Global Medium Range NWP to Seasonal Verification

Joe Tribbia NCAR

Talk given at ICAP workshop Boulder, Co
22 October 2014

Outline

- ICAP questions in Medium to Seasonal Range
- Hurricane Cartoon
- Forecast Verification Medium Range
- Back to Cartoon
- Monthly to Seasonal Prediction Challenges
- Seasonal Verification Challenges
- What it means for ICAP

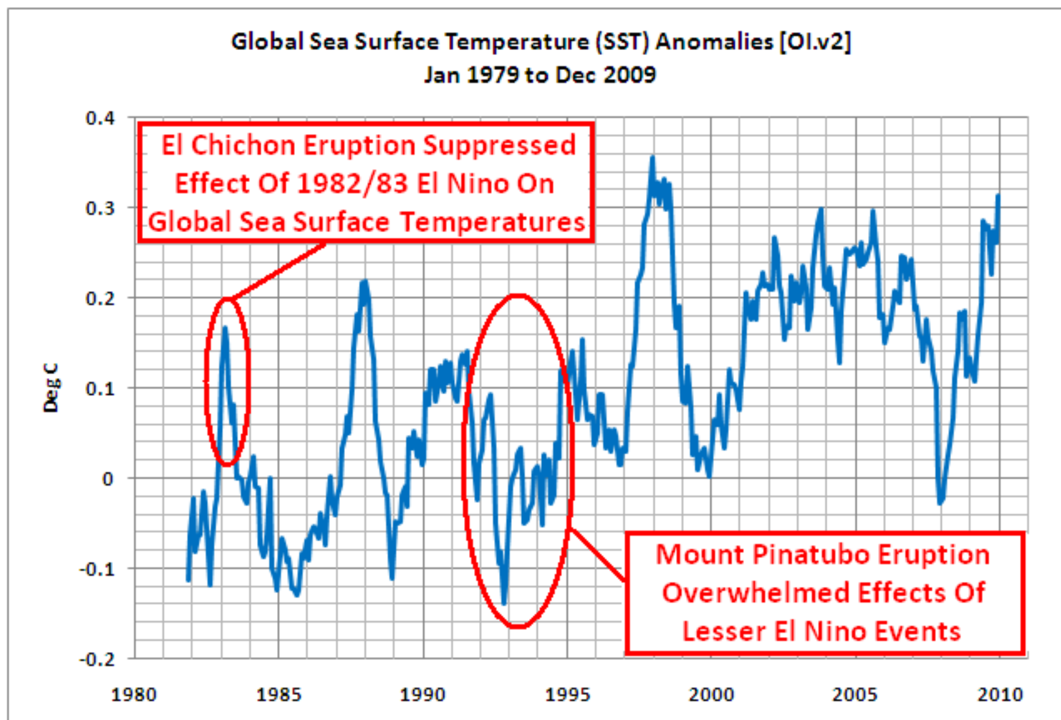
Medium Range to Monthly Aerosol Prediction
WHY?

Country with serious air quality issues

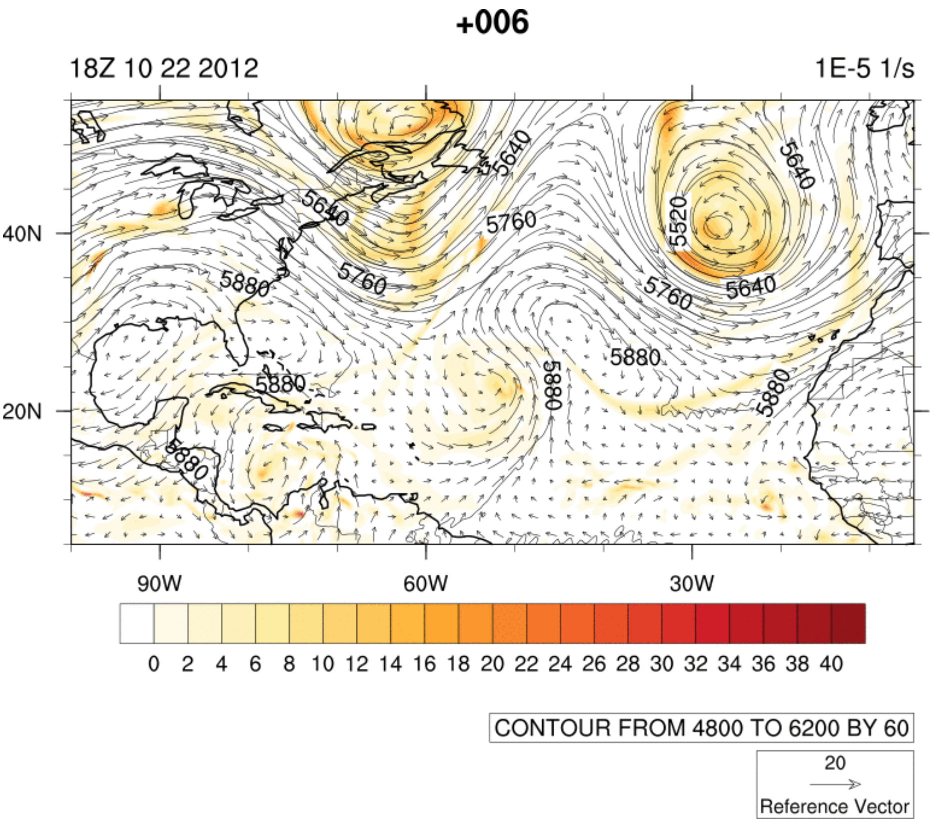
Is hosting a high – profile international even in a major city. Should mitigation efforts be undertaken and if so when?

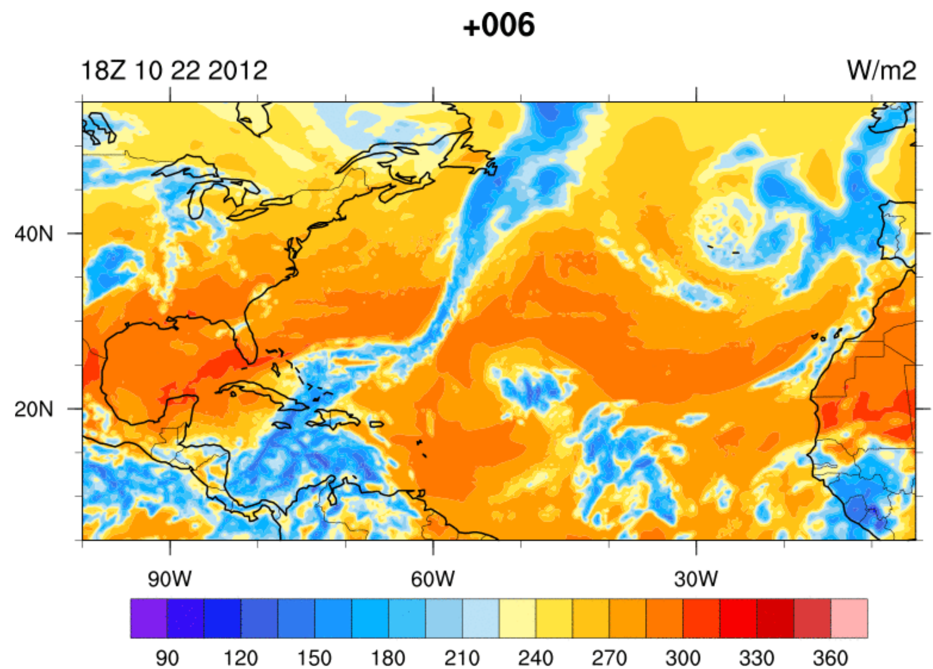
Seasonal Aerosol Prediction
WHY?

Impact of surprise aerosol emission on seasonal forecast



Cartoon: CAM-SE forecasting hurricane Sandy (Courtesy of Colin Zarzycki)



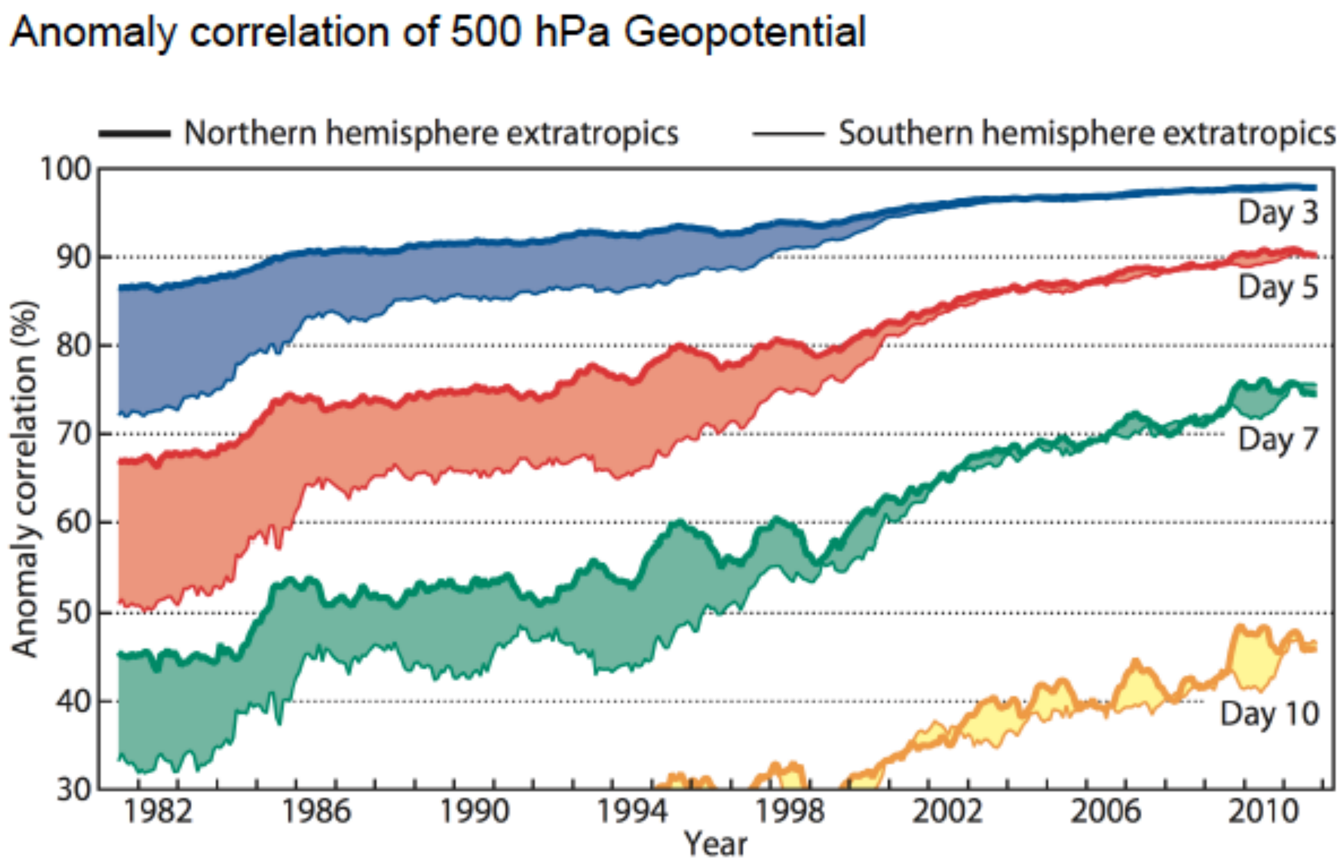


Hurricane Sandy, 13 km forecast initialized 10/22/12 12Z

Medium to Monthly Prediction

- Range: Day 5 to Day 30
- Deterministic success unlikely-> Ensemble
- Atmospheric Phenomenology: Blocking, Planetary waves, Teleconnections, MJO (anything with persistence)
- Probabilistic Verification Measures: Talagrand Diagram, Brier, ROC (c.f. Popper for the philosophical challenge)

Normally for 500 hPa height
Skill drops quickly from 5-days on



ECMWF 2012

Reliability of the ensemble spread

- Consider ensemble variance (spread) for an M-member ensemble

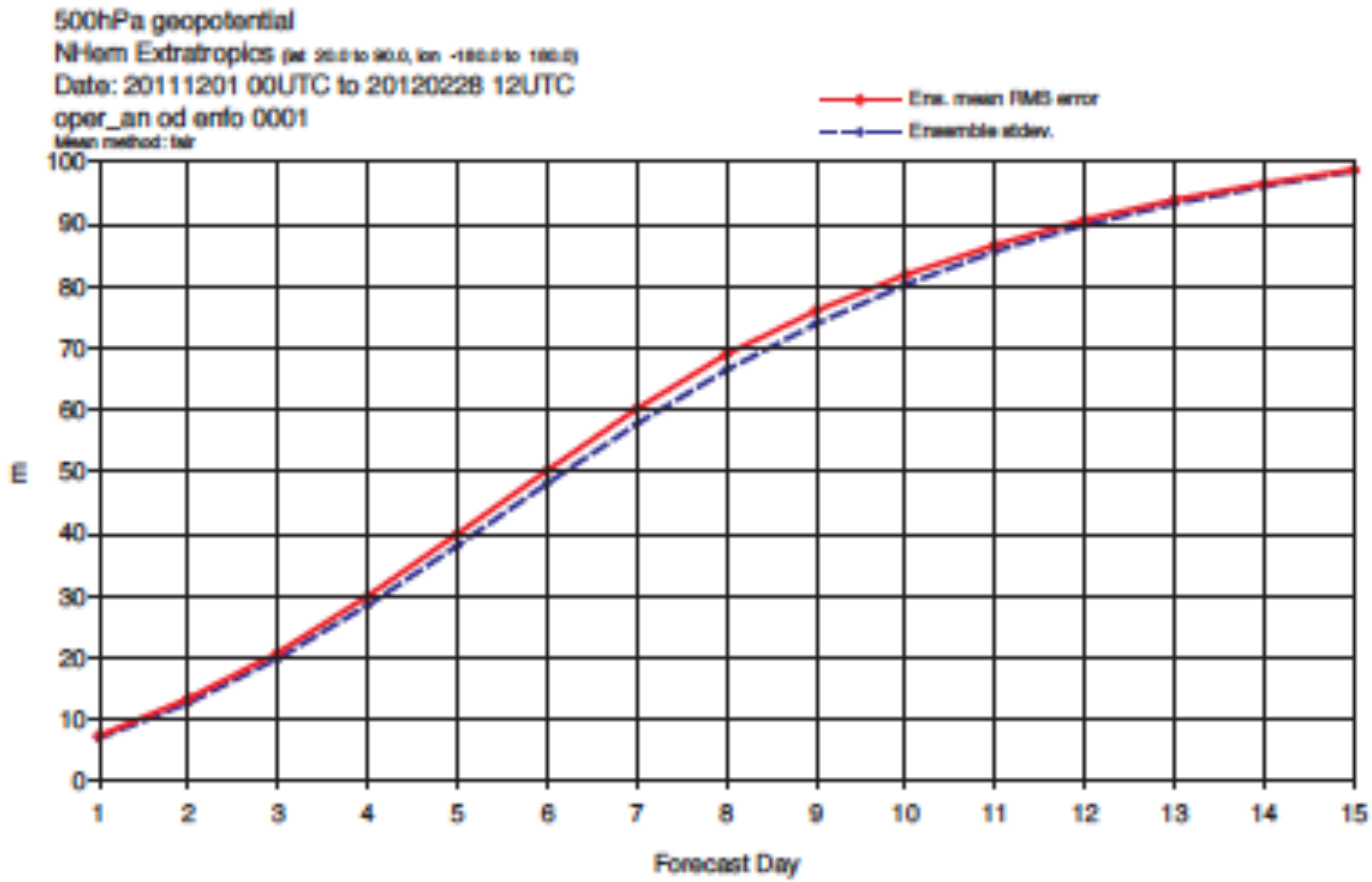
$$\frac{1}{M-1}\sum_{j=1}^{M}(x_j-\overline{x})^2$$

- and the squared error of the ensemble mean

$$(\overline{x}-y)^2$$

- Average the two quantities for many locations and/or start times.
- The averaged quantities have to match for a reliable ensemble (within sampling uncertainty).
- Finite ensemble size can be corrected for in the estimation of the error of the ensemble mean and the ensemble variance

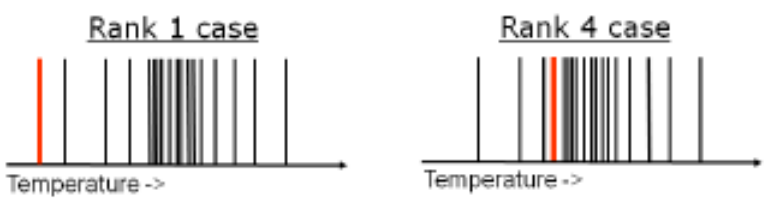
ECMWF EPS 500 hPa

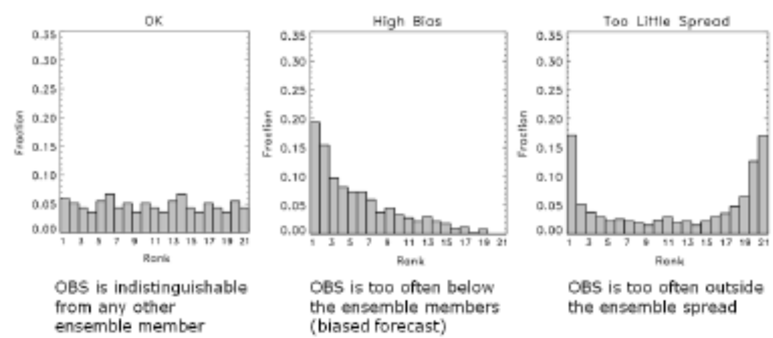


Rank Histogram (Talagrand)

Determine where observation lies relative to the ensemble

Flat histogram necessary for reliable ensemble





Yes/No forecast
Freeze warning

Reliability Diagram

Plot Forecast Frequency versus

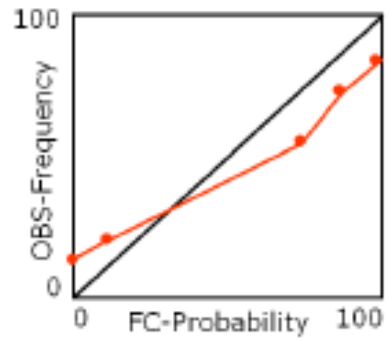
Observed Frequency

Reliable Probabilistic Forecast

System should lie on diagonal

This example shows overconfident forecast system

Using bins any forecast can be turned into a Yes/No forecast ($T_1 \leq T < T_2$)



FC Prob.	# FC	OBS-Frequency (perfect model)	OBS-Frequency (imperfect model)
100%	8000	8000 (100%)	7200 (90%)
90%	5000	4500 (90%)	4000 (80%)
80%	4500	3600 (80%)	3000 (66%)
....	....	....	....
....	....	....	....
....	....	....	....
10%	5500	550 (10%)	800 (15%)
0%	7000	0 (0%)	700 (10%)

Brier Score

- Consider an event, e, T<32°F
- Build a 2x2 Contingency table

	e observed	
e predicted	Yes	No
Yes	hits *a*	false alarms *b*
No	misses *c*	correct rejections *d*

- Hit rate H=a/(a+c)
- False alarm rate F=b/(b+d)
- Sample size N=a+b+c+d

Multiple Contingency Table

The joint distribution of forecasts and observations for a M-member ensemble can be summarized in a $(M+1) \times 2$ contingency table $\mathbf{T}$

$$\text{sample size } N = \sum_{j=0}^{M} n_j + \sum_{j=0}^{M} \tilde{n}_j$$

Each row corresponds to a probability value, e.g. $p = j/M \longrightarrow$

e pred. by m_e members	e observed Yes	No
M	n_M	$\tilde{n}_M$
$M-1$	n_{M-1}	$\tilde{n}_{M-1}$
$\dots$	$\dots$	$\dots$
j	n_j	$\tilde{n}_j$
$\dots$	$\dots$	$\dots$
1	n_1	$\tilde{n}_1$
0	n_0	$\tilde{n}_0$

Brier Score Defined

$$\mathrm{BS} = \frac{1}{N}\sum_{k=1}^{N}(p_k - o_k)^2$$

- p_k is the predicted probability of the k-th forecast and $o_k = 1$ (0) if the event occurred (did not occur)
- The Brier score BS is the **mean squared error** of the probability forecast.
- The BS can be decomposed in three components that measure
  - reliability
  - resolution
  - uncertainty

Brier Score Decomposed
BS=REL-RES+UNC

Reliability: deviation of observed
Relative frequency from forecasted
Probability

$$REL = \frac{1}{N}\sum_{j=0}^{M} l_j(\overline{o}_j - p_j)^2$$

Resolution: ability of forecast system to
Recognize when the observed probability
Differs from average

$$RES = \frac{1}{N}\sum_{j=0}^{M} l_j(\overline{o}_j - \overline{o})^2$$

Uncertainty: Variance of the
Observed (0/1) in the sample

$$UNC = \overline{o}(1 - \overline{o})$$

N=total number of cases
M= number of probability bins
p_j= j/M the probability in bin j
l_j=number of cases in bin j
$\overline{o}_j$ =n_j/l_j frequency of event occurring when forecasted with probability p_j
$\overline{o}$ = event frequency in the whole sample

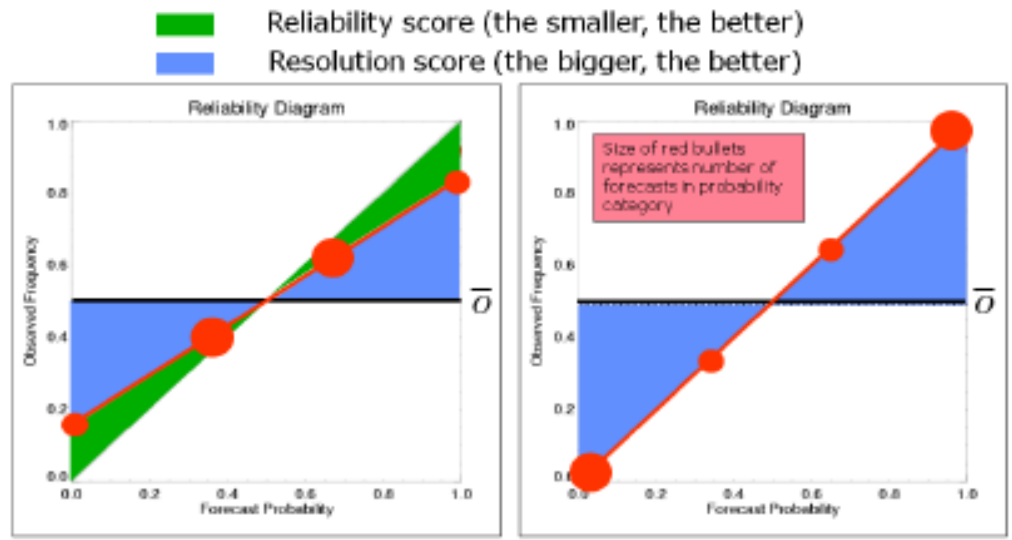
Brier Skill Score

- $BSS = 1 - BS/BS_{REF}$
- Skill scores are used to compare the performance of forecasts with that of a reference forecast (e.g. climatological distribution)
- Has an element of Information/Entropy skill score
- They are defined so that the perfect forecast has a skill score of 1 and the reference forecast has the skill score of 0
- skill score =(actual fc –ref)/(perfect fc- ref)
- positive (negative) BSS -> forecast is better (worse) than the reference forecast

Brier Score

Attributes Diagram

Brier Skill in terms of (relative) Reliability and Resolution



$BSS = 1 - BS/BS_{REF}$

$= 1 - (REL - RES + UNC)/UNC$

$= (RES - REL)/UNC$

Discrimination and ROC

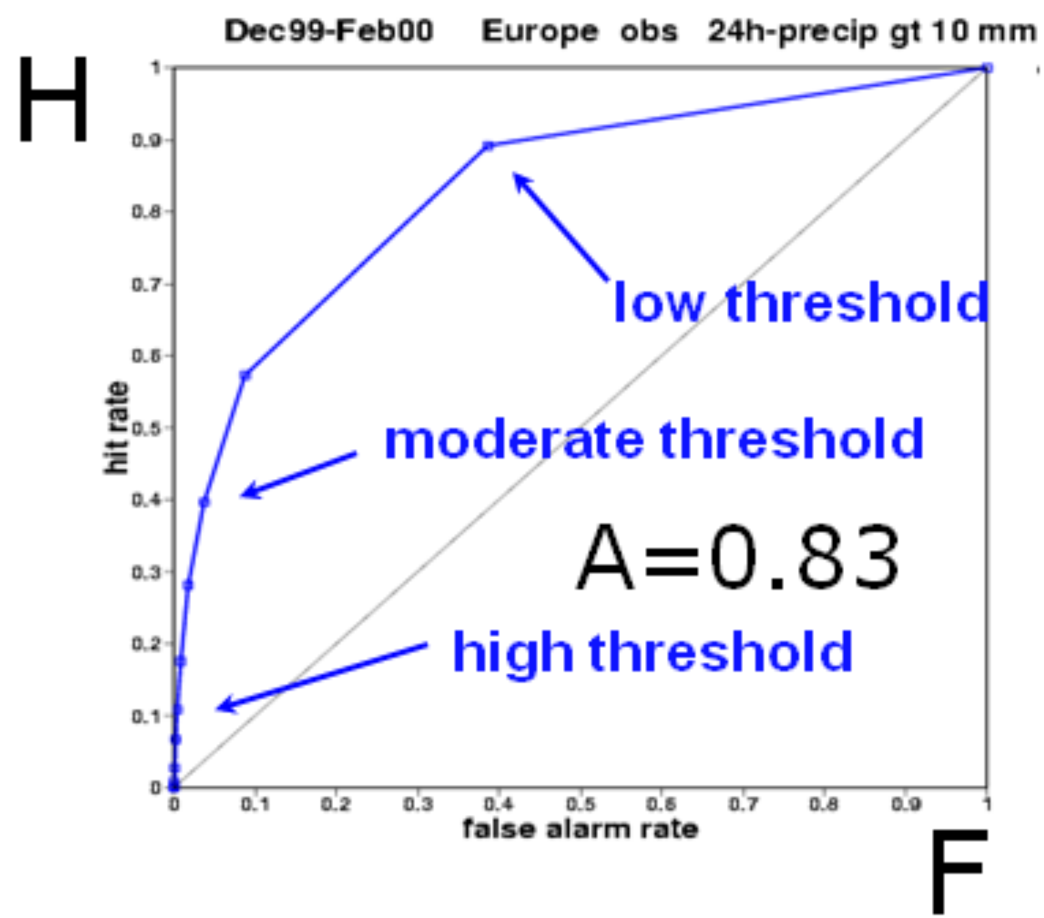
- Until now, we looked at the question: What is the distribution of observations o if the forecast system predicts an event to occur with probability p?
- To measure the ability of a forecast system to discriminate between occurrence and non-occurrence of an event, one has to ask:

  What distributions of probabilities have been predicted when the event occurred and when it did not occur?

For any probability threshold p_i one can then determine the hit rate $H_i = a/(a+c)$ and the false alarm rate $F_i = b/(b+d)$

- The relative operating characteristic (ROC, also referred to as receiver operating characteristic) is the diagram that shows H versus F for all probability thresholds

Relative Operating Characteristic
ROC



random forecast (independent of observed event) on diagonal
summary measure: area under the ROC in the interval [0.5,1.0]

Logarithmic Score (LS)
An Entropy-like Score

- also known as ignorance score (Good 1952, Roulston and Smith 2002)

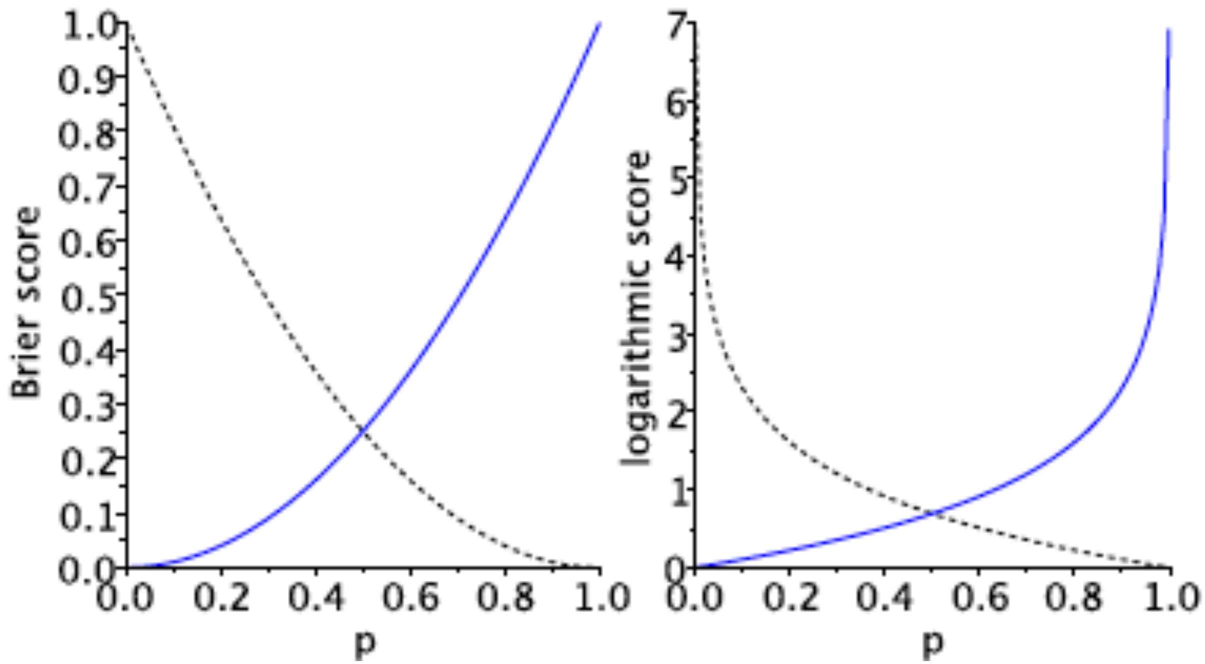
$$LS = \frac{1}{N}\sum_{k=1}^{N} o_k \log p_k + (1-o_k)\log(1-p_k)$$

- The score ranges between 0 and 1. The latter happens if the predicted probability is zero and the event occurs (or if p = 1 and the event does not occur).
- The ignorance score is more sensitive to the cases with probability close to 0 and close to 1 than the Brier score.

Brier score versus logarithmic score

- event occurs (dotted) ; event does not occur (solid)

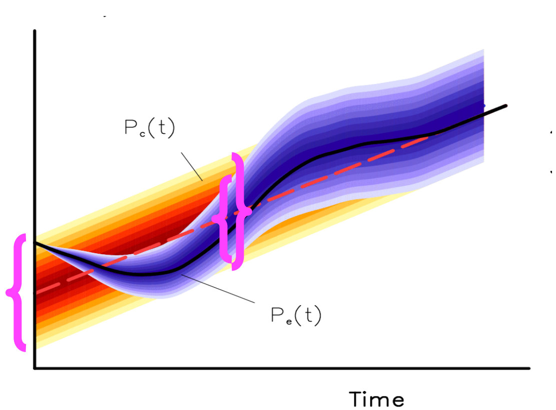
  $(p-1)^2$ and p^2 -log(p) and -log(1 -p)

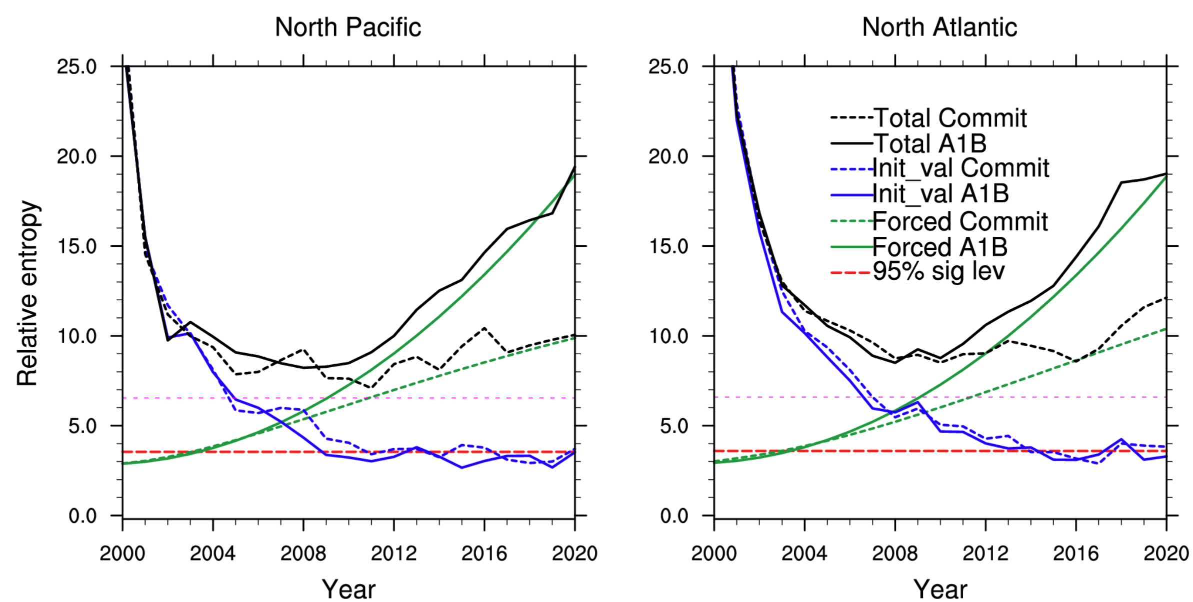


Relative Entropy Metric

$$R = \int_S P_e(s) \log_2 \left[\frac{P_e(s)}{P_c(s)} \right] ds$$

Examine Sources of Skill : Example Decadal Predictability

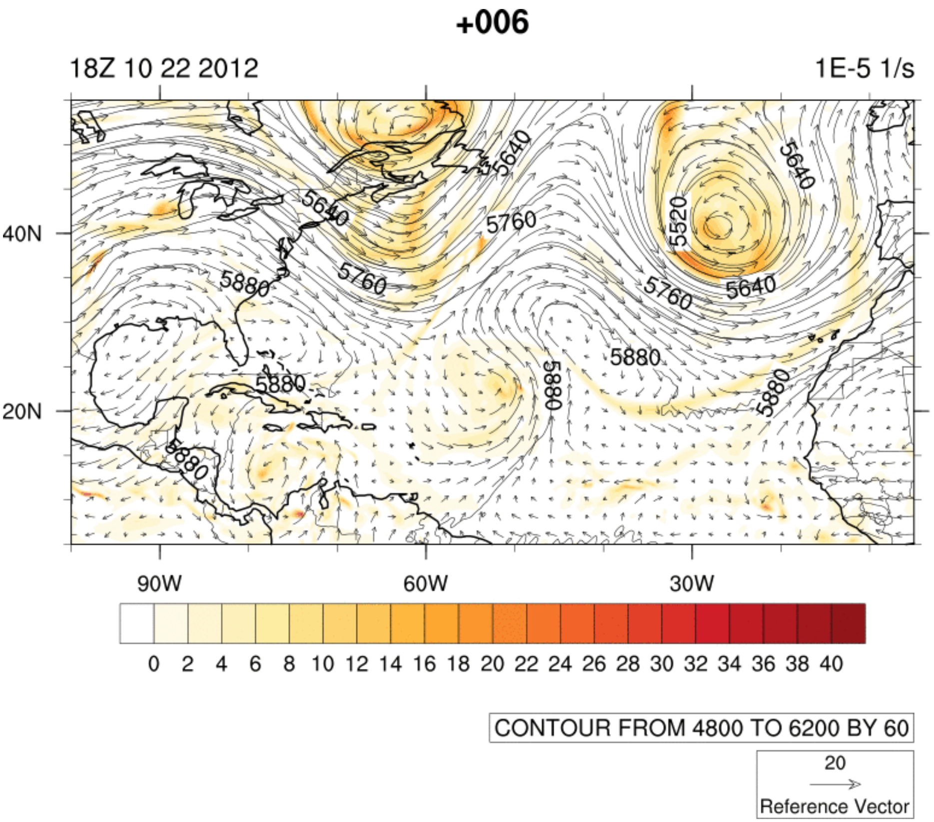


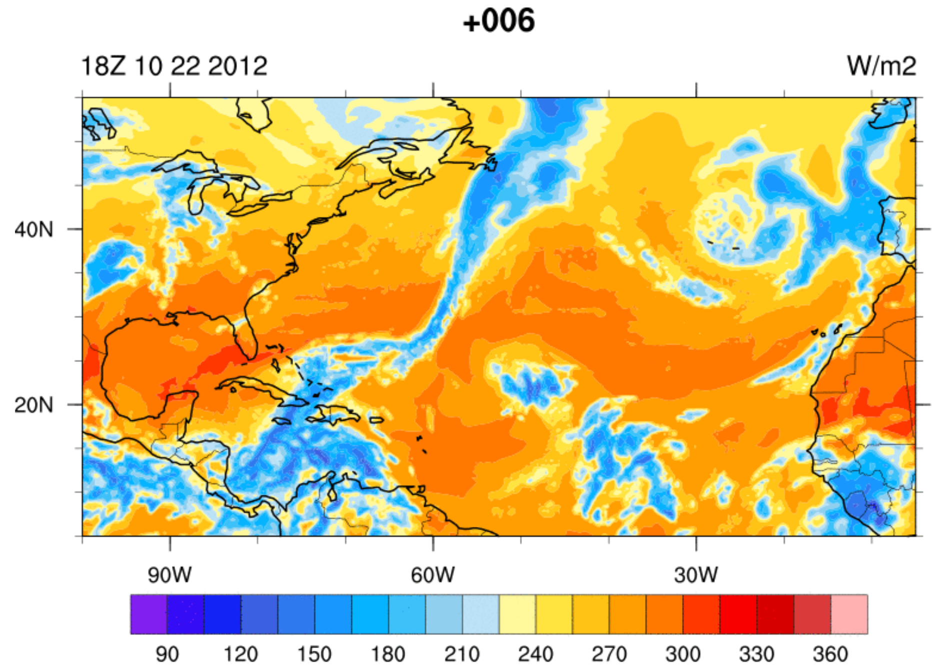


Probabilistic Measures of Ensemble Skill- What is wrong with these?

- Mostly look at the ability of the forecast to reproduce observed statistics
- This is proper for a probability prediction
- But it misses the relative importance of rare events-things we care about
- Symptom of this: BSS saturates at an ensemble size of about 50 members
- Inadequate size for extremes or structured PDFs
- *Notes: Martin Leutbecher*

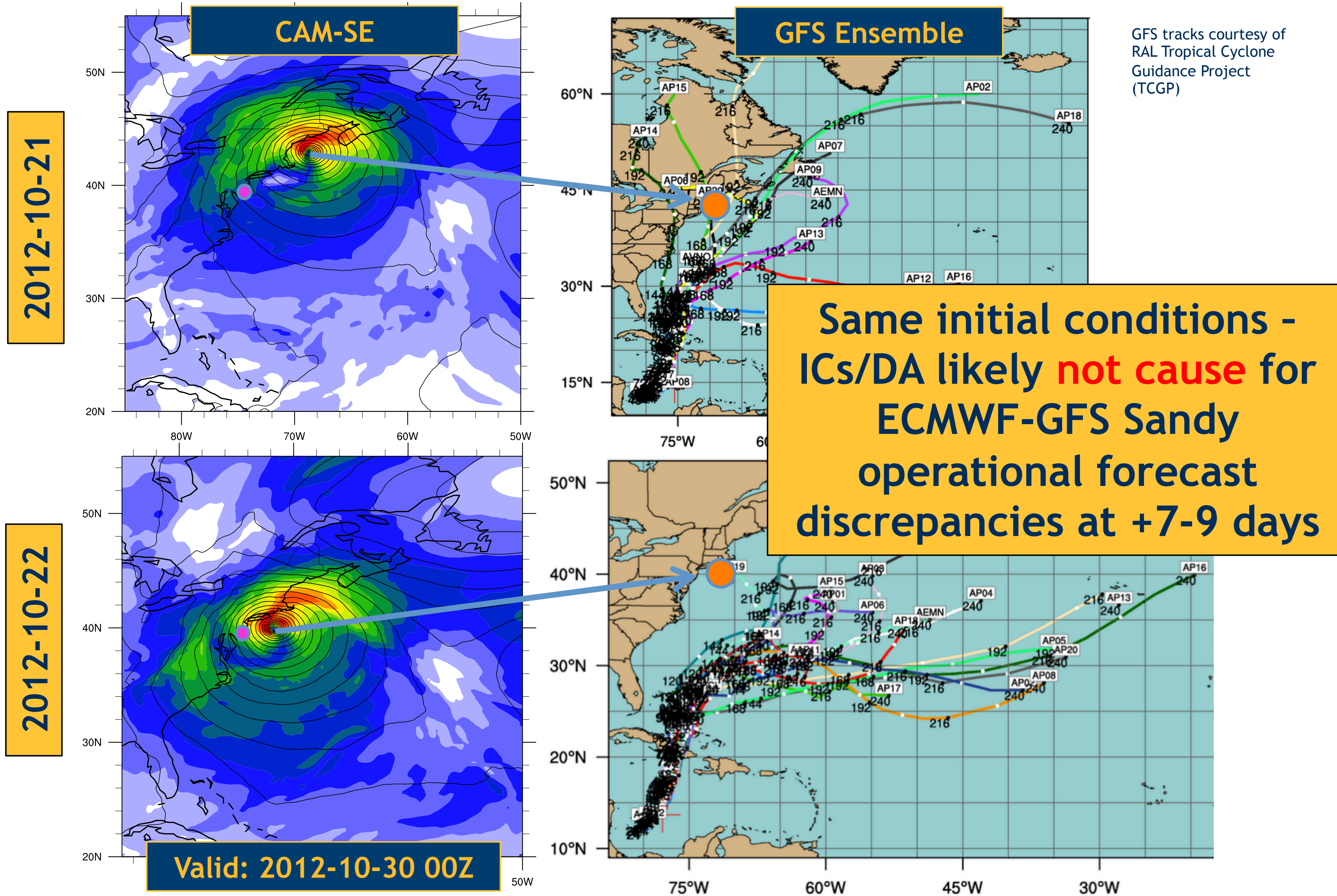
Back to the Cartoon





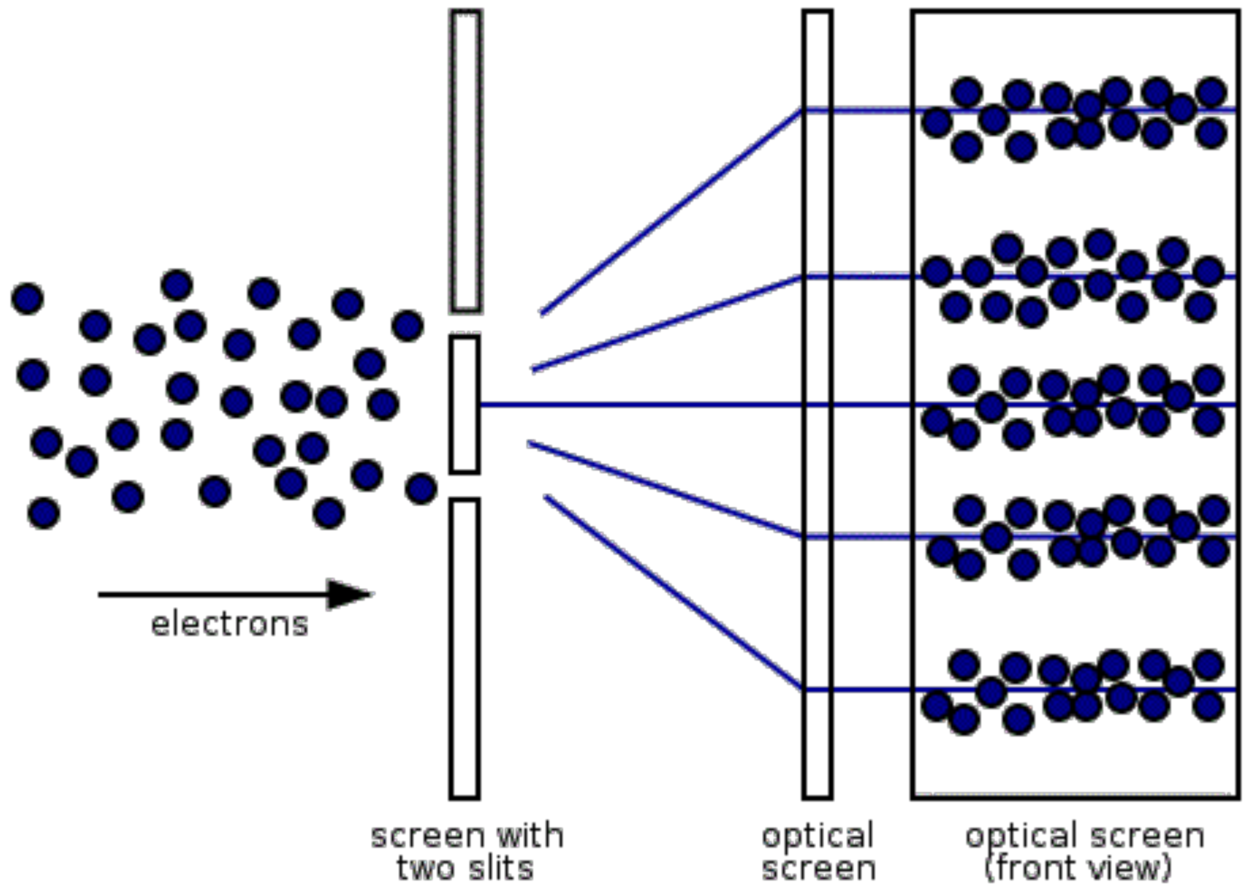
Hurricane Sandy, 13 km forecast initialized 10/22/12 12Z

EMC ensemble does not forecast recurvature at 10 day lead
IS THIS A BAD ENSEMBLE SYSTEM? This question not asked/answered



GFS tracks courtesy of RAL Tropical Cyclone Guidance Project (TCGP)

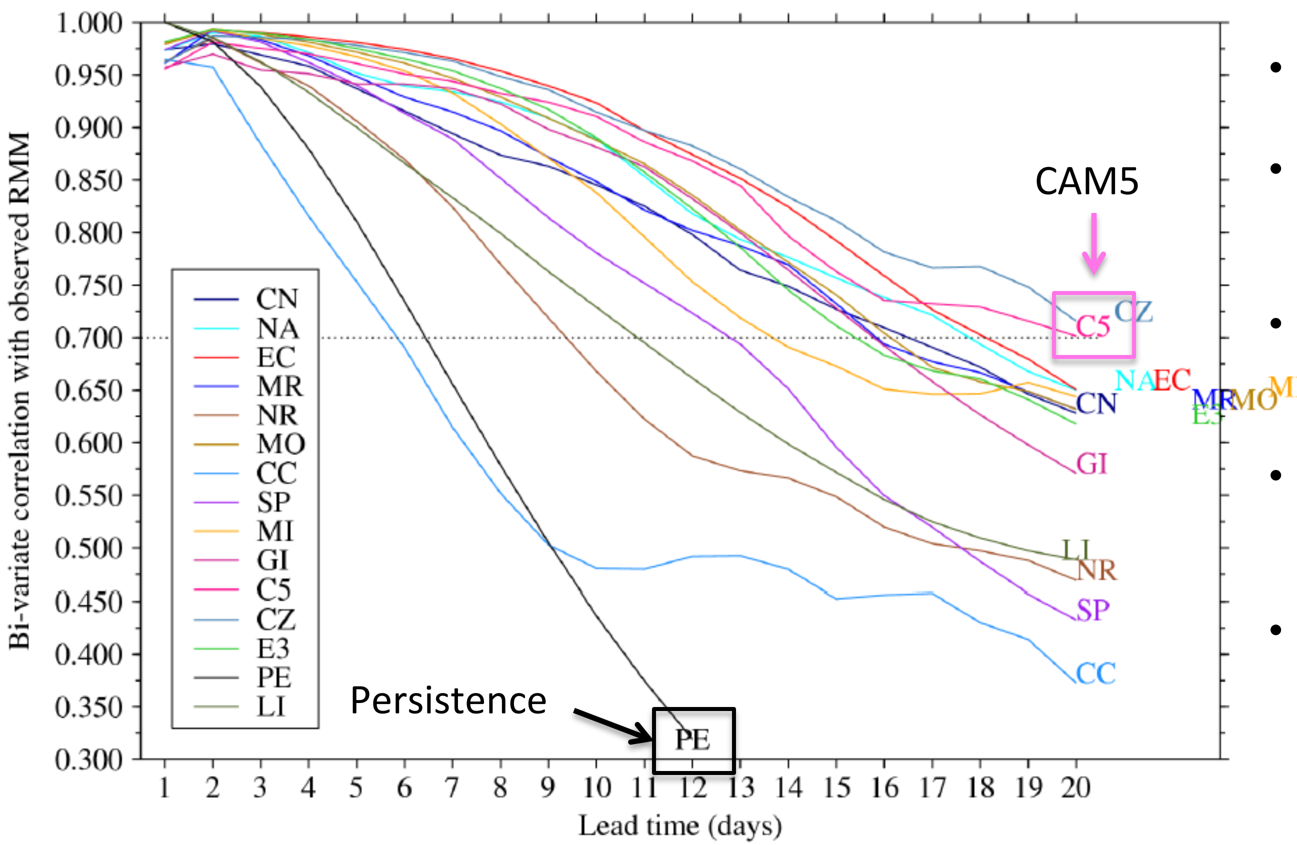
Akin to complaining that quantum mechanics is wrong after 100 electrons



What about Sub-Seasonal to Seasonal Probabilistic Prediction

- Same issues with ensemble verification
- Same skill metrics can be used
- Different fields for verification (SST, time mean continental T_{2m} and Precipitation)
- El Nino largest signal -- intermittent
- Low frequency modes (PNA, NAO, AO, MJO)
- Forecasts done using Coupled System

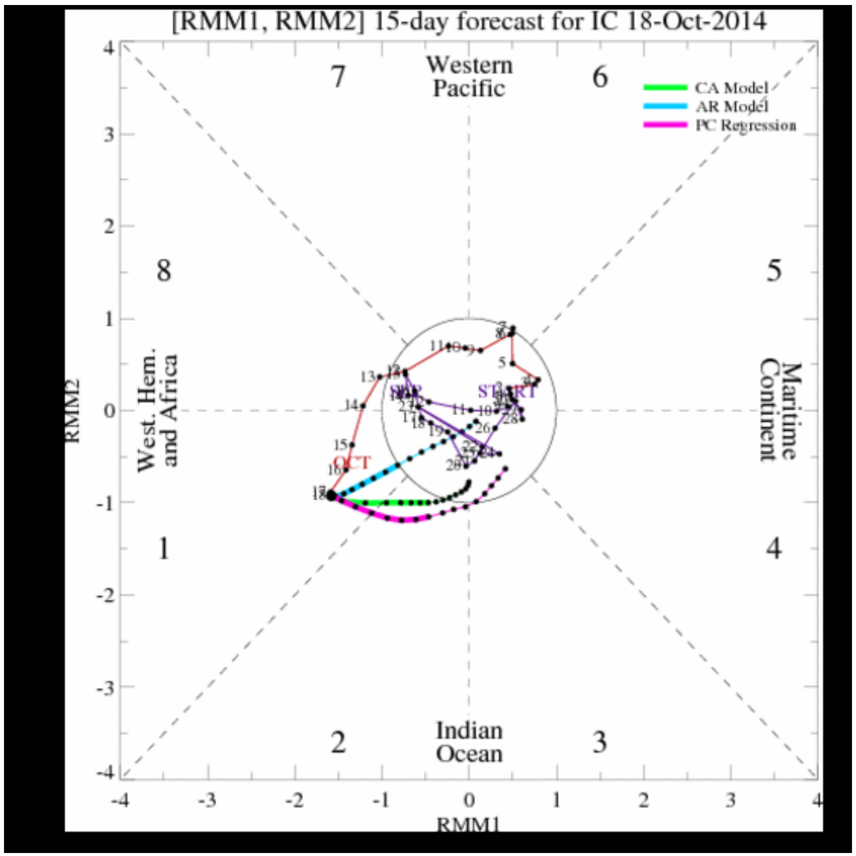
Madden Julian Oscillation (MJO) Hindcasts



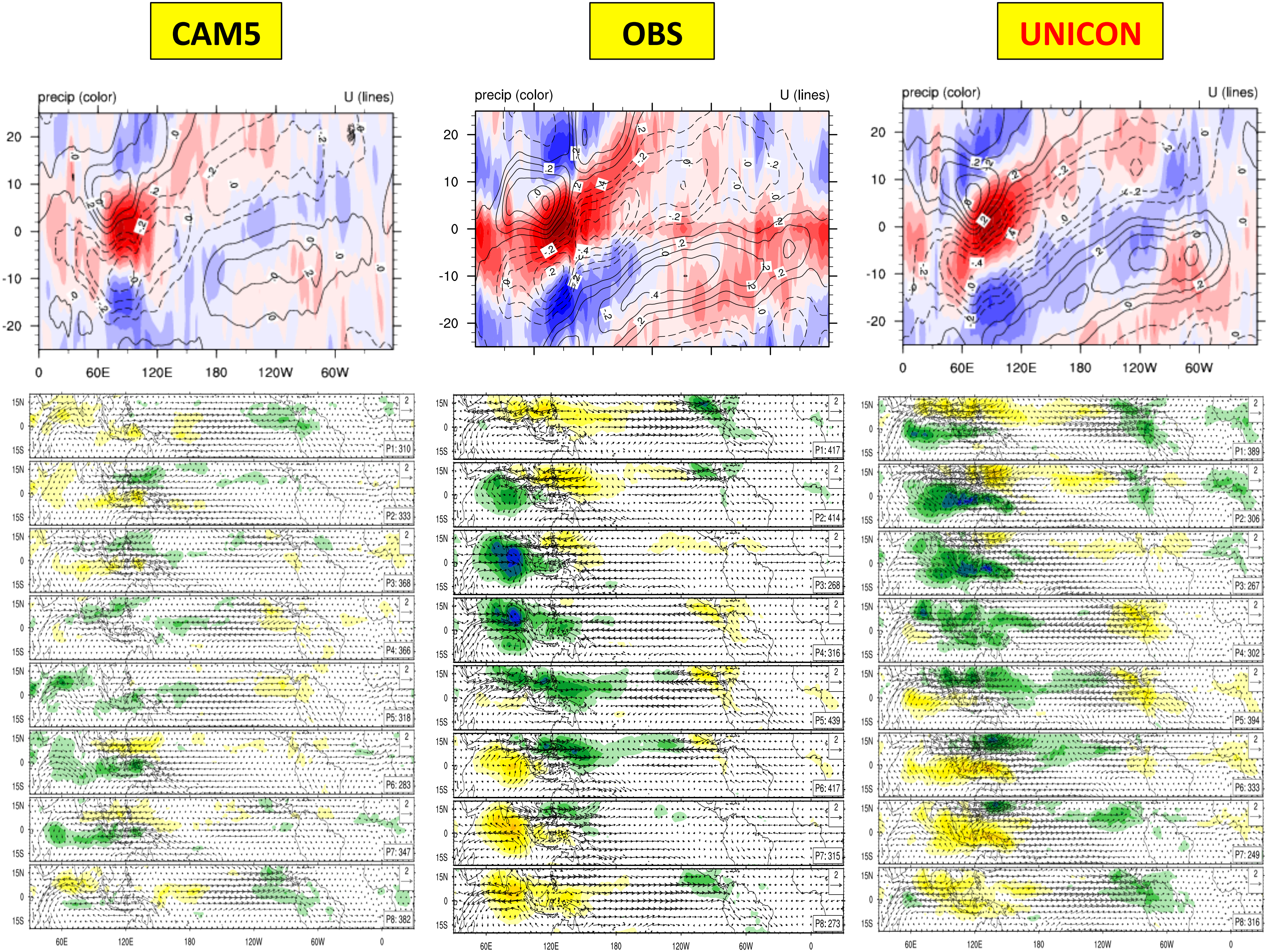
- Initial forecast mode (CAPT)
- During MJO-DYNAMO Campaign
- Combined bivariate mode of MJO variability (RMM)
- CAM5 only model to retain skill out to 20 days.
- Top performer among participating CMIP5 models.

Courtesy: Nick Klingaman, U. Reading, UK

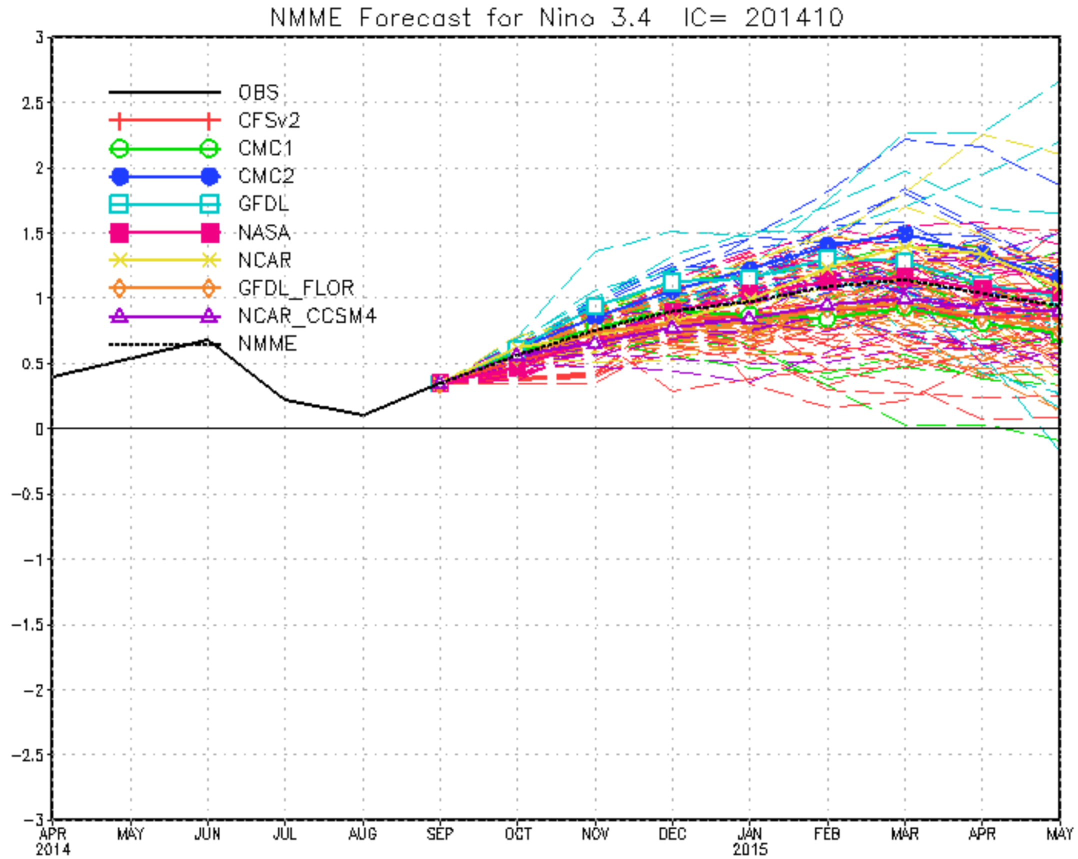
MJO verification using phase diagram



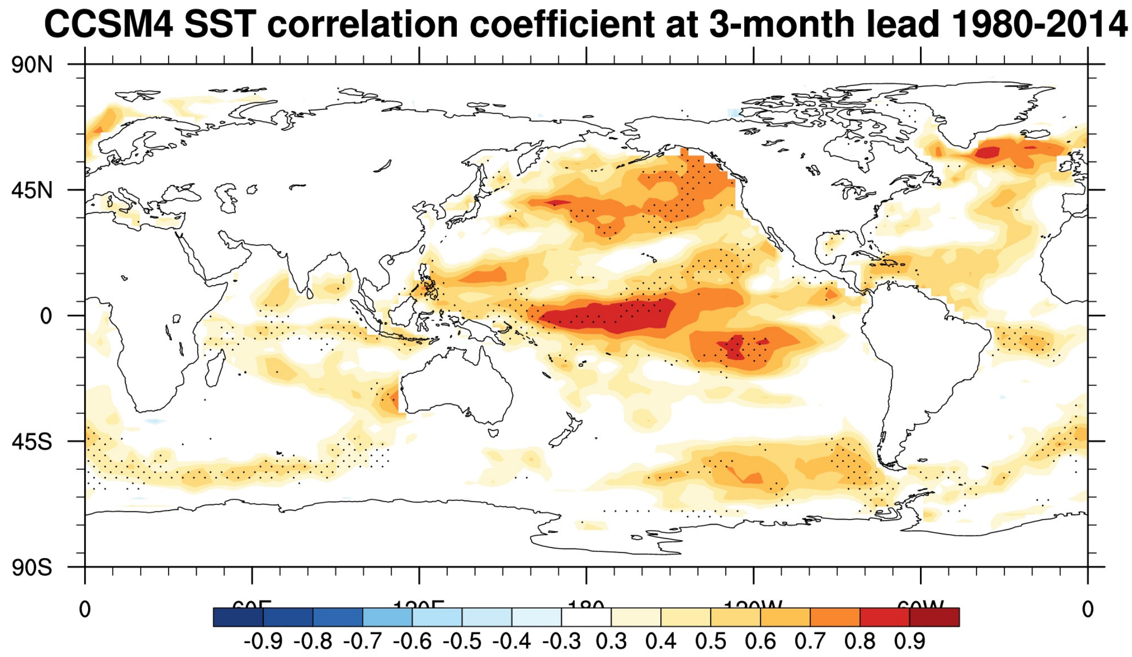
- RMM1 and RMM2 are multivariate EOFs that capture MJO wind and OLR (precip) anomalies
- An MJO event is a counter-clockwise rotation in RMM1, RMM2 plane


CAM5
OBS
UNICON
precip (color)
U (lines)
precip (color)
U (lines)
precip (color)
U (lines)
0
60E
120E
180
120W
60W
P1: 310
P2: 333
P3: 368
P4: 366
P5: 318
P6: 283
P7: 347
P8: 382
P1: 417
P2: 414
P3: 268
P4: 316
P5: 439
P6: 417
P7: 315
P8: 273
P1: 389
P2: 306
P3: 267
P4: 302
P5: 394
P6: 333
P7: 249
P8: 316

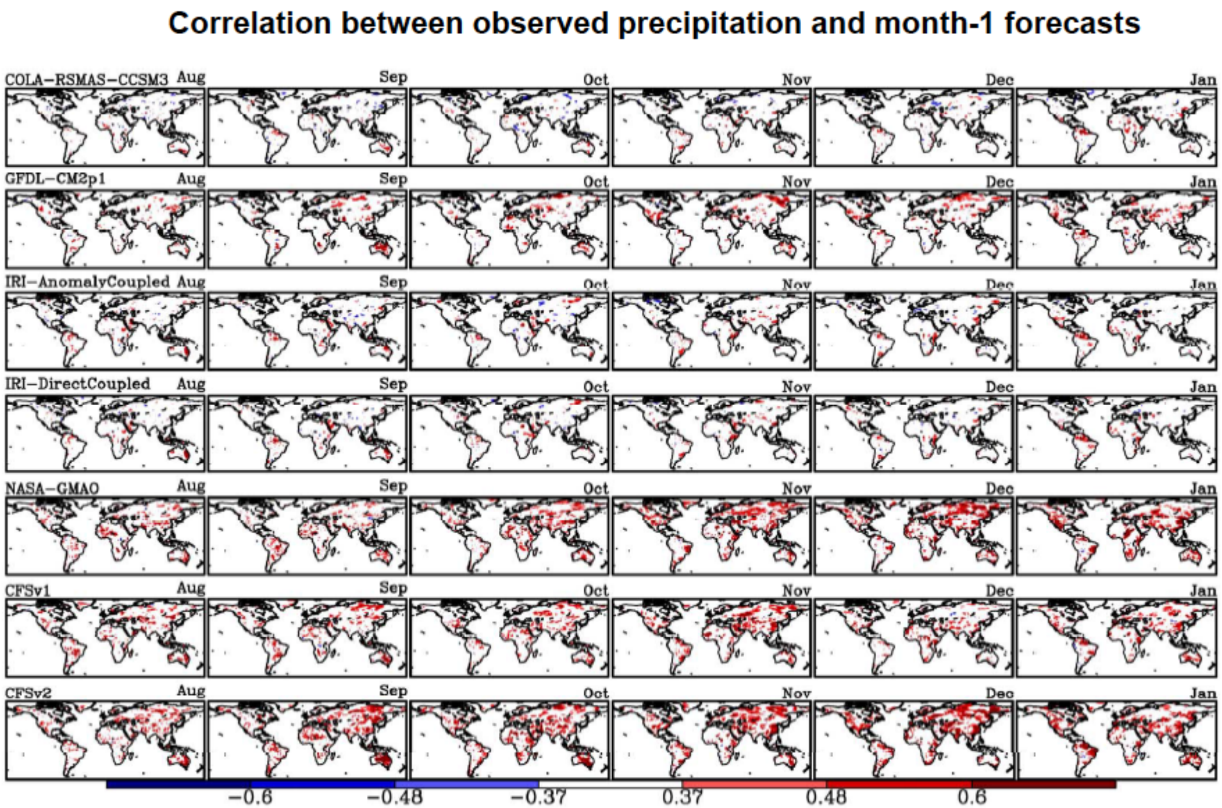
(MM)Ensemble of Nino 3.4 SST



Typical model AC score for SST
30 year hindcast

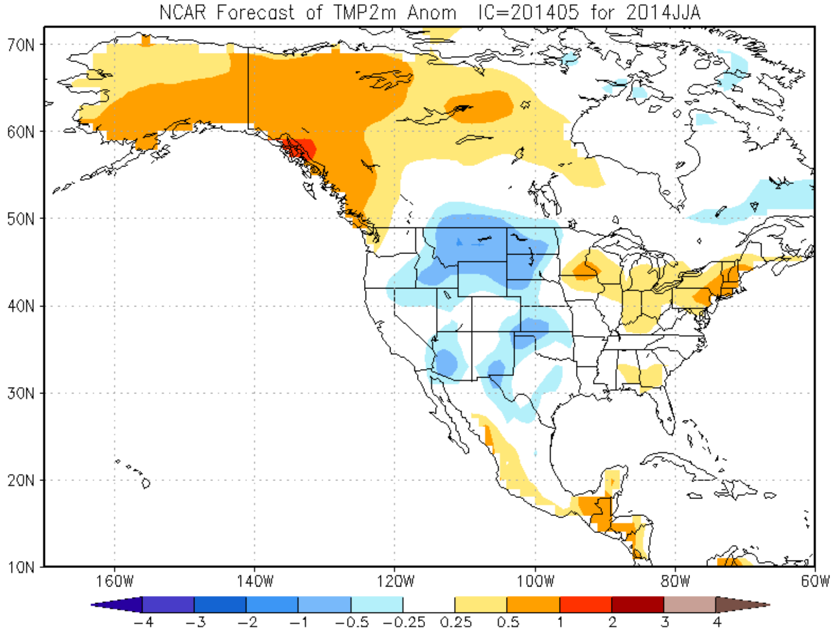


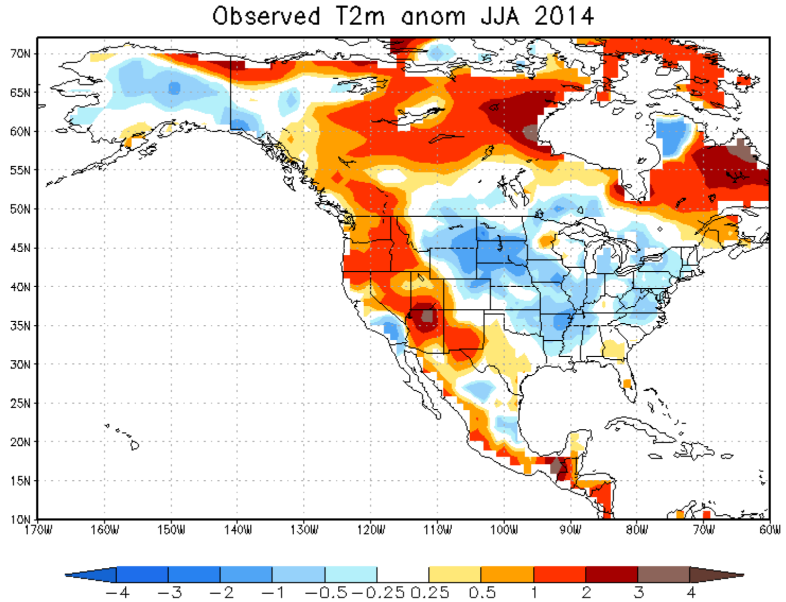
30 year hindcast results Continental Precipitation



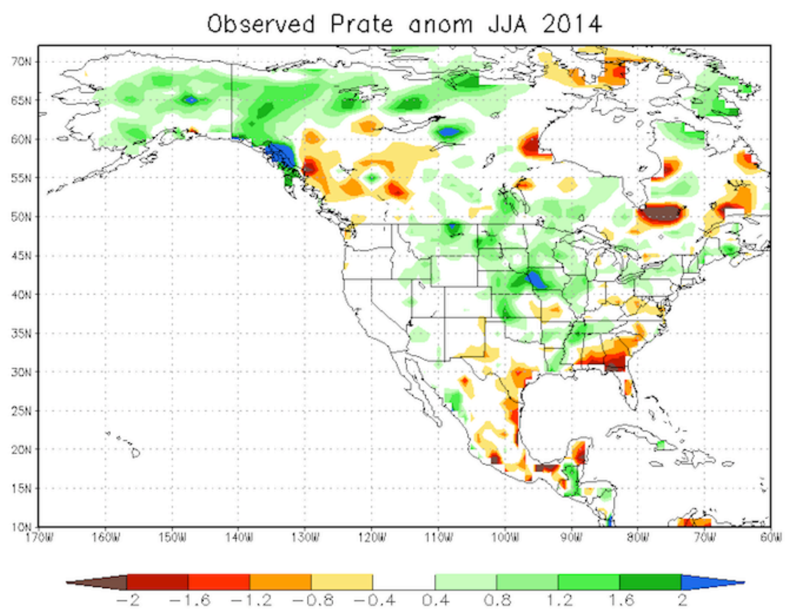
PRINCETON UNIVERSITY

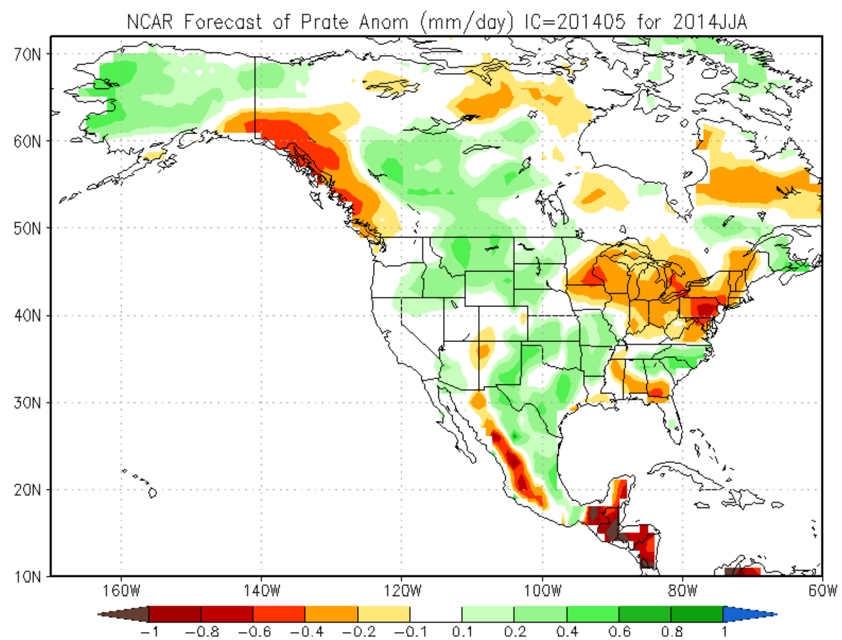
Tropical SST is NOT a Seasonal Forecast



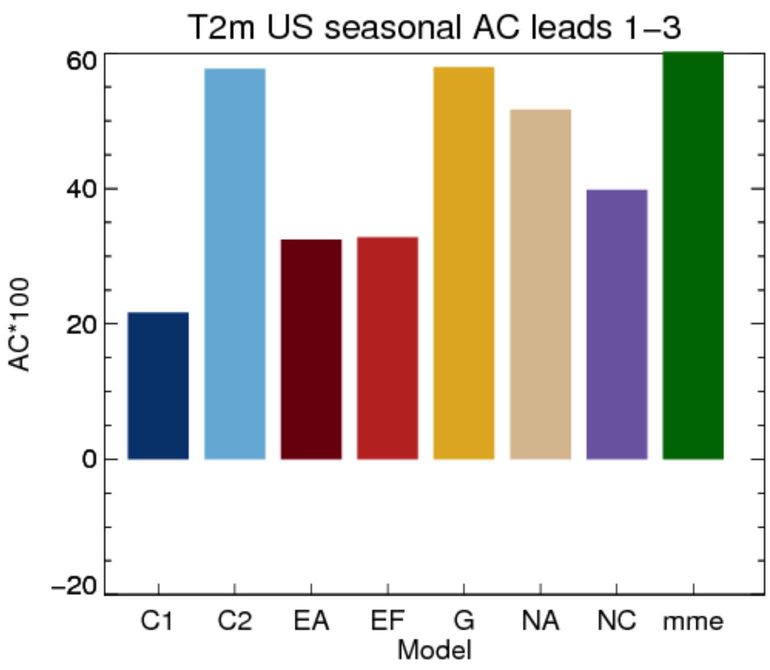


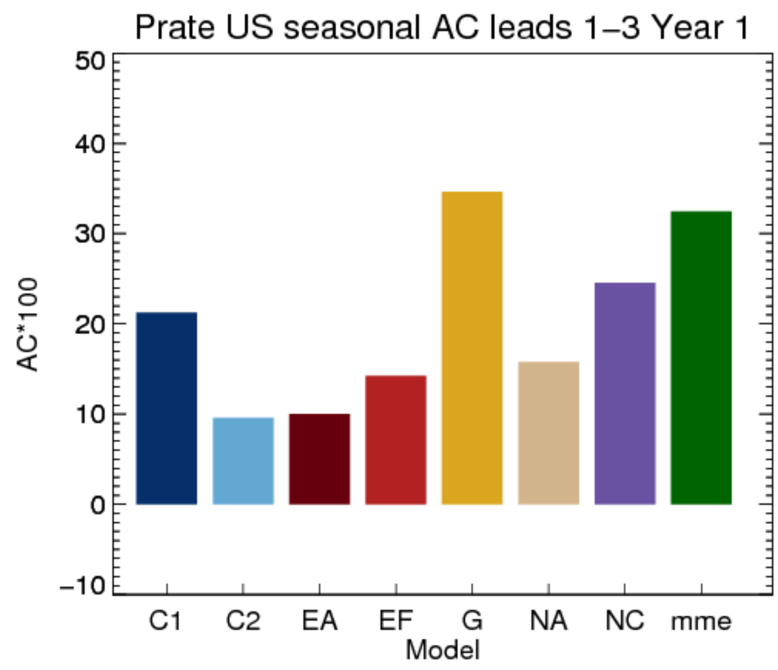
Tropical SST is NOT a Seasonal Forecast





Real time Verification of AC Seasonal T_{2m} and P_{rate}





Problems beyond medium range

- Bias in coupled models is large
- Forecasts must be bias corrected and (re)calibrated
- Number of samples of independent events is small. Hindcasts used for calibration and skill determination are woefully inadequate
- Even though the forecasts are probabilistic- pressure for deterministic verification and extremes

Final Comments

- At medium range out to a month good probabilistic forecasting systems (reliable and good resolution) are usually adequate for use. (Exception is extreme events)
- There are sufficient samples at this range to make use of verification for model development
- At monthly to seasonal range coupled models are not yet good enough.
- Bias is problematic and there are too few samples for verification and development to synergize.

"You can't always get what you want
But if you try sometimes you just might find"

"You get what you need"…The Rolling Stones

or

"You get what you can"….Medium-seasonal Verification

The End