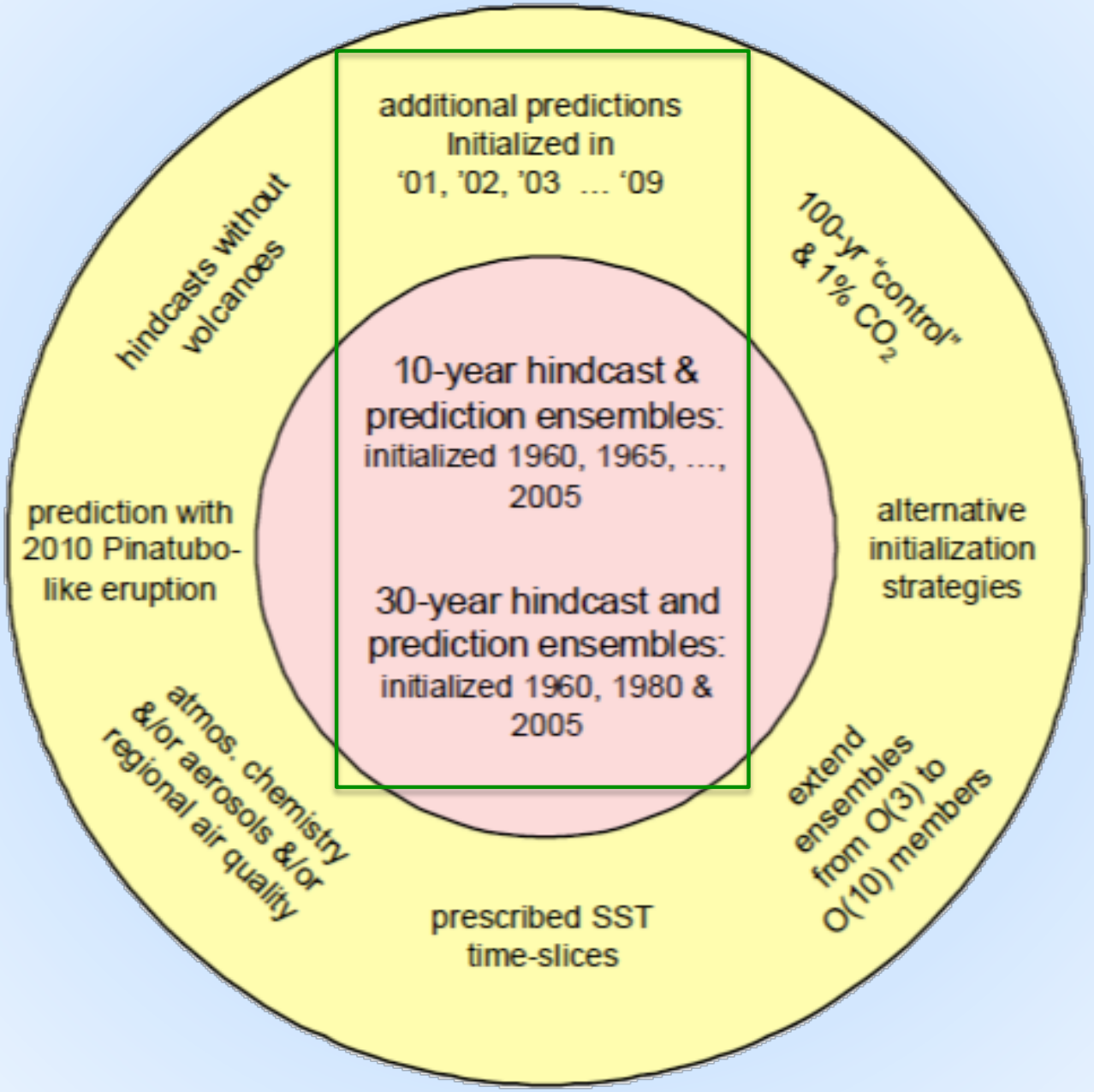


Two important aspects:

a. **Decadal-multidecadal fluctuations**
b. **Long-term trend**

(Courtesy: Joe Tribbia, NCAR)

CMIP5 Experimental Prediction Design



http://www.clivar.org/organization/wgcm/references/Taylor_CMIP5.pdf

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.

US CLIVAR Working Group on Decadal Predictability:

Proposed FRAMEWORK for Verification:

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

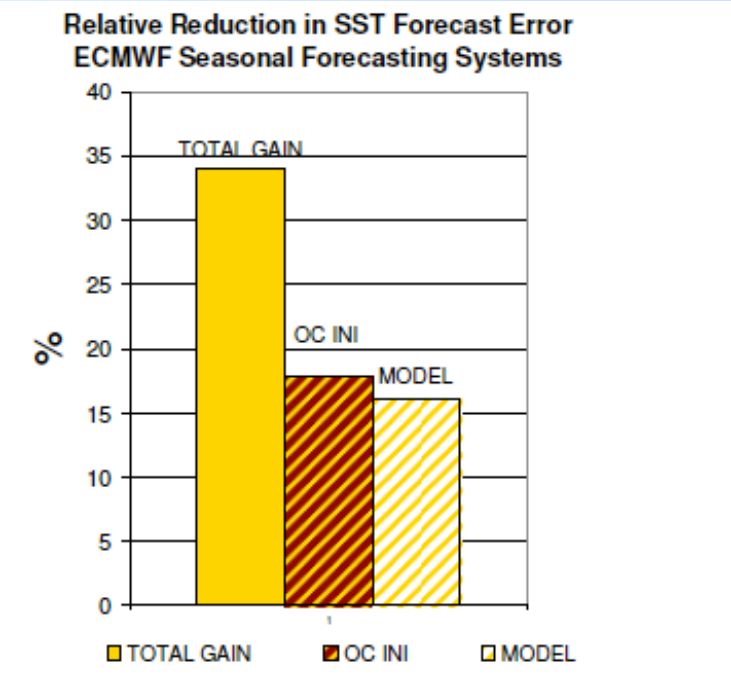
2. **Prediction skill** – Quality of system; quality of information
   - Where? What space & time scales?
   - Actual anomalies & 'decadal scale trends'
   - Conditional skill?
   - Values of ICs: higher correlations, lower RMSEs

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

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

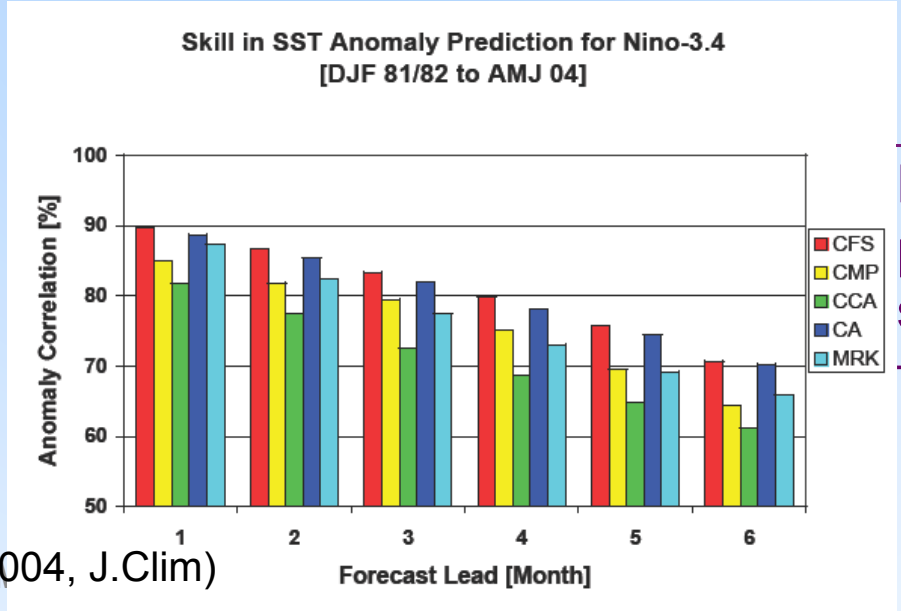
Motivation: Forecasts need verification

… 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

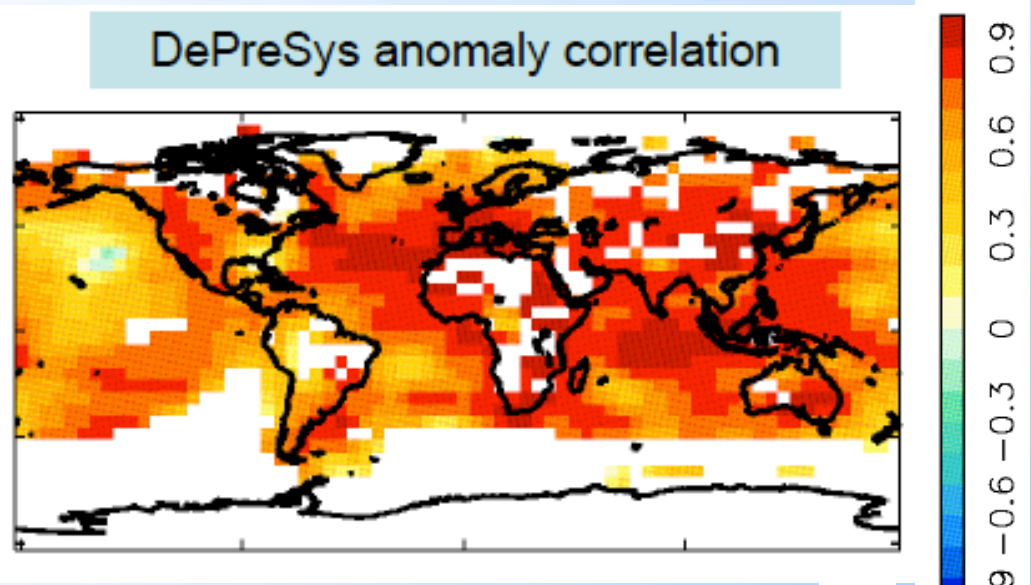


(Saha et al. 2004, J.Clim)

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

How "good" are they?: Deterministic Metrics

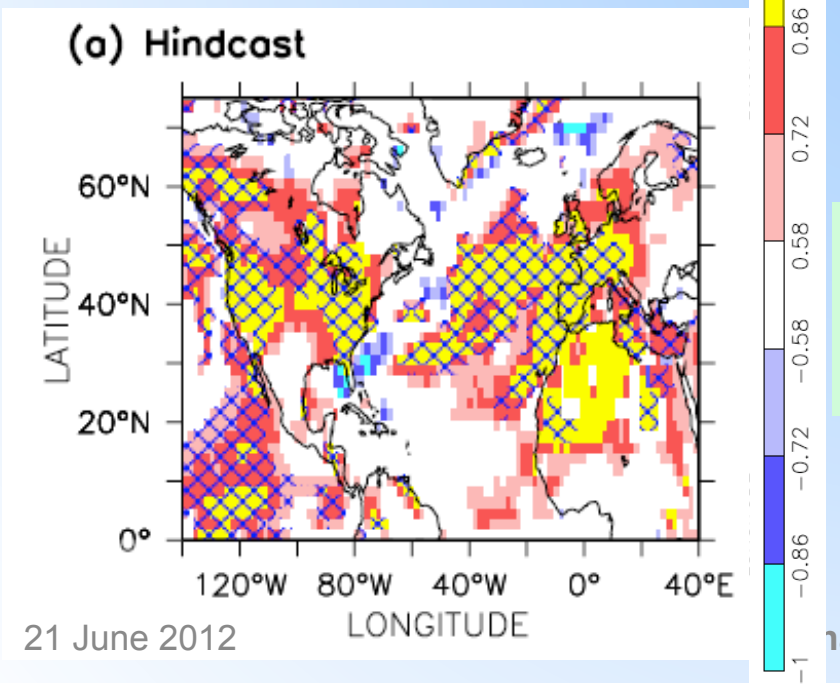
Regional Average (15°x15°); 5-Year Means:



- 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)

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 6-9, 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

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

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

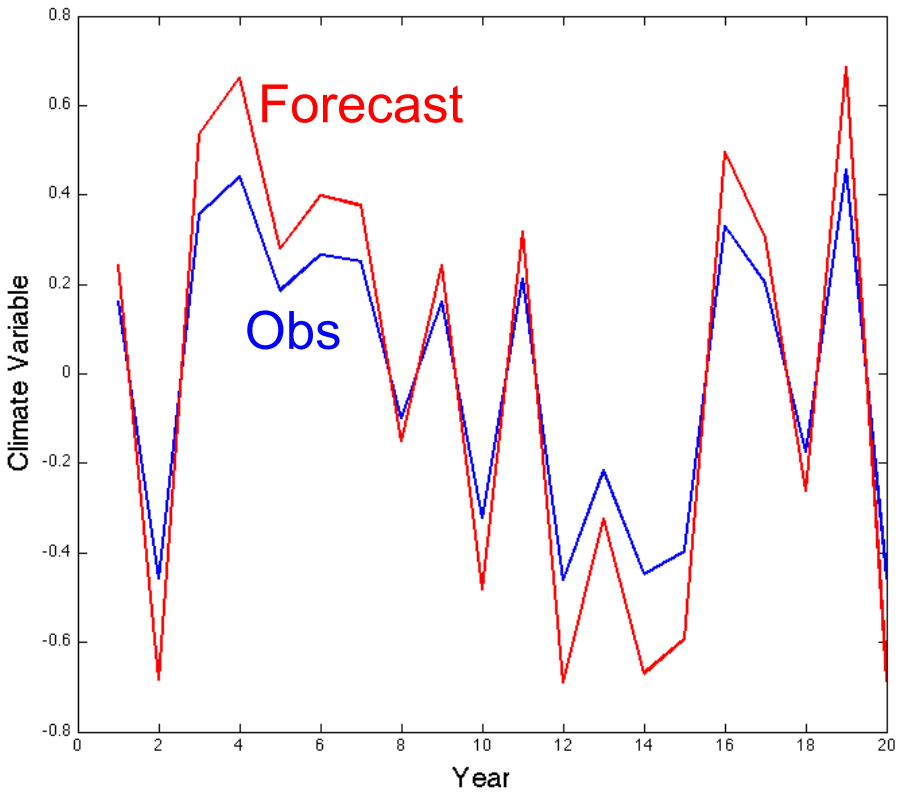
$$MSSS = 1 - \frac{MSE_{init}}{MSE_{uninit}}; \quad \text{here ref} = \text{uninitialized hindcasts}$$

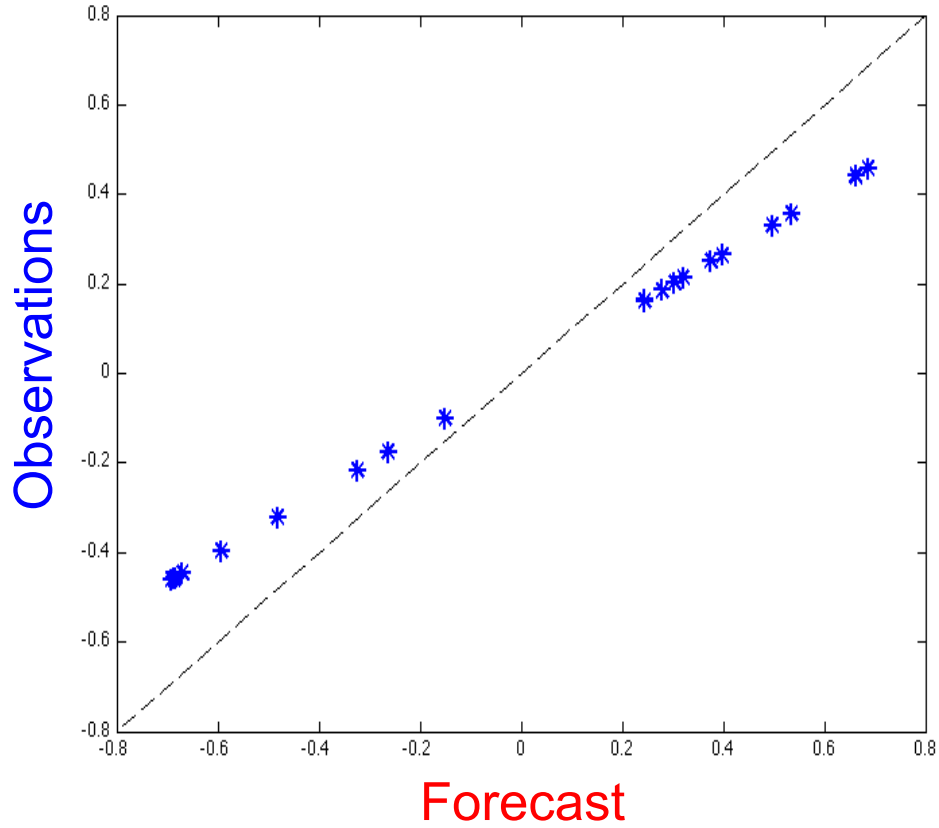
$$MSSS(f,r,x) = \frac{MSSS(f,\bar{x},x) - MSSS(r,\bar{x},x)}{1 - MSSS(r,\bar{x},x)}$$

(from Murphy, Mon Wea Rev, 1988)

Elaboration on "Conditional Bias"

$$\text{Conditional Bias} = [r_{fx} - (s_f / s_x)]$$





Perfect Correlation, but Conditional Bias because $s_f > s_x$

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

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

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

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

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

(from Murphy, Mon Wea Rev, 1988)

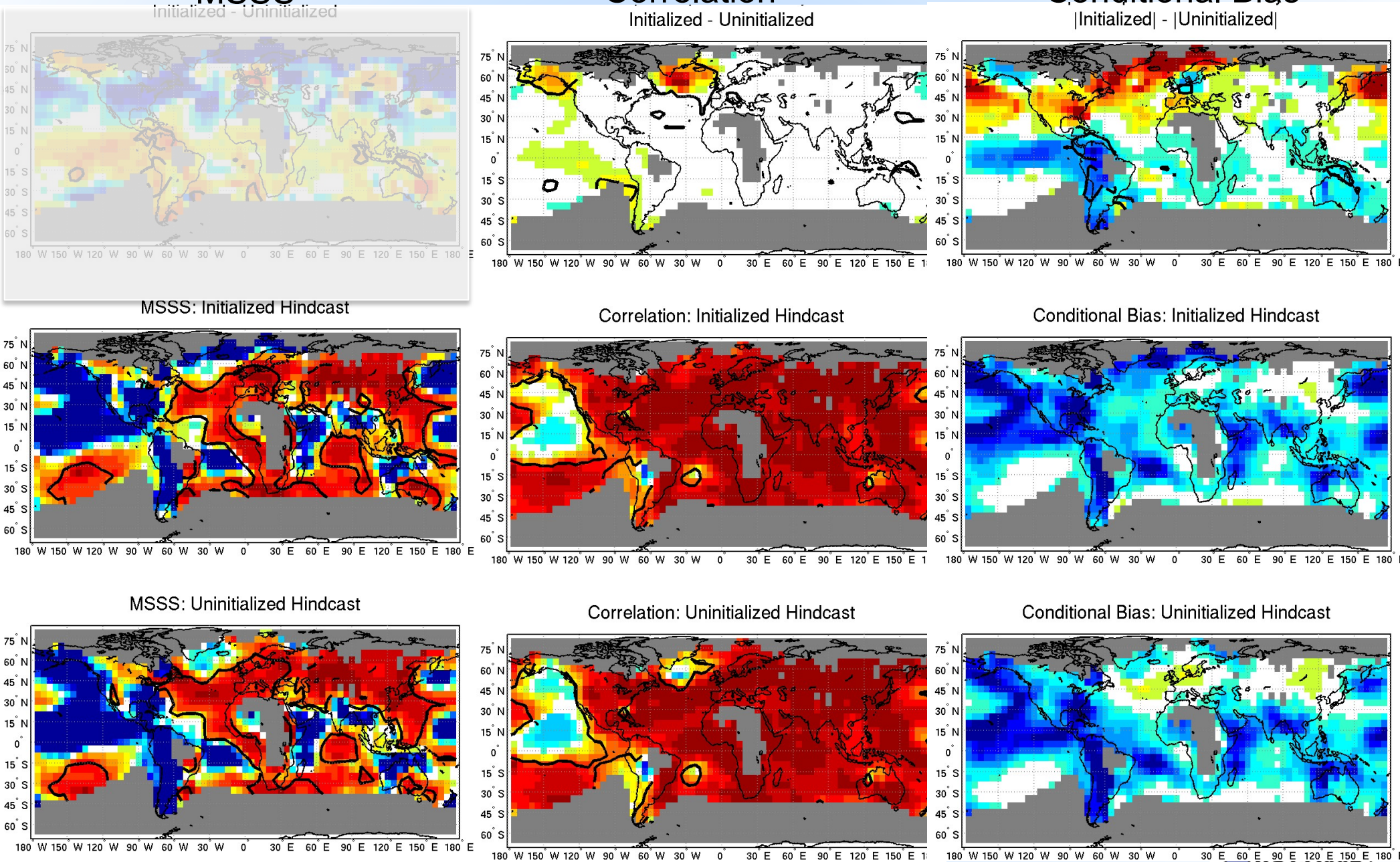
Deterministic Metrics: Mean Squared Skill Score (MSSS)

GFDL2.1

MSSS

Correlation

Conditional Bias

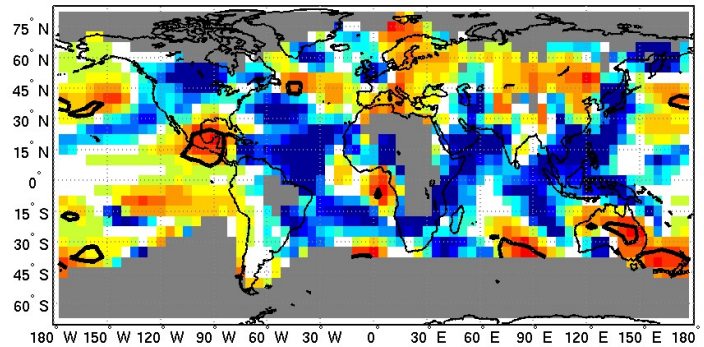


Deterministic Metrics: Mean Squared Skill Score (MSSS)

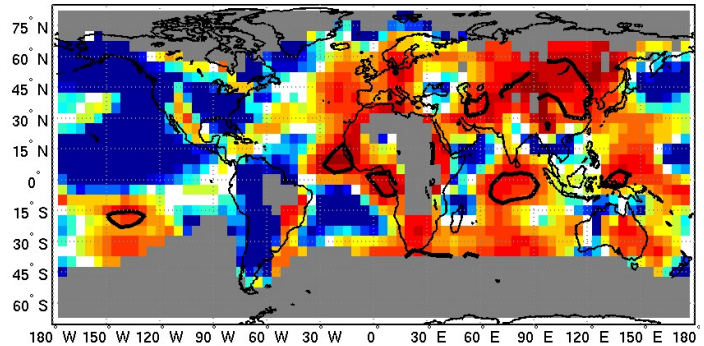
DePreSys

MSSS

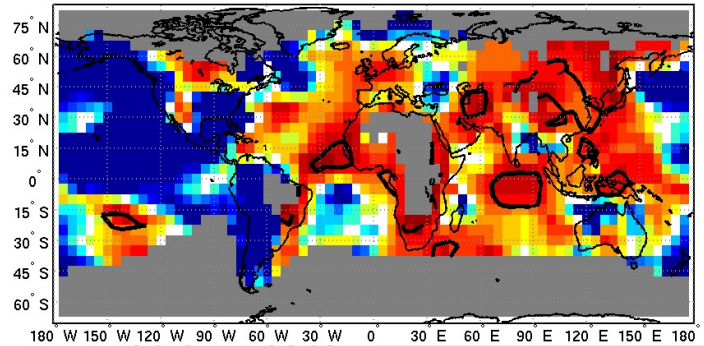
Initialized - Uninitialized



MSSS: Initialized Hindcast

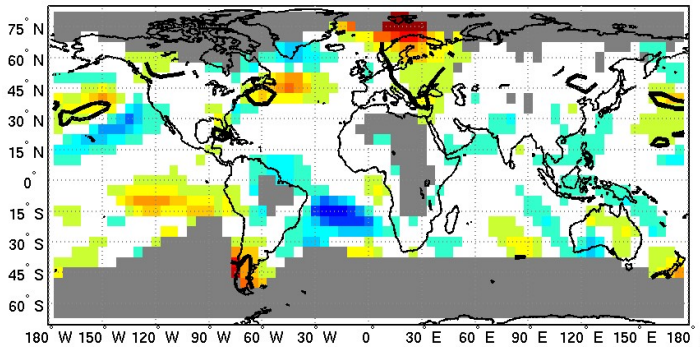


MSSS: Uninitialized Hindcast

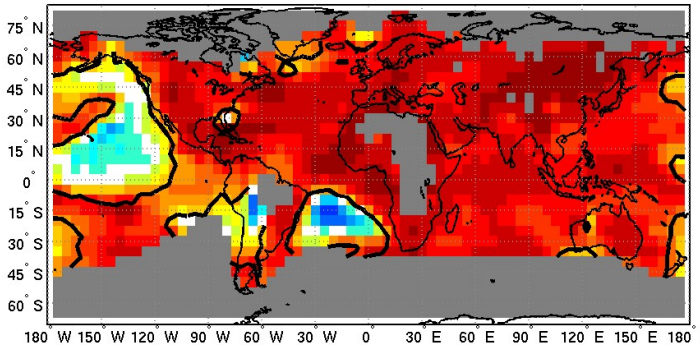


Correlation

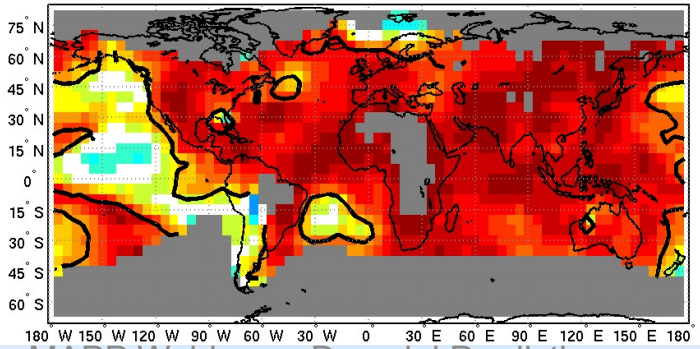
Initialized - Uninitialized



Correlation: Initialized Hindcast

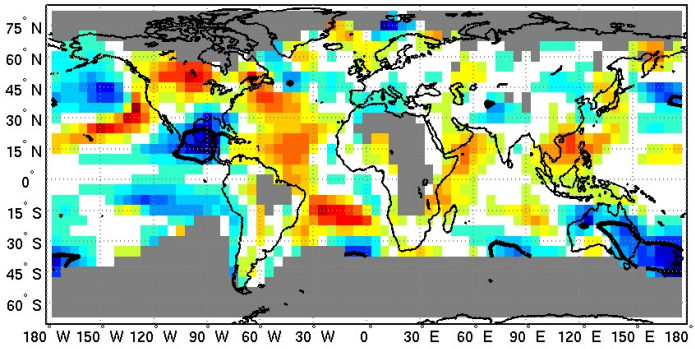


Correlation: Uninitialized Hindcast

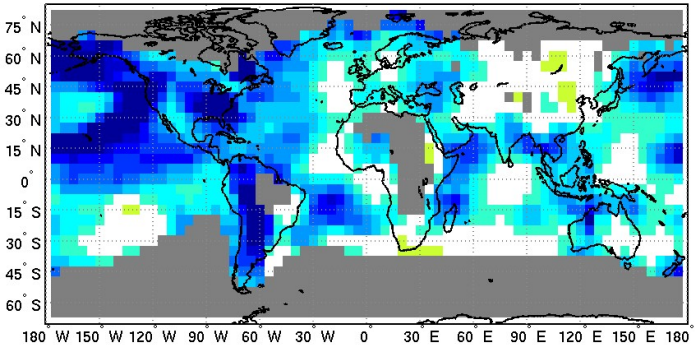


Conditional Bias

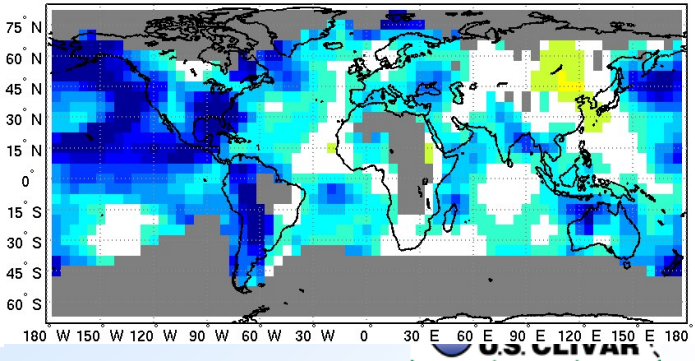
|Initialized| - |Uninitialized|



Conditional Bias: Initialized Hindcast



Conditional Bias: Uninitialized Hindcast

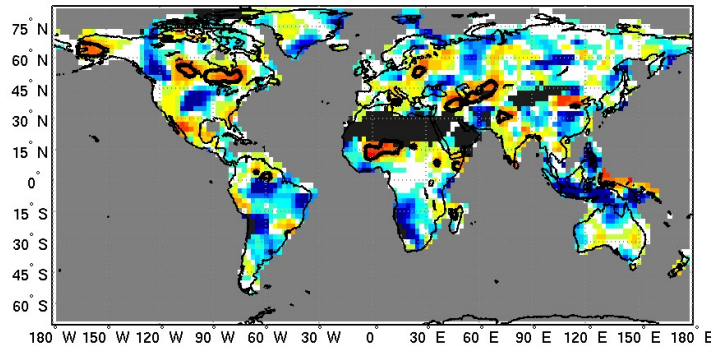


Deterministic Metrics: Mean Squared Skill Score (MSSS)

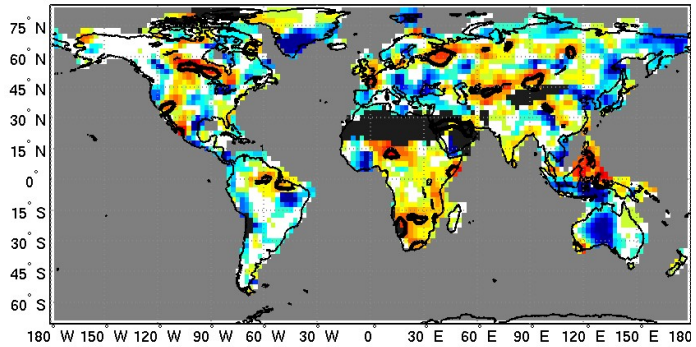
GFDL2.1

Year 2-5 | Year 6-9 | Year 2-9

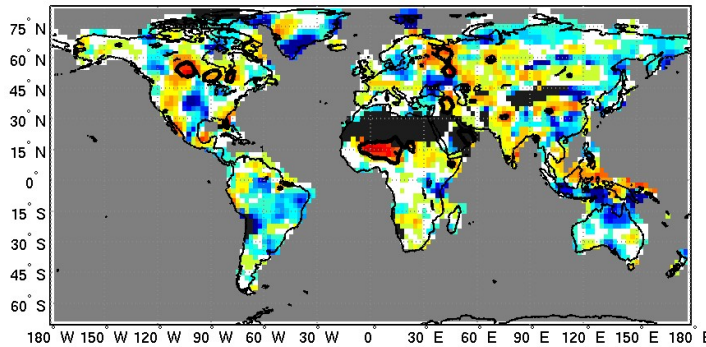
Initialized - Uninitialized



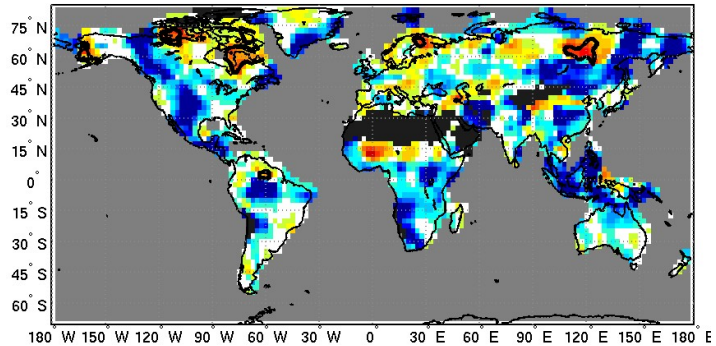
Initialized - Uninitialized



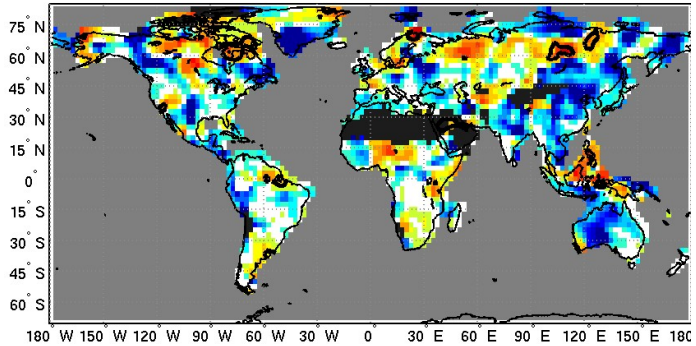
Initialized - Uninitialized



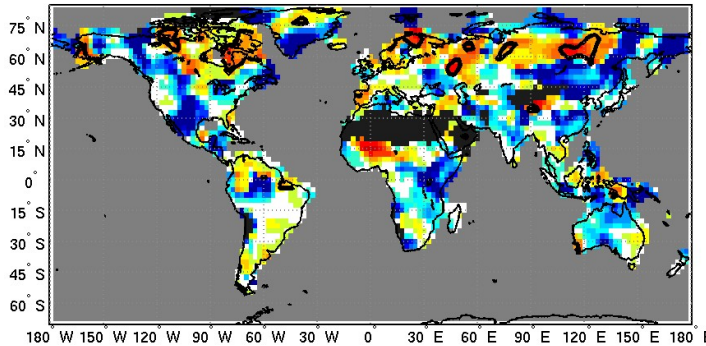
MSSS: Initialized Hindcast



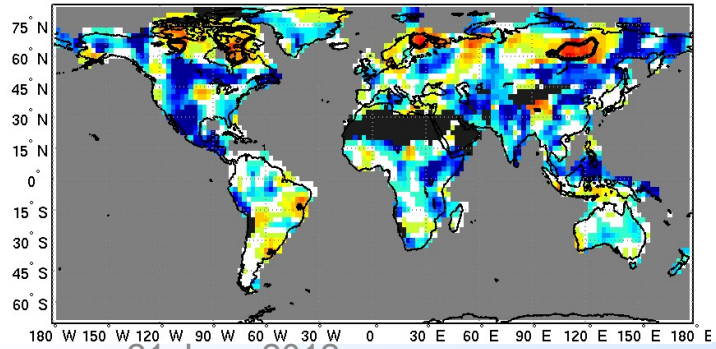
MSSS: Initialized Hindcast



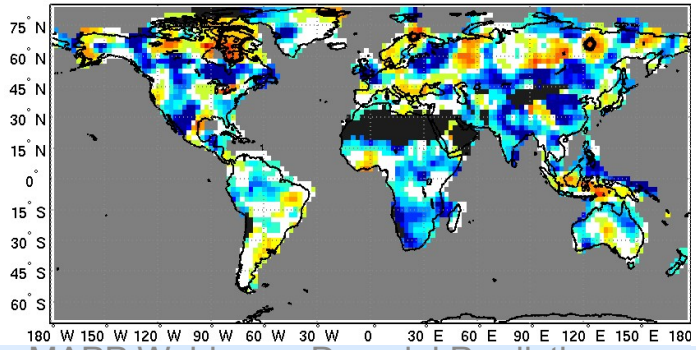
MSSS: Initialized Hindcast



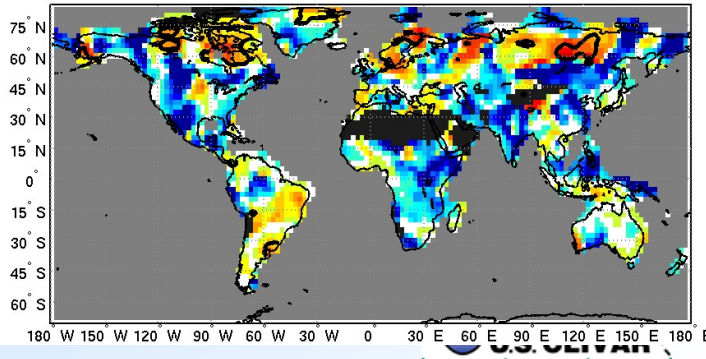
MSSS: Uninitialized Hindcast



MSSS: Uninitialized Hindcast



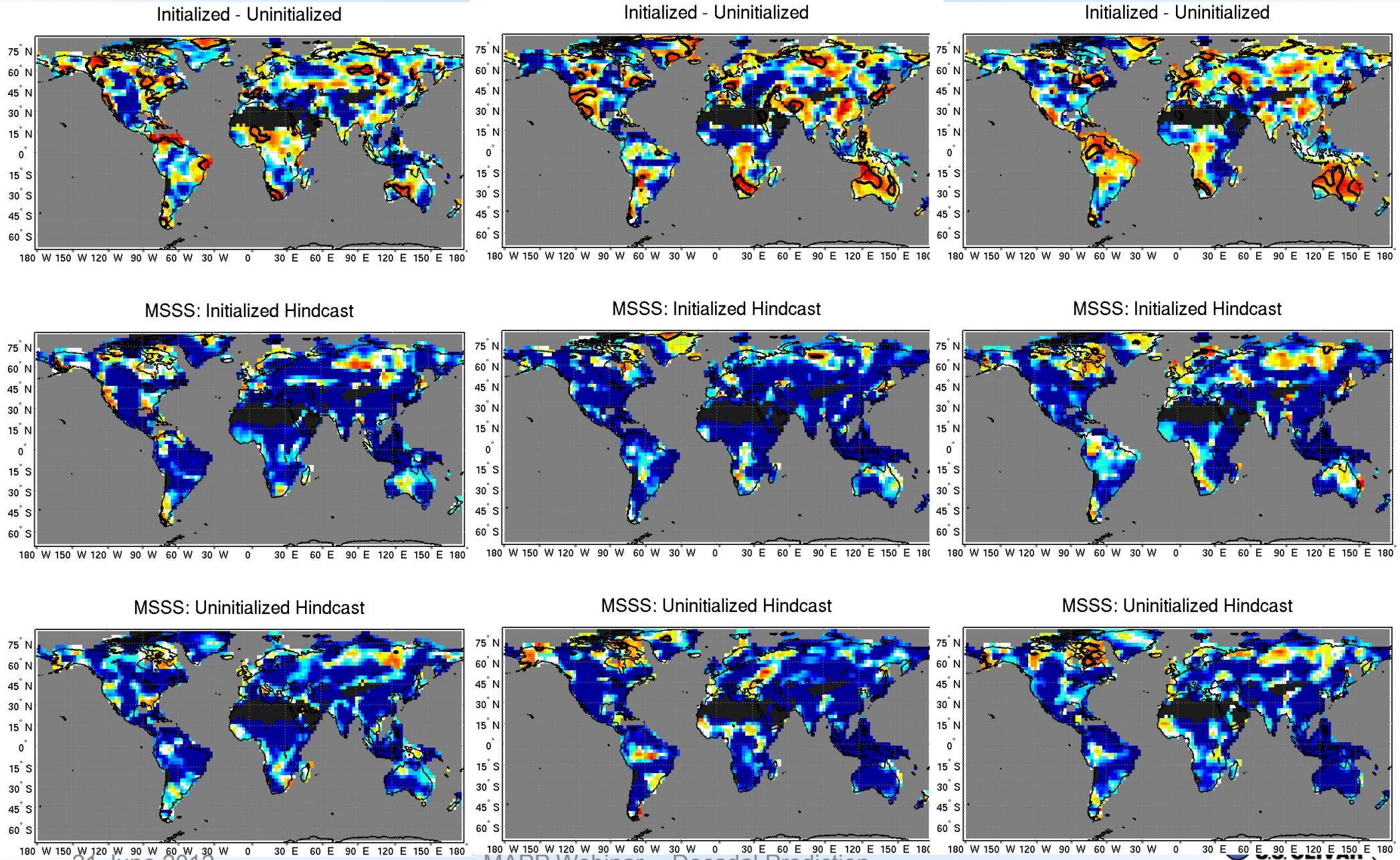
MSSS: Uninitialized Hindcast



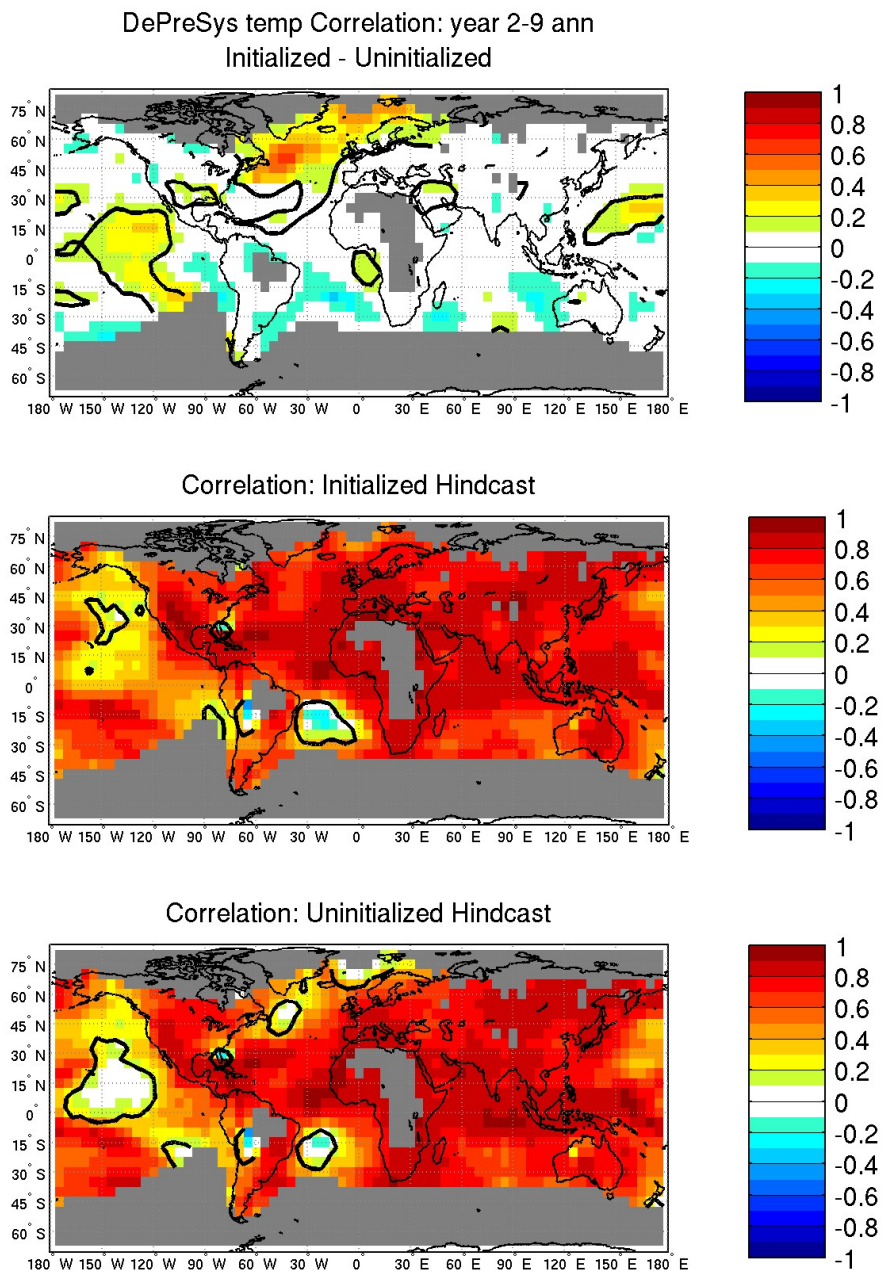
Deterministic Metrics: Mean Squared Skill Score (MSSS)

DePreSys

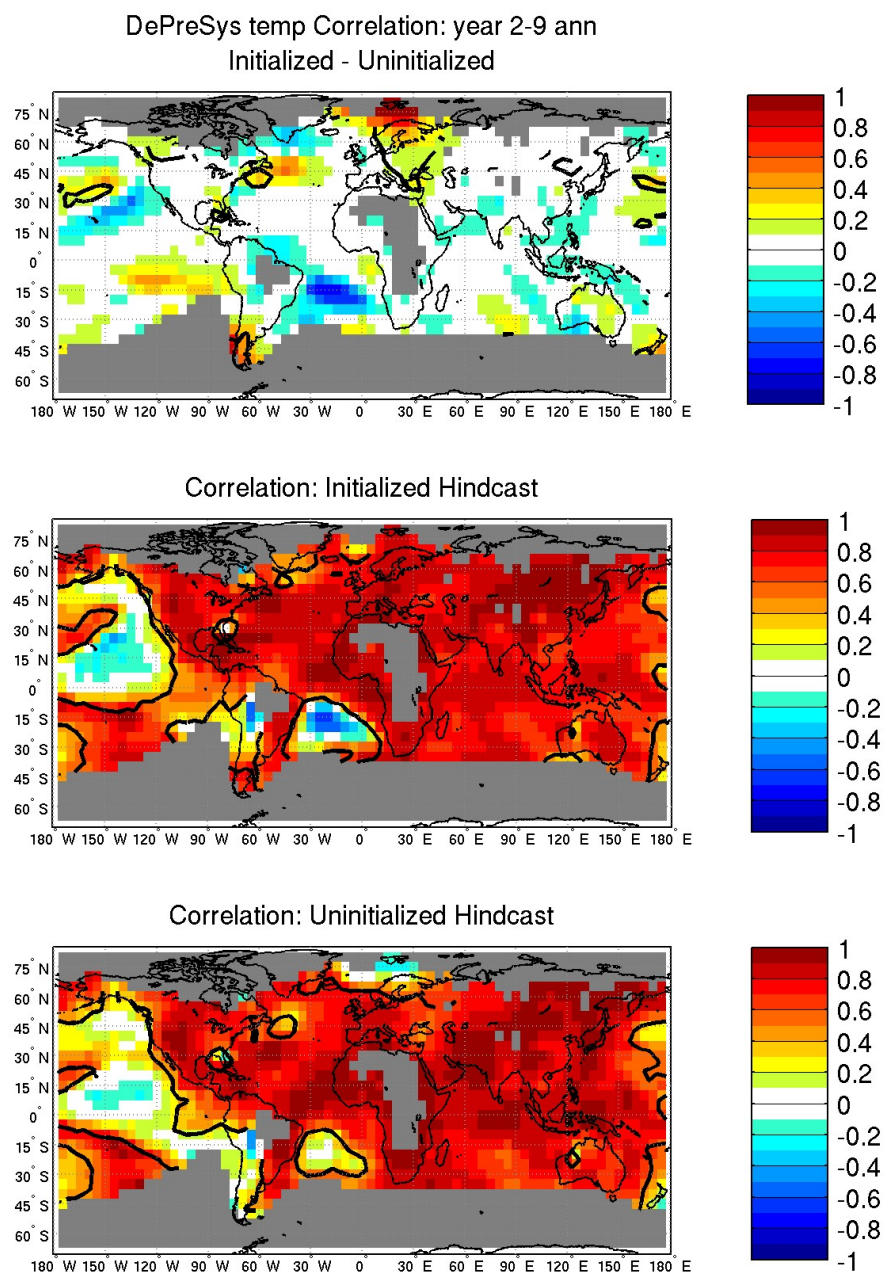
Year 2-5 | Year 6-9 | Year 2-9



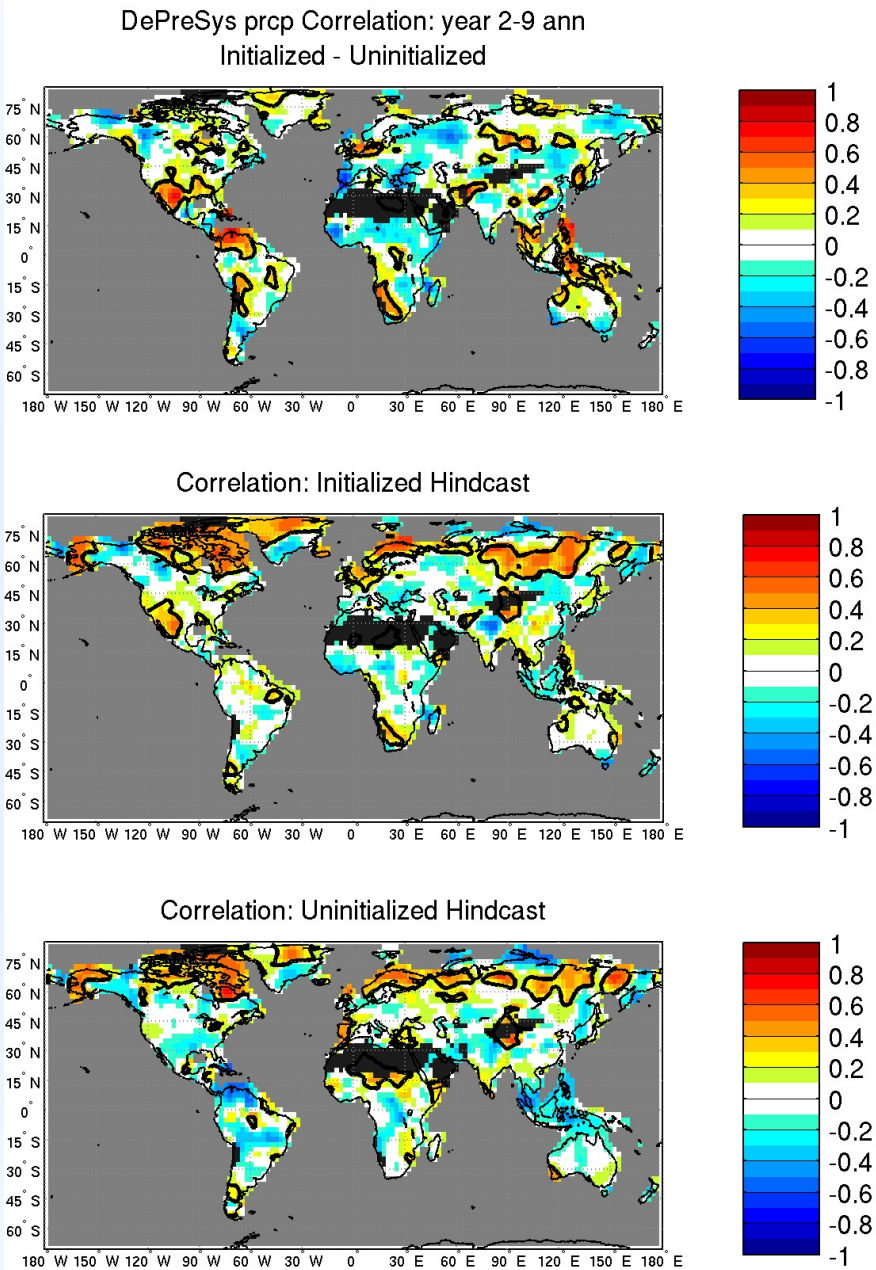
Using ALL start years



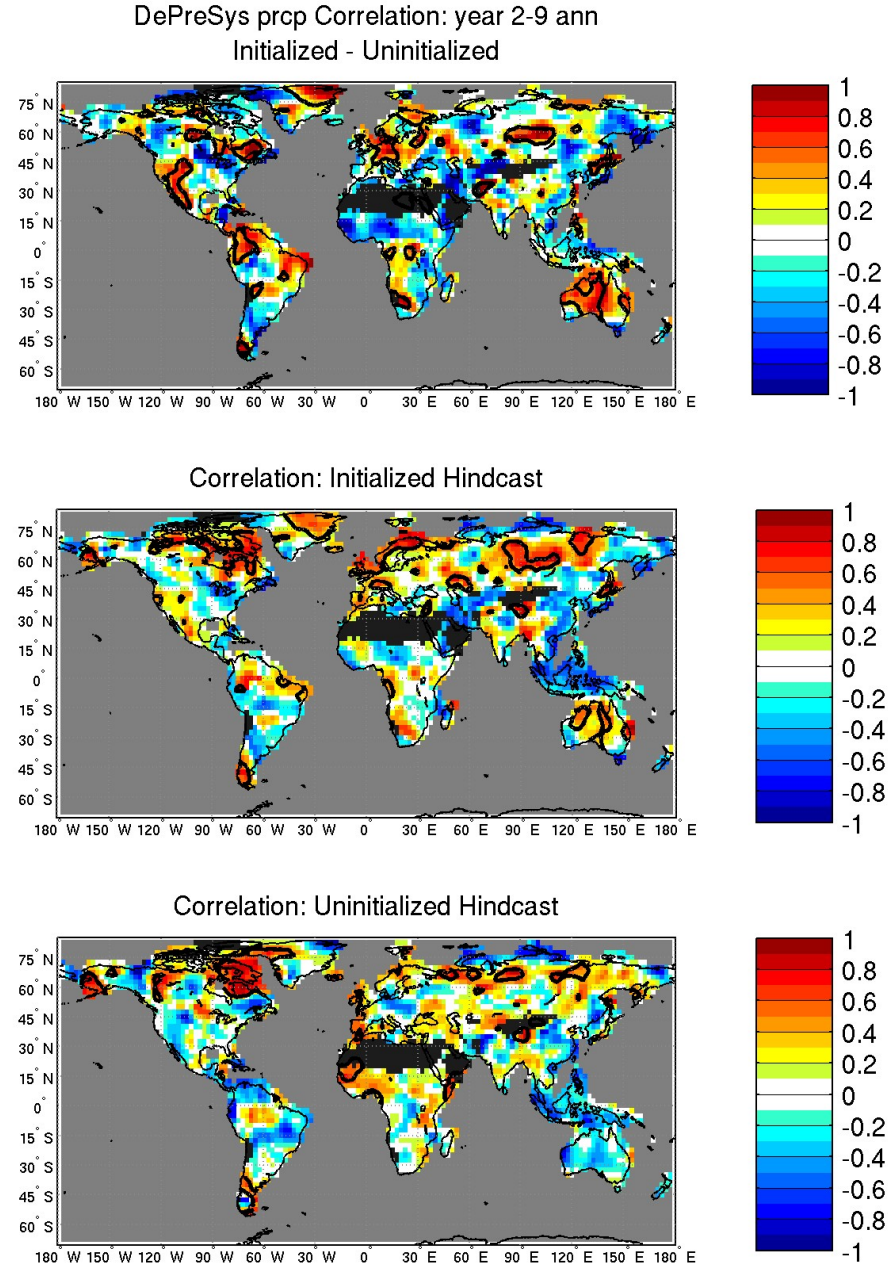
Using Every-5-year starts



Using ALL start years



Using Every-5-year starts



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 6-9, Years 2-9

Spatial scale: Grid scale, spatially-smoothed

Probabilistic Metrics: TEMPERATURE

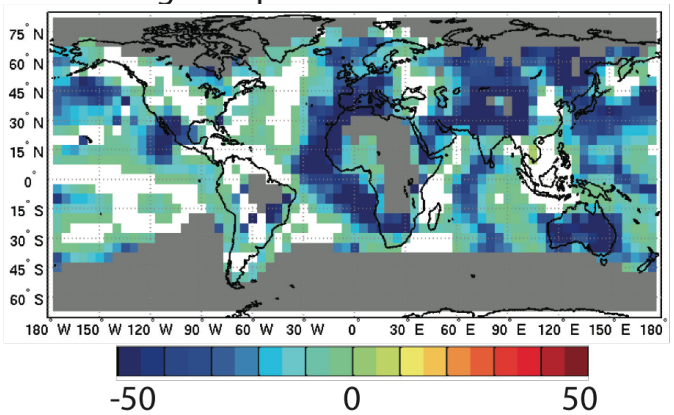
Ens Mean + Std Error
vs
Ens Mean + Ens Spread

Ens Mean + Ens Spread
vs
Climo Distribution

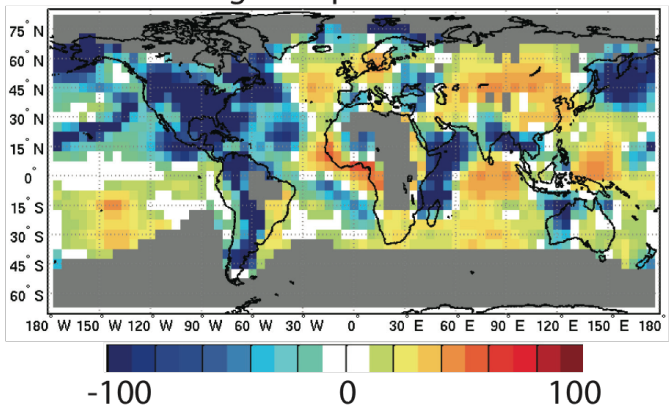
Ens Mean + Std Error
vs
Climo Distribution

DePreSys CRPSS (%): Years 2-9

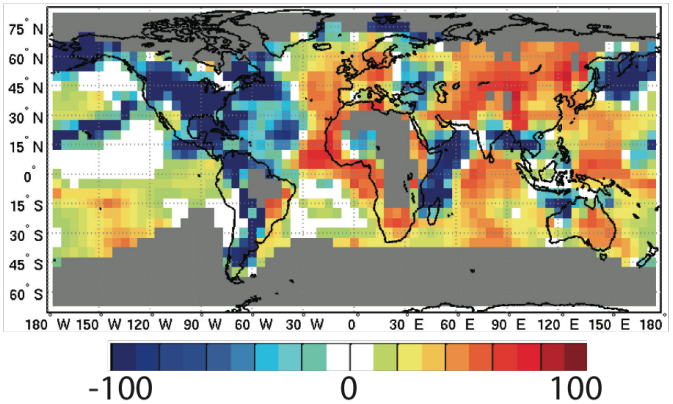
Avg Ens spread vs Standard Error



Avg Ens Spread v Clim

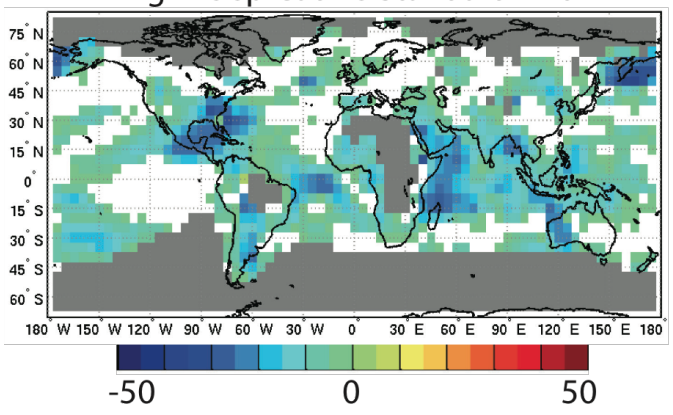


Ens Mean Standard Error vs Clim

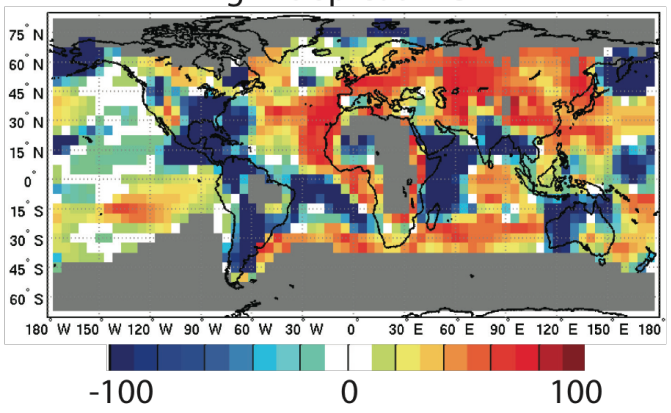


CanCM4 CRPSS (%): Years 2-9

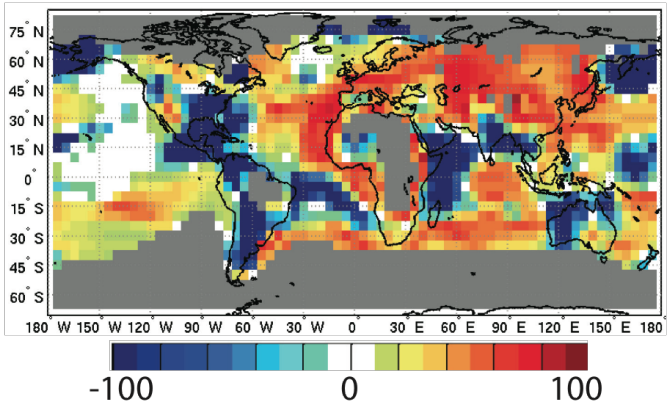
Avg Ens spread vs Standard Error



Avg Ens Spread v Clim



Ens Mean Standard Error vs Clim



Probabilistic Metrics: PRECIPITATION

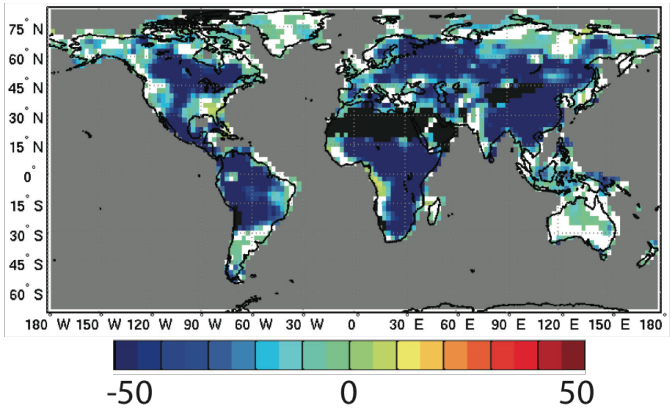
Ens Mean + Std Error
vs
Ens Mean + Ens Spread

Ens Mean + Ens Spread
vs
Climo Distribution

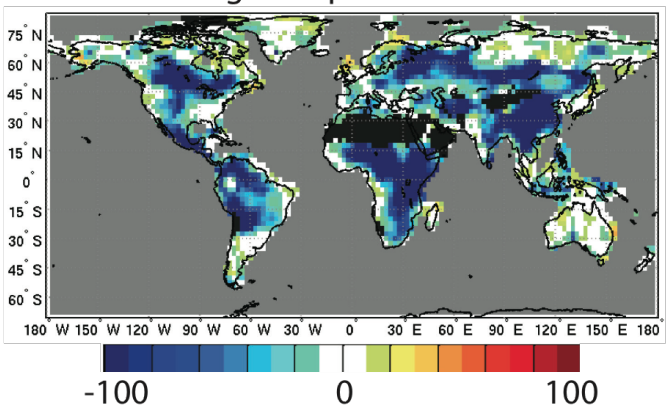
Ens Mean + Std Error
vs
Climo Distribution

DePreSys CRPSS (%): Years 2-9

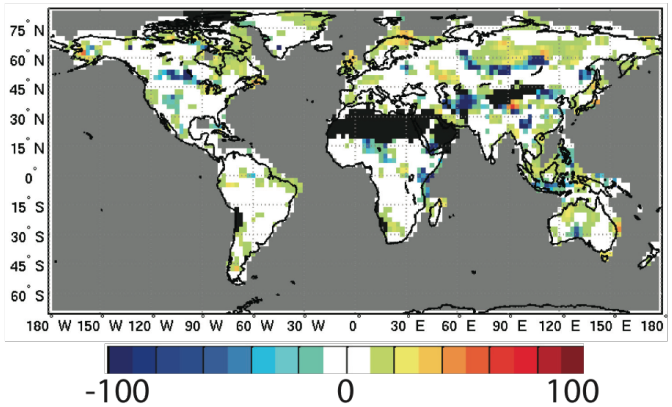
Avg Ens spread vs Standard Error



Avg Ens Spread v Clim

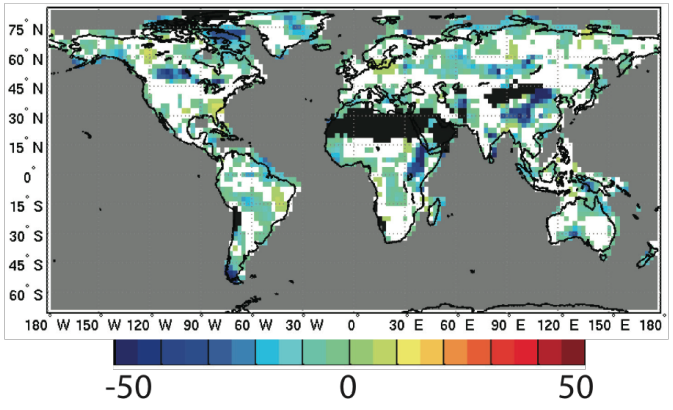


Ens Mean Standard Error vs Clim

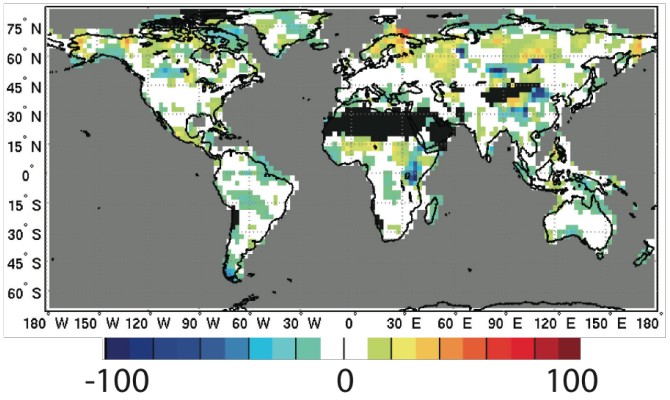


CanCM4 CRPSS (%): Years 2-9

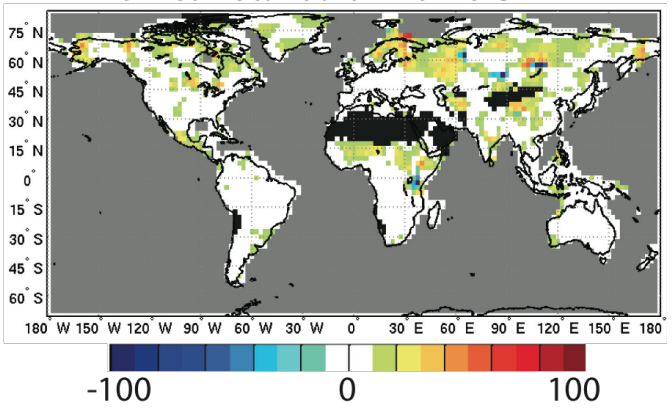
Avg Ens spread vs Standard Error



Avg Ens Spread v Clim



Ens Mean Standard Error vs Clim



Preliminary Conclusions on Skill

- Different hindcasts differ in where they have skill
  - They also differ on which regions improve due to initialization
- Increased sample size (more start years) gives more robust skill estimates
- Accuracy is an important consideration in use of hindcast data
  - Both Correlation & Bias are involved
- Uncertainty is also important
  - Ensemble spread $\neq$ uncertainty

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 in review for Climate Dynamics.

For more on skill of these hindcasts and other CMIP5 decadal hindcast experiments visit http://clivar-dpwg.iri.columbia.edu

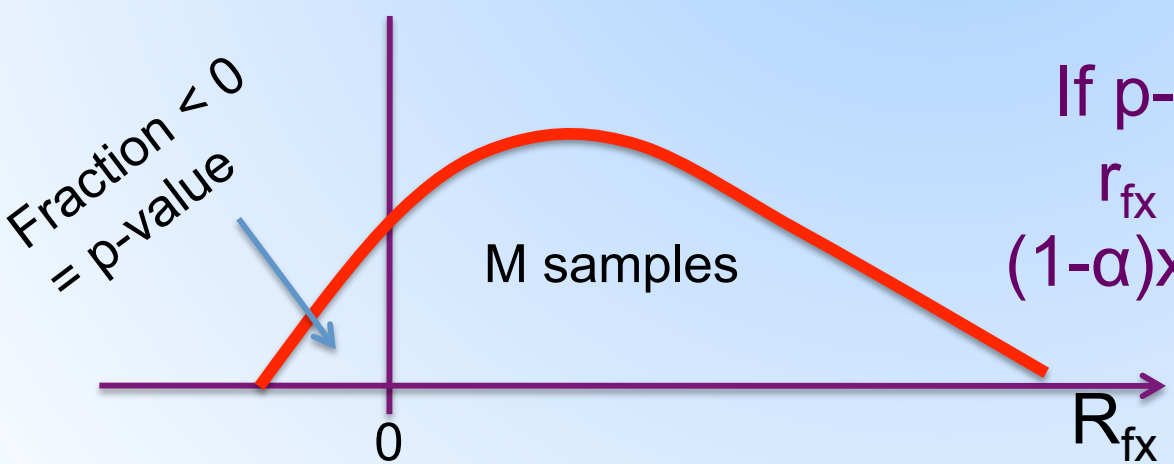
Statistical Significance: Non-parametric bootstrap

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