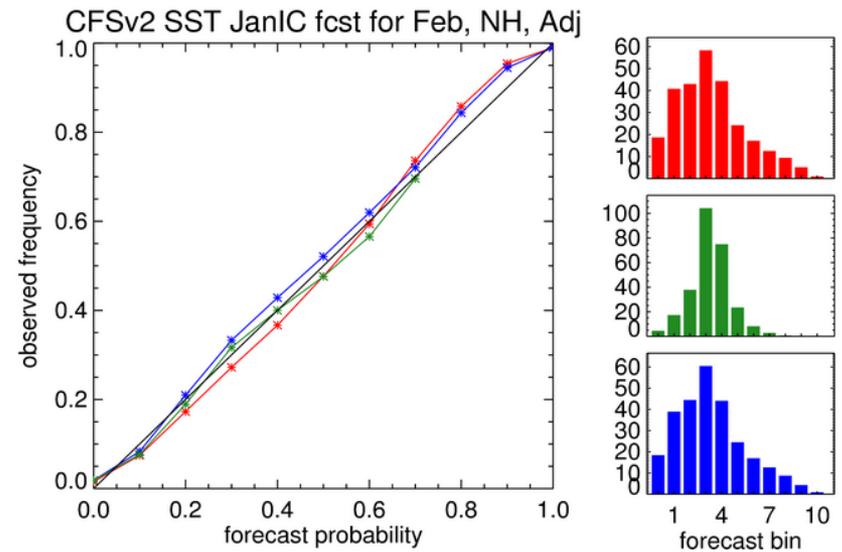
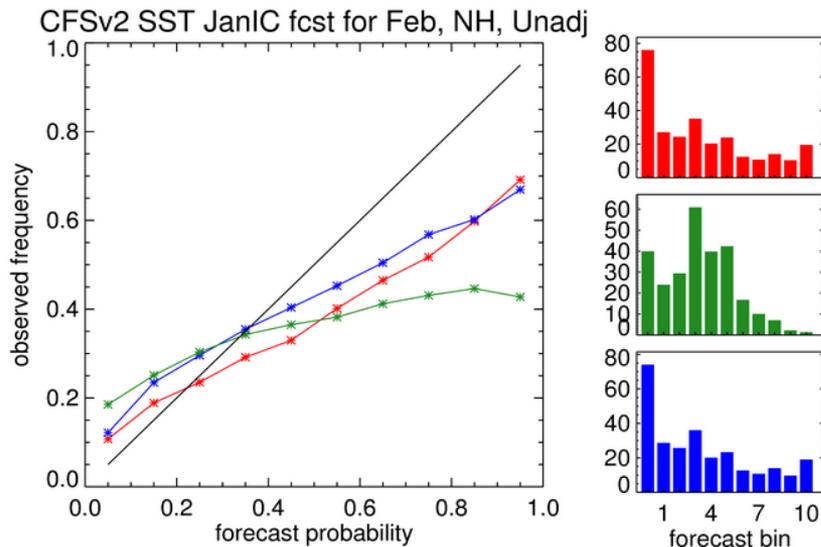


# Improving probabilities for seasonal prediction based on the North American Multi Model Ensemble (NMME)

7th NOAA Test Bed and Proving Ground (TBPG) workshop  
April 5-6, 2016, College Park, MD



Transition plan for MAPP/CTB Funded Proposal:

*Improved probabilistic forecast products for the NMME seasonal forecast system*

Anthony Barnston (PI), Huug van den Dool, Emily Becker, Michael Tippett, Shuhua Li

The project aims to improve the NMME probabilistic forecasts by addressing systematic biases in both forecast anomalies and categorical probabilities. The corrected NMME forecasts will have improved reliability and accuracy. The improvements will come about due to (1) xyz (developed at IRI), and (2) [refinements to local probability anomalies \(developed at CPC\)](#).

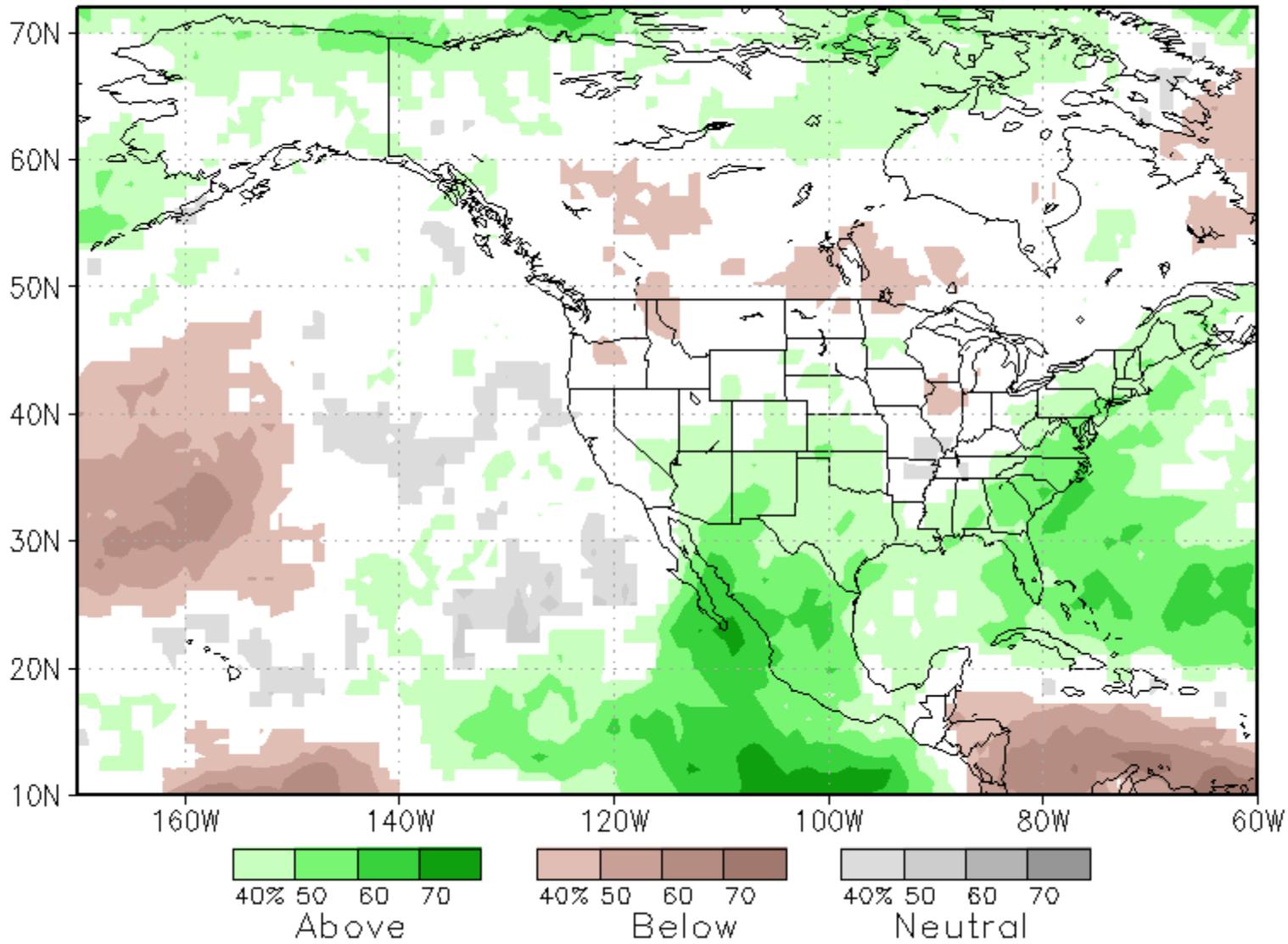
Huug van den Dool, Emily Becker  
Li-Chuan Chen and Qin Zhang

*The Probability Anomaly Correlation,  
the PAC, applied to NMME.*

*What is NMME???*

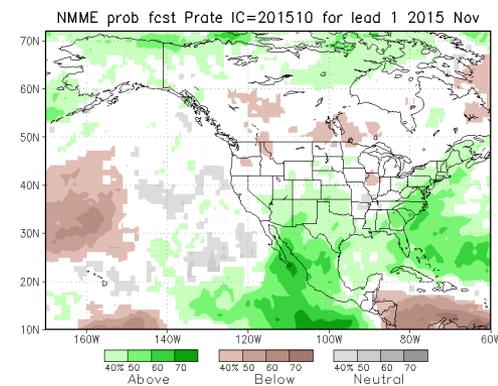
Model	Hindcast Period	No. of Members	Arrangement of Members	Lead (month)	Model resolution (atmos)	Model resolution (ocean)	Reference
<b>Active</b>							
NCEP/CFSv2	1982-2010	24 (28)	4 members (0, 6, 12, 18z) every 5 <sup>th</sup> day	0-9	T126L64	MOM4L40 .25deg Eq	Saha et al (2010)
GFDL/CM2.1	1982-2010	10	All 1 <sup>st</sup> of the month 0Z	0-11	2x2.5degL24	MOM4L50 .3deg Eq	Delworth (2006)
GFDL/CM2.5 (FLOR)	1982-present	24	All 1 <sup>st</sup> of the month 0Z	0-11	C18L32 (50km)	MOM5 L50 0.30 deg Eq 1degPolar1.5	Vecchi et al (2014)
CMC1-CanCM3	1981-2010	10	All 1 <sup>st</sup> of the month 0Z	0-11	CanAM3 T63L31	CanOM4L40 .94deg Eq	Merryfield et al (2013)
CMC1-CanCM4	1981-2010	10	All 1 <sup>st</sup> of the month 0Z	0-11	CanAM4 T63L35	CanOM4L40 .94deg Eq	Merryfield et al (2013)
NCAR/CCSM4	1982-2010	10	All 1 <sup>st</sup> of the month 0Z	0-11	0.9x1.25degL26	POPL60 .25deg Eq	Kirtman et al. (in prep)
NASA/GEOS5	1981-2010	11	4 mems every 5 days; 7 mems on last day of last month	0-9	1x1.25 deg L72	MOM4L40 .25deg Eq	Vernieres et al (2012)
<b>Retired</b>							
NCEP/CFSv1	1982-2009	15	1 <sup>st</sup> 0Z +/-2 days, 21 <sup>st</sup> 0z +/-2d, 11 <sup>th</sup> 0z +/-2d	0-8	T62L64	MOM3L40 0.30 deq Eq	Saha et al (2006)
NCAR/CCSM3	1982-2010	6	All 1 <sup>st</sup> of the month 0Z	0-11	T85L26	POPL42 0.3deg Eq	Kirtman and Min2009)
IRI-ECHAM4f	1982-2010	12	All 1 <sup>st</sup> of the month 0Z	0-7	T42L19	MOM3L25(1.5x0.5)	DeWitt (2005)
IRI-ECHAM4a	1982-2010	12	All 1 <sup>st</sup> of the month 0Z	0-7	T42L19	MOM3L25 (1.5x0.5)	DeWitt (2005)
<b>Planned</b>							
NCAR/CESM1	1982-2010	10	All 1 <sup>st</sup> of the month 0Z	0-11	0.9x1.25degL30	POPL60 .25deg Eq	Tribbia et al.

# NMME prob fcst Prate IC=201510 for lead 1 2015 Nov

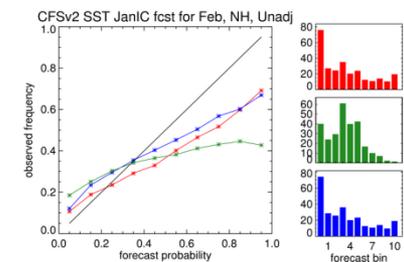


THIS IS WHAT WE DO IN  
REAL TIME

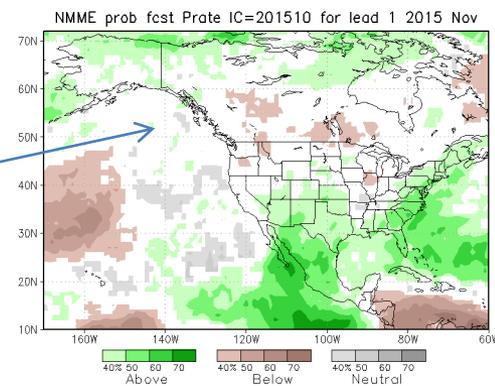
# Even before the PAC adjustment



- Determine tercile limits at each gridpoint, based on 1982-2010 hindcasts appropriate for target month/season and lead (SST, T2m, prate)
- Apply the count method to a new independent forecast. Each ensemble member is mapped onto two 0s and one 1.
- Add up all counts for each model, then across all models in use in NMME. Express as % for each of three classes.
- Please note implicit correction of mean and pdf.
- Please note how models are added together into overall NMME probabilities.
- Remember: this is BEFORE the PAC related adjustment. The reference we have to beat is already cleaned up and scoring well.

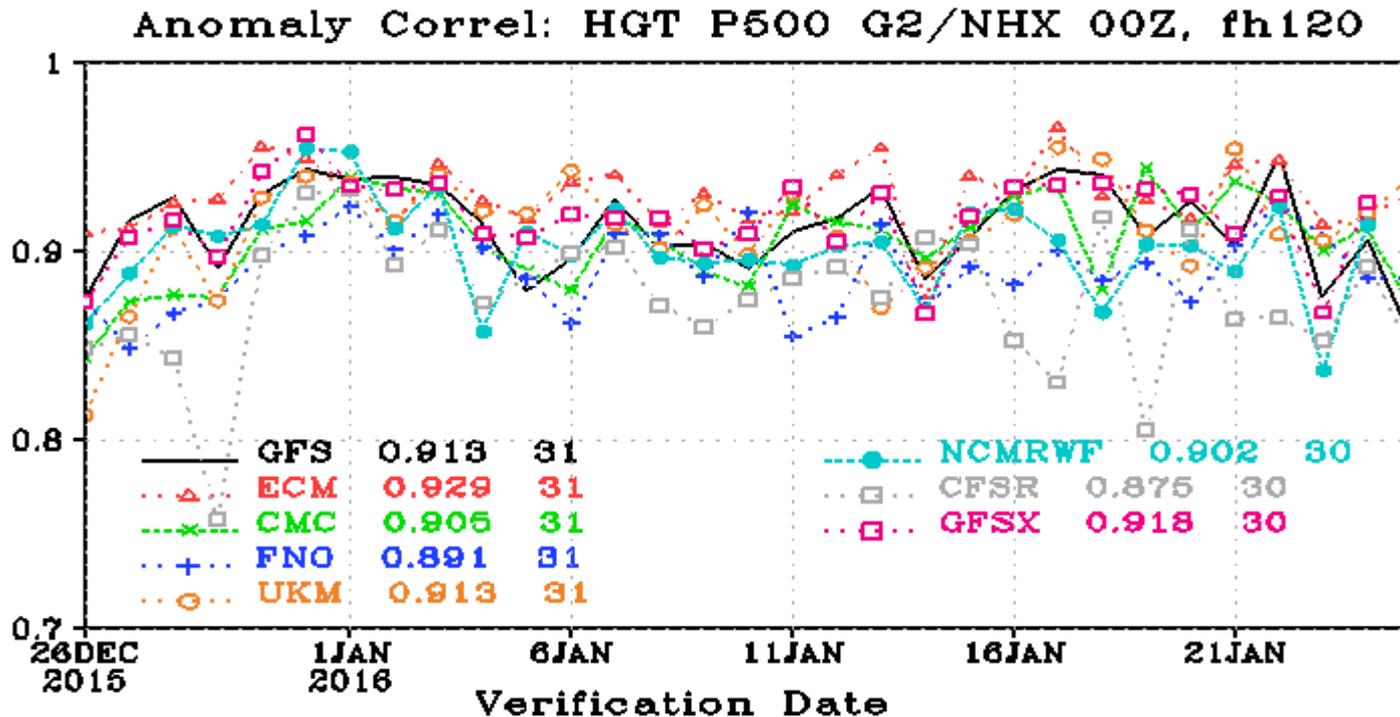


# Note also



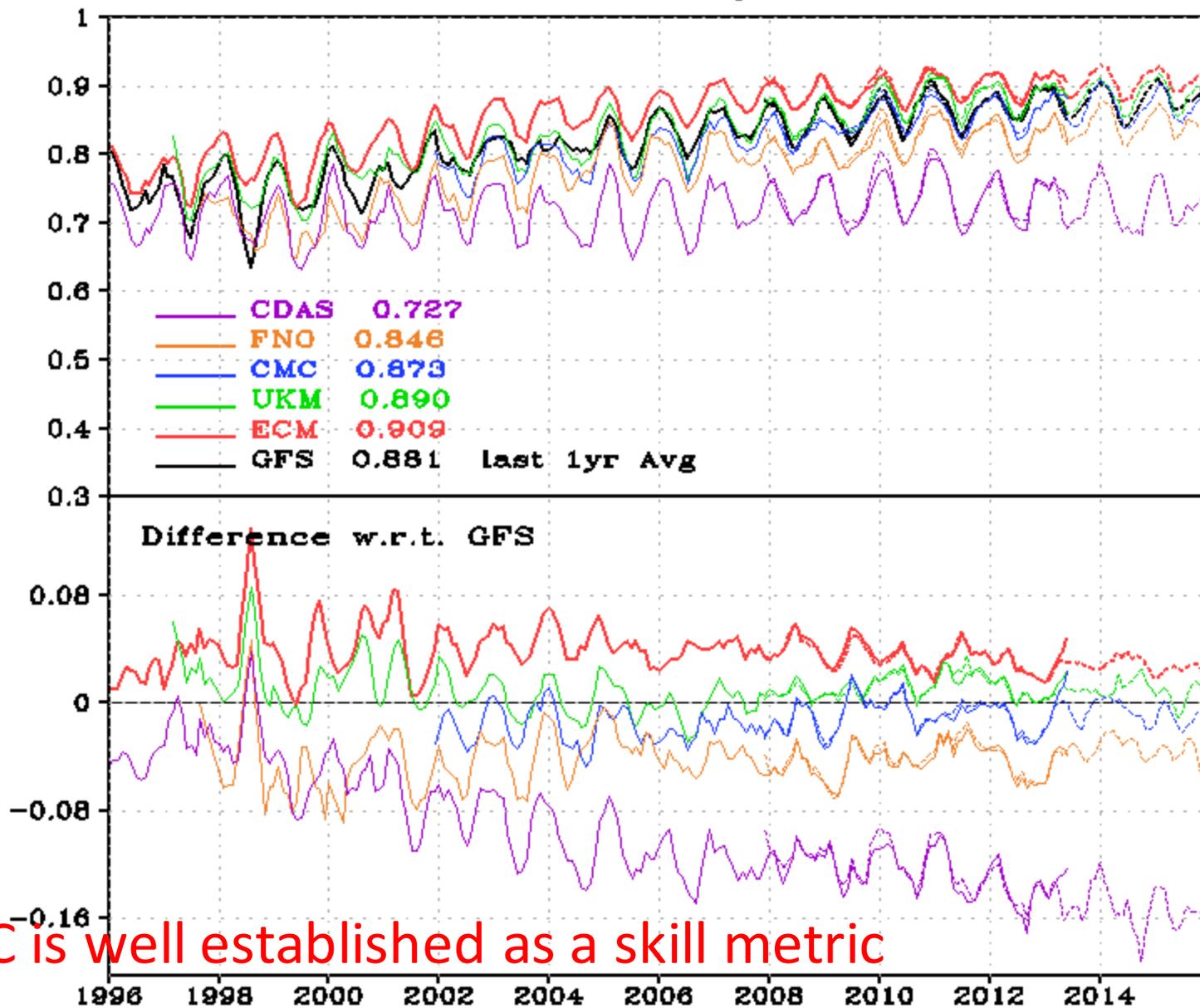
- This Slide is really about predictability, created without any reference to observations, independent of any skill models may have (i.e. as established by a verification against observations)
- The notion probability anomaly (PA), departure from  $1/3^{\text{rd}}$ .
- The count method has a large round-off problem.
- Name of the game: The need to smooth probabilities, i.e. damp PA. Suppose we had a single low skill model with just one member.....
- We use the traditional terciles, but nothing we say depends on how many classes one uses.

# A little excursion about the AC, the anomaly correlation



The AC is well established as a skill metric

### NH HGT AC: 500hPa Day5, 3-Mon Mean



The AC is well established as a skill metric

# Hidden meaning of a correlation:

A correlation tells you by how much forecast anomalies should be damped in order to minimize the MSE. (Damp towards climatology).

One knows the answer without actually having to do the damping.

I.e. AC indicates the inherent skill one has.

{according to  $(\text{MSE}_{\text{control}} - \text{MSE}_{\text{Forecast}}) / \text{MSE}_{\text{control}}$  }.

Mean square error (MSE) is a very very basic verification attribute.

{{Damping forecast anomalies is not everybody's favorite activity.

Because ? it weakens the weather in weather maps.}}

# By extension

- The PAC damps the probability anomalies so as to minimize the Probability version of MSE, called Brier Score.

# The Brier Score

- Mean square error of a probability forecast

$$BS = \frac{1}{N} \sum_{i=1}^N (p_i - o_i)^2$$

where  $N$  is the number of realizations,  $p_i$  is the probability forecast of realization  $i$ .  $O_i$  is equal to 1 or 0 depending on whether the event (of realization  $i$ ) occurred or not.

$$PAC = (\sum_i p_i' o_i') / [ (\sum_i p_i' p_i') (\sum_i o_i' o_i') ]$$

Where ' is departure from 1/3rd .

Index  $i$  goes across time, 1 to  $N$ .

$p$  is predicted probability,

$o$  is 0 or 1 depending on the event happening or not.

# The Brier Score

- Mean square error of a probability forecast

$$BS = \frac{1}{N} \sum_{i=1}^N (p_i - o_i)^2$$

where  $N$  is the number of realizations,  $p_i$  is the probability forecast of realization  $i$ .  $O_i$  is equal to 1 or 0 depending on whether the event (of realization  $i$ ) occurred or not.

In the same way that traditional MSE can be minimized by a regression, we here attempt to minimize the BS, i.e. the MSE for probability forecasts.

The meaning/interpretation of traditional anomaly correlation (AC)

By extension the meaning/interpretation of the probability anomaly correlation (PAC)

Tool to be applied		Result1
Deterministic	AC	Forecast gets damped towards deterministic climatology=long term mean
Probabilistic	PAC	Forecast gets damped towards probabilistic climatology=(1/3rd,1/3rd,1/3rd)

Tool to be applied		Result2
Deterministic	AC	Lower MSE
Probabilistic	PAC	Lower BS

Unanswered question: To what extent should we be ruled by verification metrics?

# CFSv2 JanIC SST forecasts for February, Northern Hemisphere Hindcasts 1982-2010

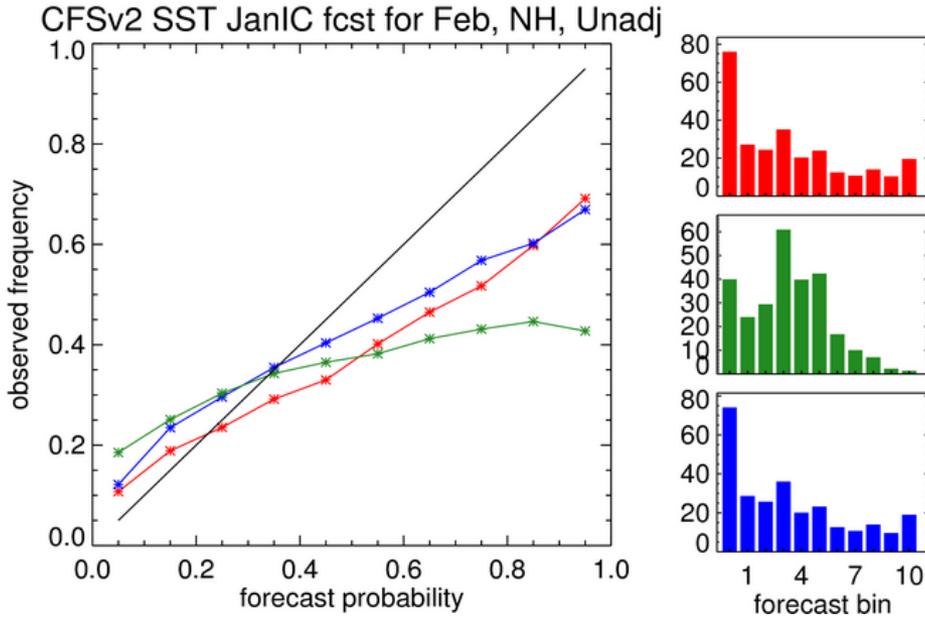
<b>Brier Score</b>	A	N	B
unadj	<b>0.187</b>	<b>0.235</b>	<b>0.201</b>
adj	<b>0.163</b>	<b>0.204</b>	<b>0.174</b>

<b>Brier Skill Sc.</b>	A	N	B
unadj	<b>0.146</b>	<b>-0.068</b>	<b>0.115</b>
adj	<b>0.257</b>	<b>0.074</b>	<b>0.232</b>

## BOTTOM LINE CONCLUSIONS

- .The PAC trick works : lower BS = higher accuracy
- .It cleans up –ve skill in N class, embarrassment avoided.
- .It tones down too bold forecasts, particularly in A&B class
- .The gain is appreciable in terms of Brier skill score.

# CFSv2 JanIC SST forecasts for February, Northern Hemisphere



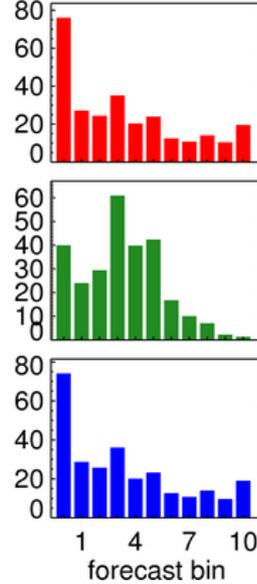
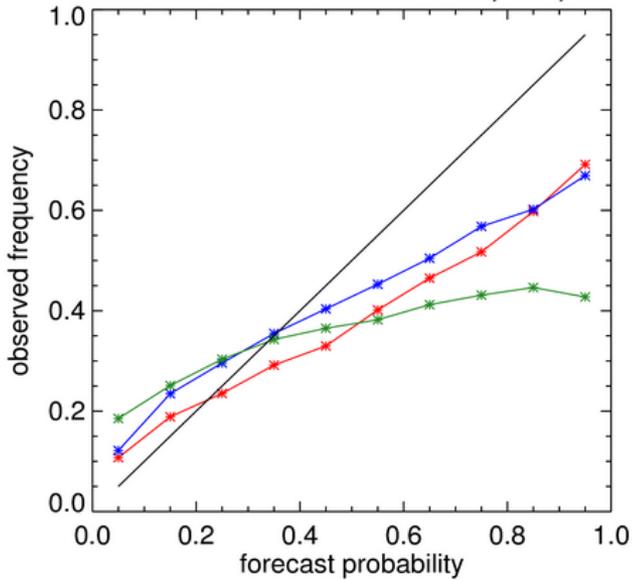
<b>Brier Score</b>	A	N	B
unadj	<b>0.187</b>	<b>0.235</b>	<b>0.201</b>
adj	<b>0.163</b>	<b>0.204</b>	<b>0.174</b>

<b>Brier Skill Sc.</b>	A	N	B
unadj	<b>0.146</b>	<b>-0.068</b>	<b>0.115</b>
adj	<b>0.257</b>	<b>0.074</b>	<b>0.232</b>

BS = Reliability minus Resolution plus Uncertainty

# CFSv2 JanIC SST forecasts for February, Northern Hemisphere

CFSv2 SST JanIC fcst for Feb, NH, Unadj

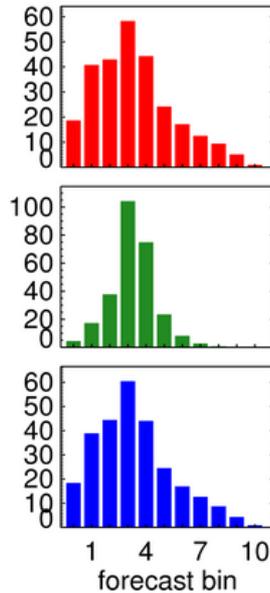
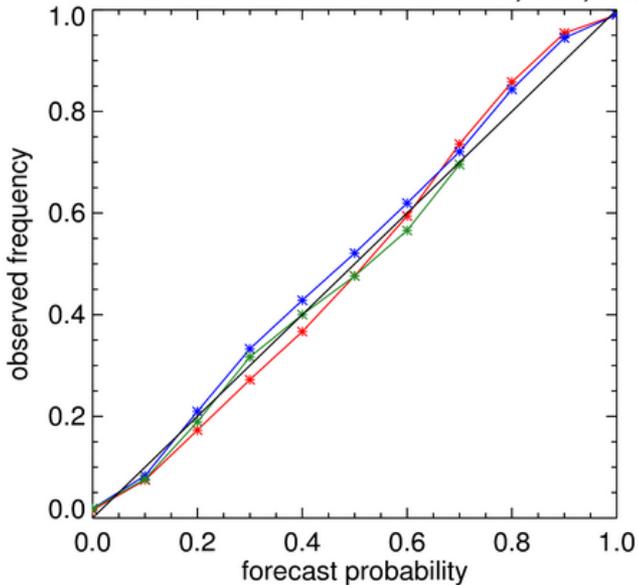


<b>Brier Score</b>	A	N	B
unadj	<b>0.187</b>	<b>0.235</b>	<b>0.201</b>
adj	<b>0.163</b>	<b>0.204</b>	<b>0.174</b>

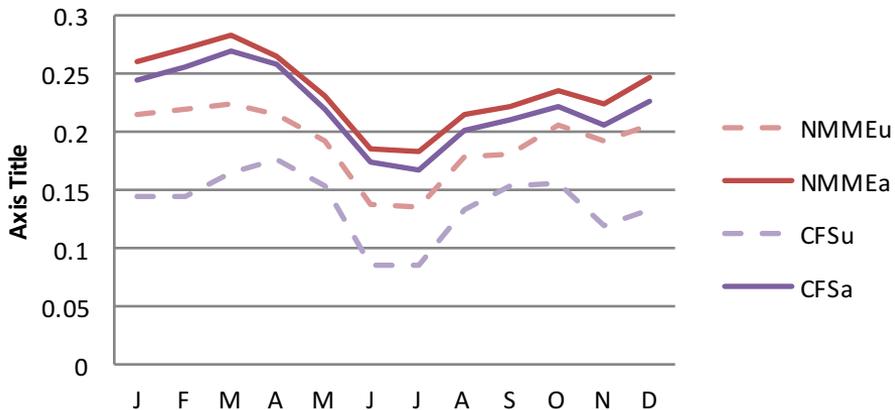
<b>Brier Skill Sc.</b>	A	N	B
unadj	<b>0.146</b>	<b>-0.068</b>	<b>0.115</b>
adj	<b>0.257</b>	<b>0.074</b>	<b>0.232</b>

BS = Reliability minus Resolution plus Uncertainty

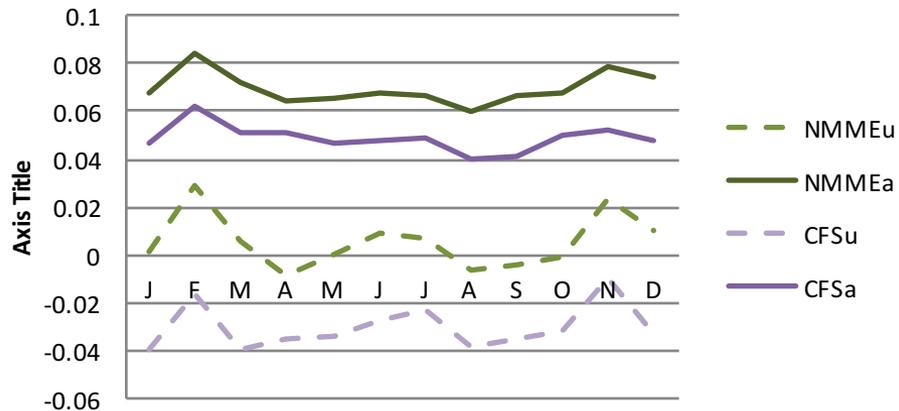
CFSv2 SST JanIC fcst for Feb, NH, Adj



**BSS Lead-1 monthly SST North. Hem.  
"above normal" category**

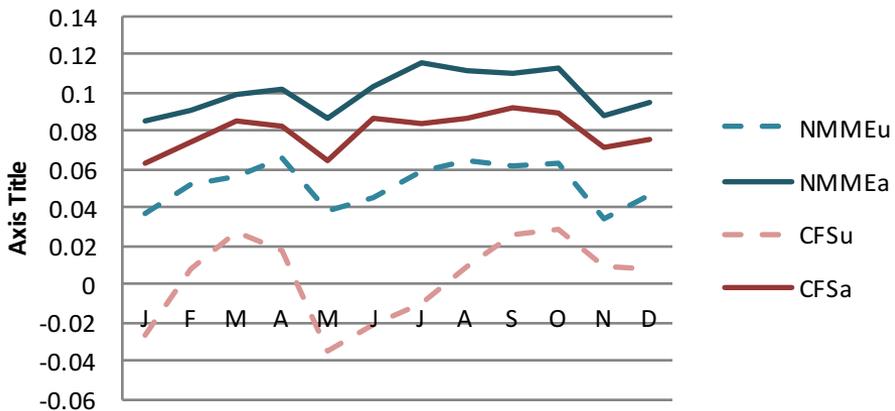


**BSS Lead-1 monthly precip rate North Amer.  
"above normal" category**

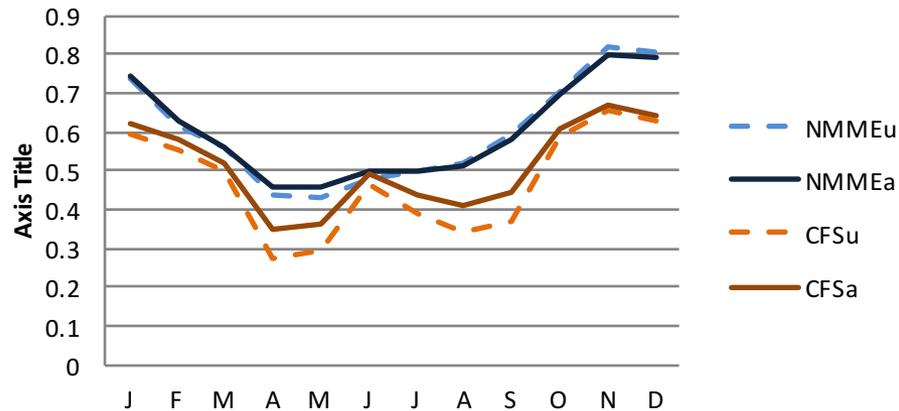


This slide added Jan, 12, 2016

**BSS Lead-1 monthly Tmp2m, North. Hemisphere  
"above normal" category**

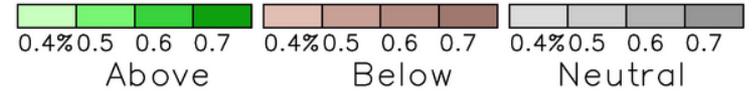
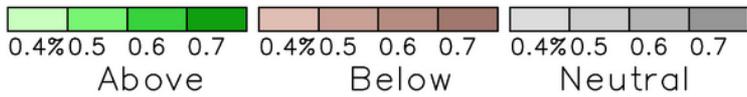
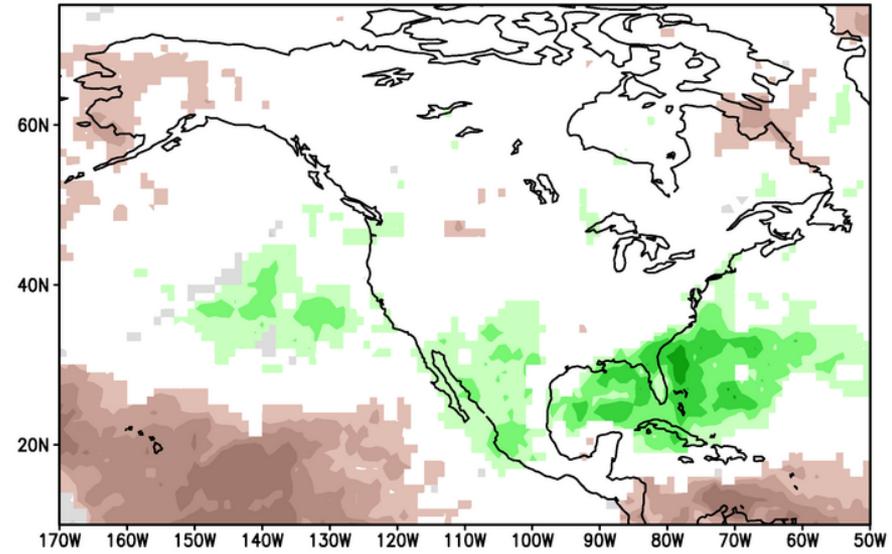
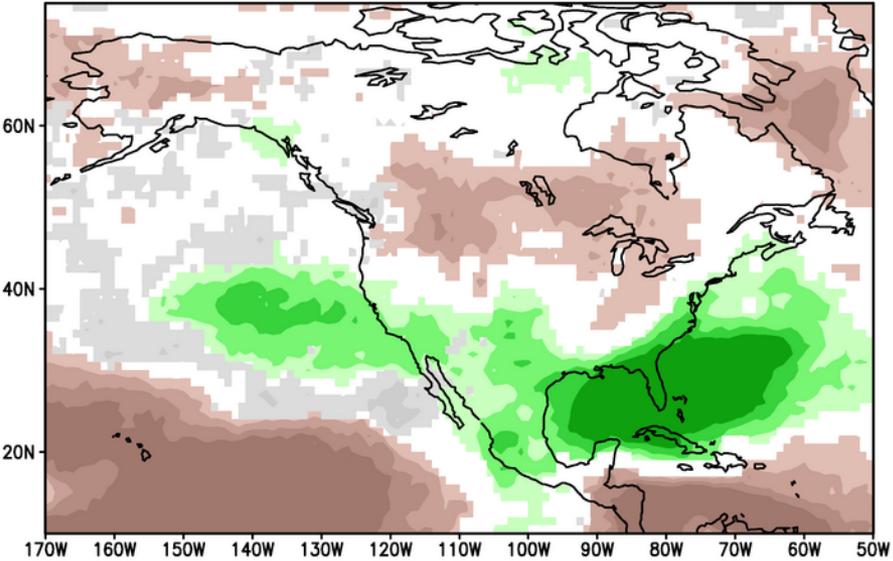


**BSS Lead-1 monthly SST, Nino34  
"above normal" category**



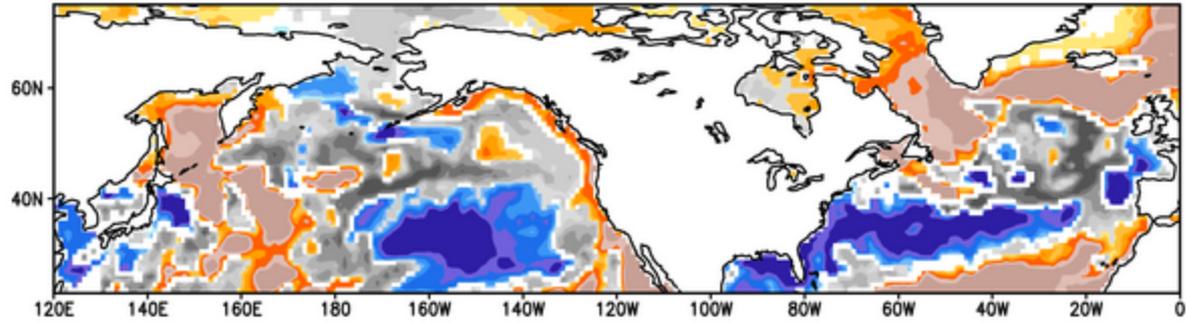
Original forecast for Jan 1983

Adjusted forecast for Jan 1983

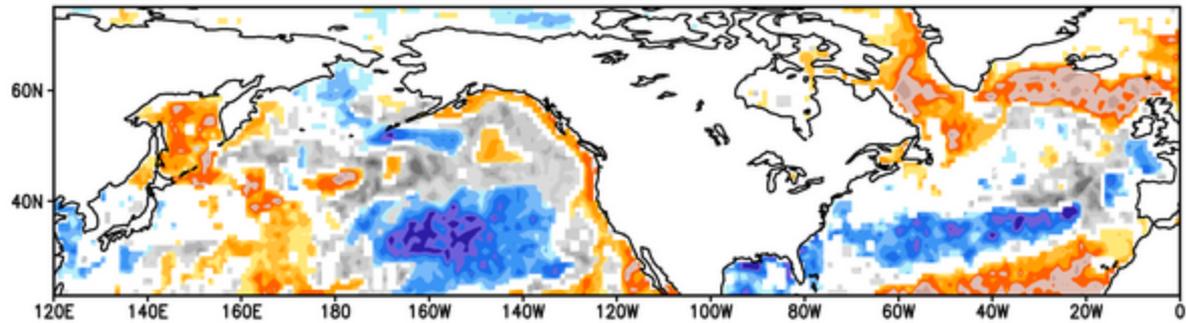


Real life example

Original forecast for Feb 2010



PAC corrected forecast for Feb 2010



Another real life example

# In conclusion

- Probability Anomaly Correlation approach yields a lower BS (as expected) for any model and for the NMME collection. This can be implemented!
- PAC is easy to understand and implement
- PAC approach has a large impact on the reliability-resolution diagram. Both reliability and resolution improve. Improvement BSS is very good.
- PAC has an interesting outside the box application in verification of ENSO composites (Li-Chuan Chen). % vs %. Not % vs (0 or 1).
- An adjustment to our most beloved conclusion
- Gone live as of April 2016

# afterthoughts

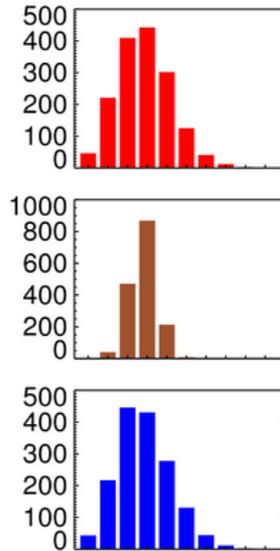
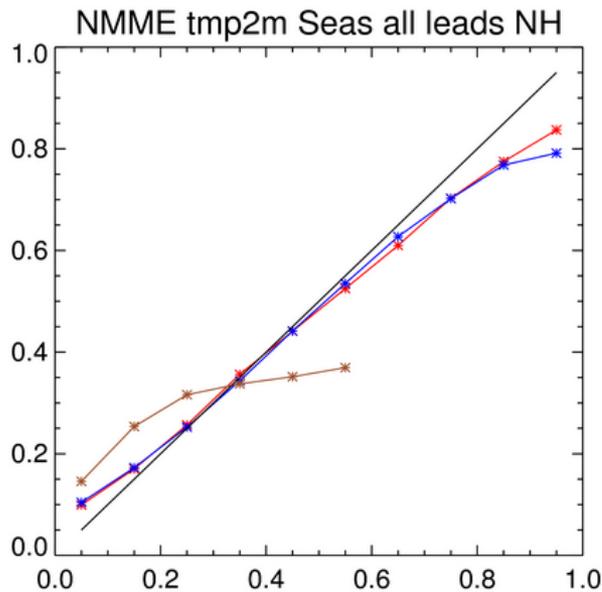
- To damp (or regress) is actually to inflate under very rare circumstances.
- From a single model to the collection of models is not always so simple.
- Details of CV have yet to be settled.
- PAC by terciles, or aggregated across three terciles. ??
- PAC is done gridpointwise. Is that OK?

# critique

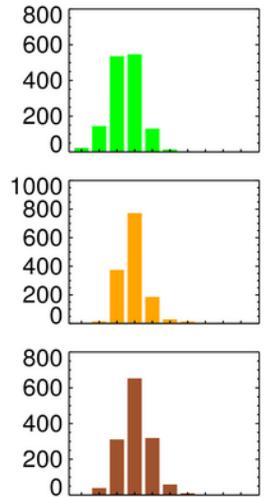
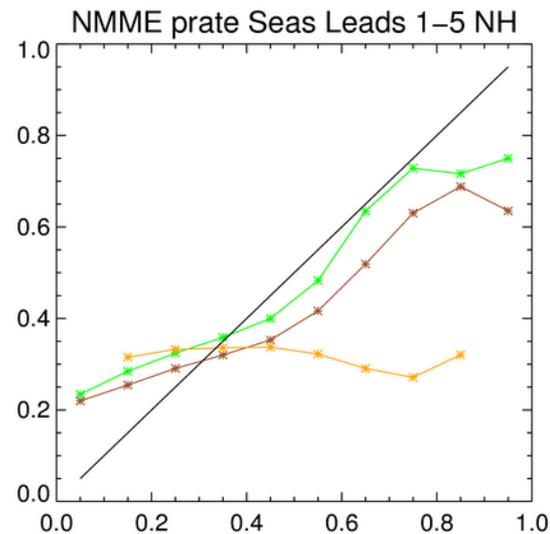
- PAC may be new, but unadvisable. You need logistic regression. Answer...true but
- We have already methods at CPC to smooth PA, like ensemble regression. Answer...true, but

XTRAs

# Managing expectations



- .How much improvement to be expected by (any) calibration?
- .Correctibility? Inherent skill  $\geq$  threshold
- .Individual Models may improve more than the NMME collection.



# Components of the Brier Score

$$BS = \frac{1}{N} \sum_{i=1}^N (p_i - o_i)^2$$

Decomposed into 3 terms for  $K$  probability classes and a sample of size  $N$ :

$$BS = \frac{1}{N} \sum_{k=1}^K n_k (p_k - \bar{o}_k)^2 - \frac{1}{N} \sum_{k=1}^K n_k (\bar{o}_k - \bar{o})^2 + \bar{o}(1 - \bar{o})$$

## reliability

If for all occasions when forecast probability  $p_k$  is predicted, the observed frequency of the event is

$$\bar{o}_k = p_k$$

then the forecast is said to be reliable. Similar to bias for a continuous variable

## resolution

The ability of the forecast to distinguish situations with distinctly different frequencies of occurrence.

## uncertainty

The variability of the observations. Maximized when the climatological frequency (*base rate*) = 0.5

Has nothing to do with forecast quality! Use the Brier skill score to overcome this problem.

The presence of the uncertainty term means that Brier Scores should not be compared on different samples.

15

BS = Reliability minus Resolution plus Uncertainty

# Thinking outside the box

- Make ENSO probability composites for a model and for observations. Then calculate a BS and PAC from it. What is unusual (almost unheard of) is that the observations are not two 0s and a 1. The observations are probabilities too. {{We can only wish reality happens more than once. It would change our entire perspective of probability forecasts.}}

**ENSO Precipitation and Temperature Forecasts in the North American Multi-Model Ensemble: Composite Analysis and Validation**

Li-Chuan Chen<sup>1,2</sup>, Huug van den Dool<sup>2</sup>, Emily Becker<sup>2,3</sup>, and Qin Zhang<sup>2</sup>

