Toward S2S Forecast Verification

Andrew W Robertson
awr@iri.columbia.edu

International Research Institute
for Climate and Society
EARTH INSTITUTE | COLUMBIA UNIVERSITY

Advanced School and Workshop on Subseasonal to Seasonal (S2S) Prediction and Application to Drought Prediction,
ICTP, Trieste, Nov 23 – Dec 4, 2015

Outline

1. What makes a good forecast? Quality, Value, Consistency

2. Skill scores: Data and hindcast requirements for S2S

3. Verification of probabilistic forecasts: sharpness & reliability

4. The S2S Verification Sub-project

What makes a “good” forecast?

Quality forecasts should correspond with what actually happens (includes skill, reliability, sharpness, discrimination, and other forecast attributes)

Value forecasts should be potentially useful (includes salience, timeliness, specificity)

Consistency forecasts should indicate what the experts really think

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, 8, 281–293.

courtesy of Simon Mason

Skill

Is one set of forecasts better than another?

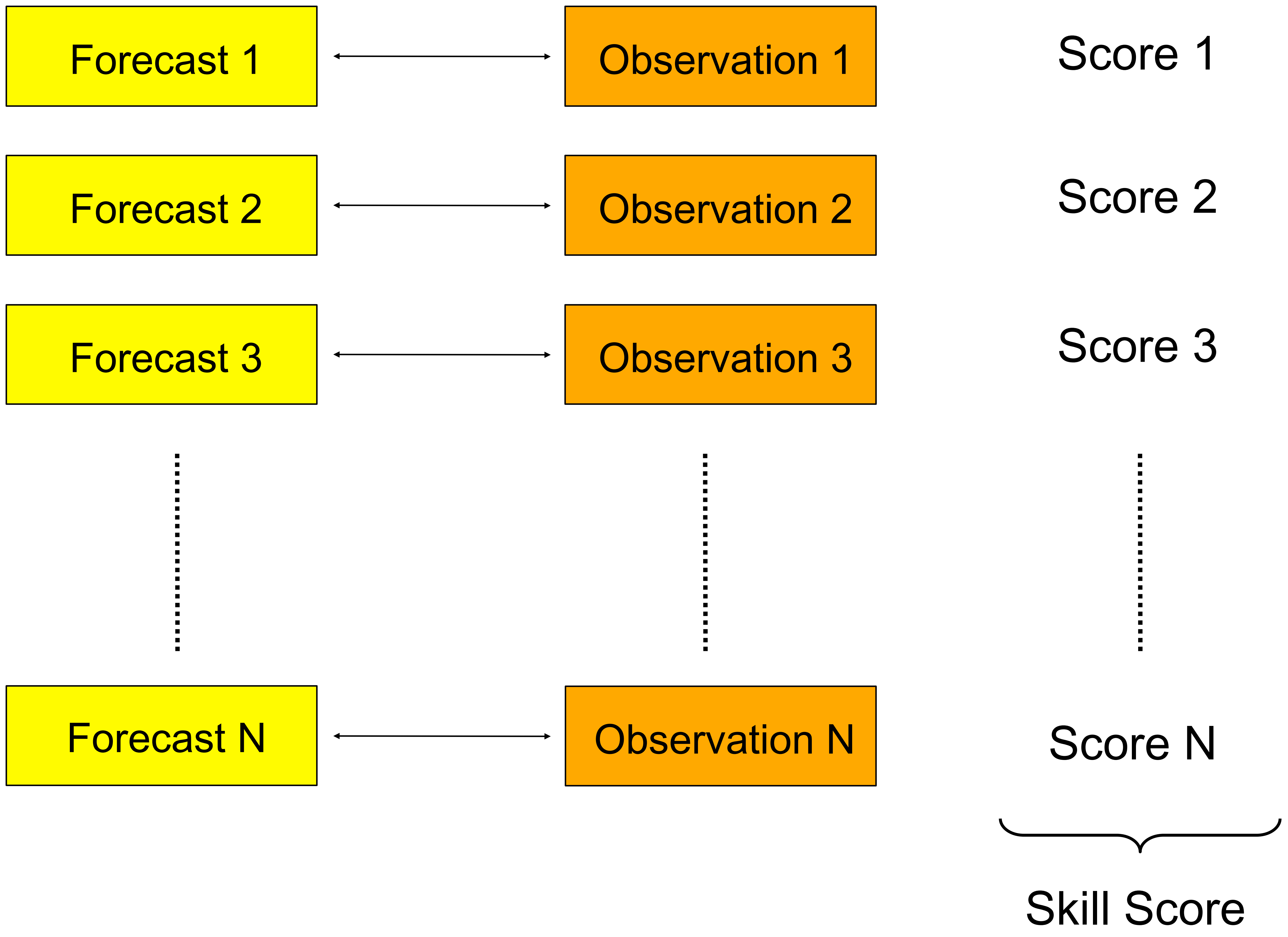
A **skill score** is used to compare the quality of one forecast strategy with that of another set (the reference set). The skill score defines the percentage improvement over the reference forecast.

Skill scores are relative measures of forecast quality.

But better in what respect? We still need to define “good” …

courtesy of Simon Mason

Skill: Assessing a set of forecasts



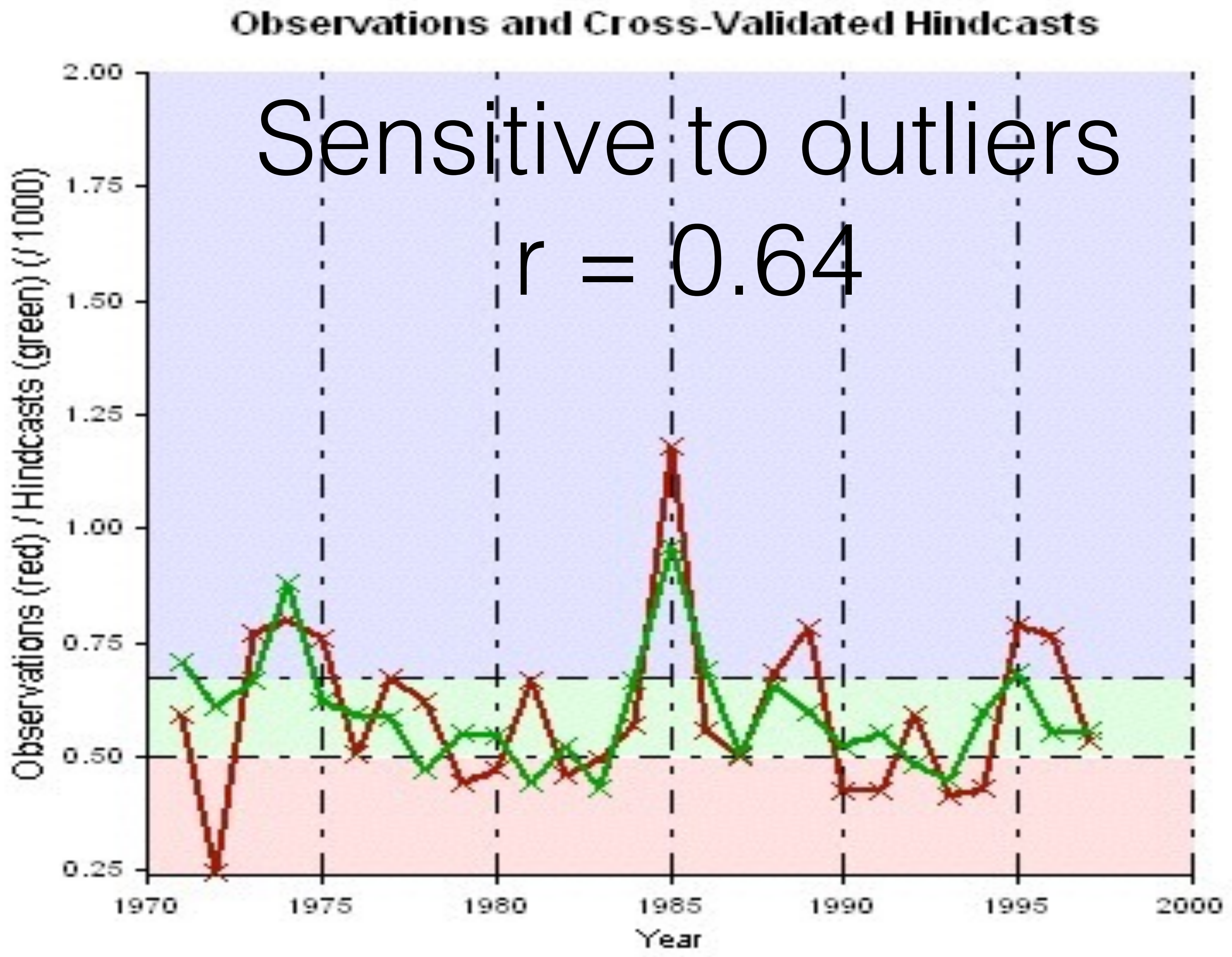
A <u>proper</u> scoring rule is designed such that quoting the true distribution as the forecast distribution is an optimal strategy in expectation

– can be based either on *real-time forecasts*,
or on hindcasts (also called re-forecasts) made retrospectively for past years

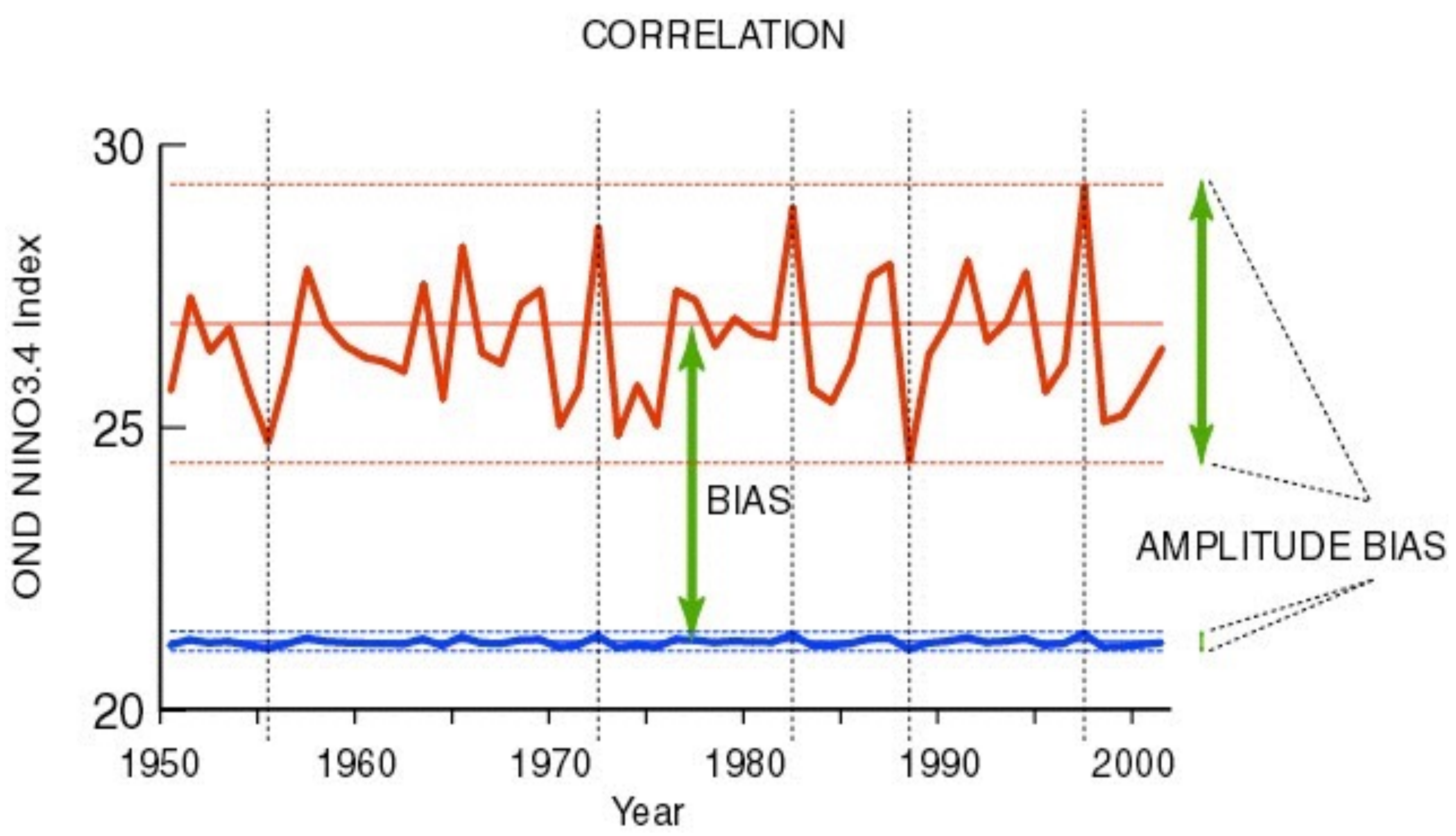
Simple deterministic score: Pearson's correlation

$$r = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i^n (x_i - \bar{x})^2}\sqrt{\sum_i^n (y_i - \bar{y})^2}}$$

- Pearson's correlation measures **association** (are increases and decreases in the forecasts associated with increases and decreases in the observations?).
- It does not measure accuracy.
- When squared, it tells us how much of the variance of the observations is correctly forecast.



courtesy of Simon Mason



Pearson's correlation measures *association* (are increases and decreases in the forecasts associated with increases and decreases in the observations?).

It does *not* measure accuracy!

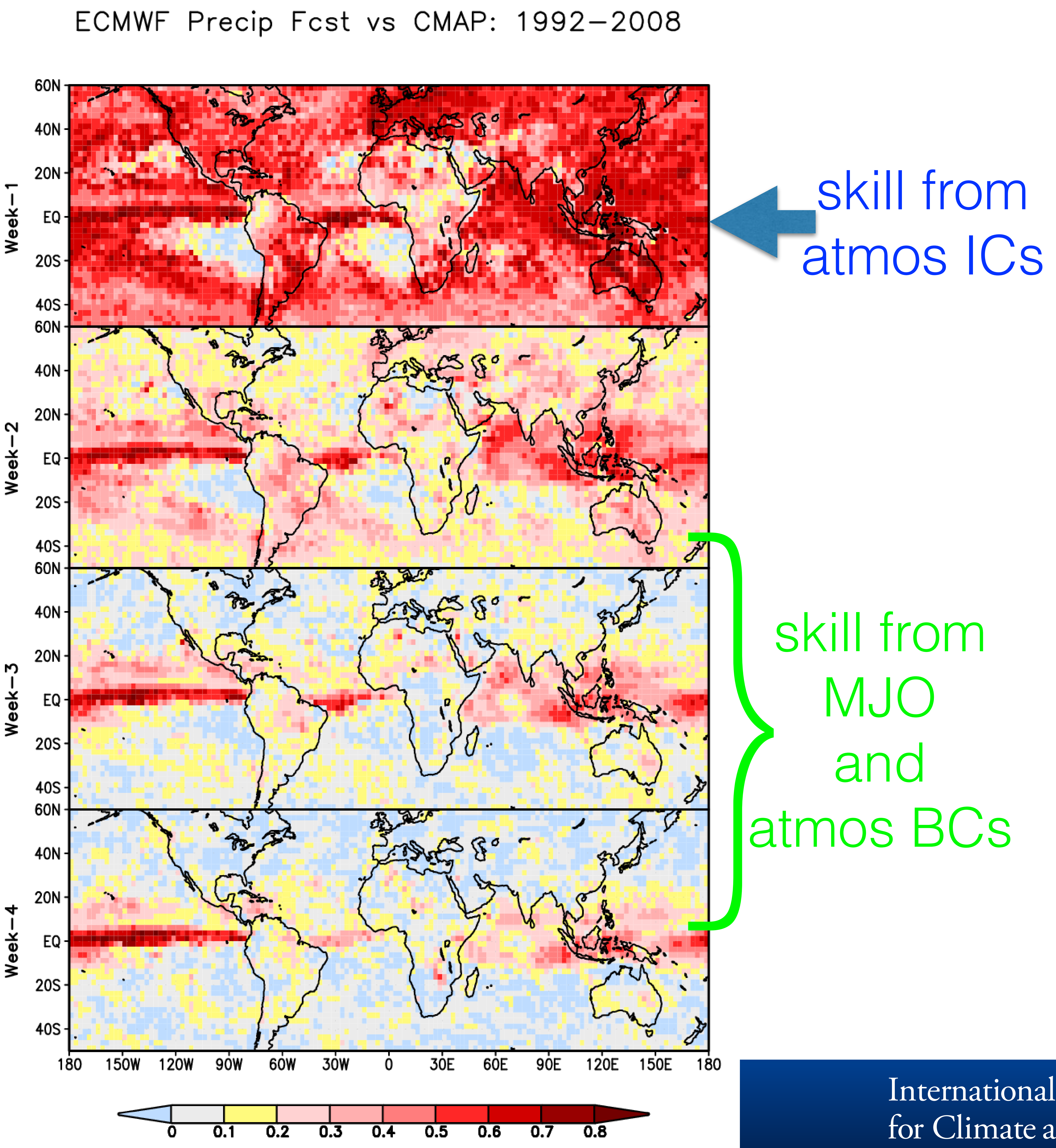
Sub-seasonal example:

ECMWF Sub-monthly forecast skill

Weekly average precip

Jun–Aug anomaly correlation skill

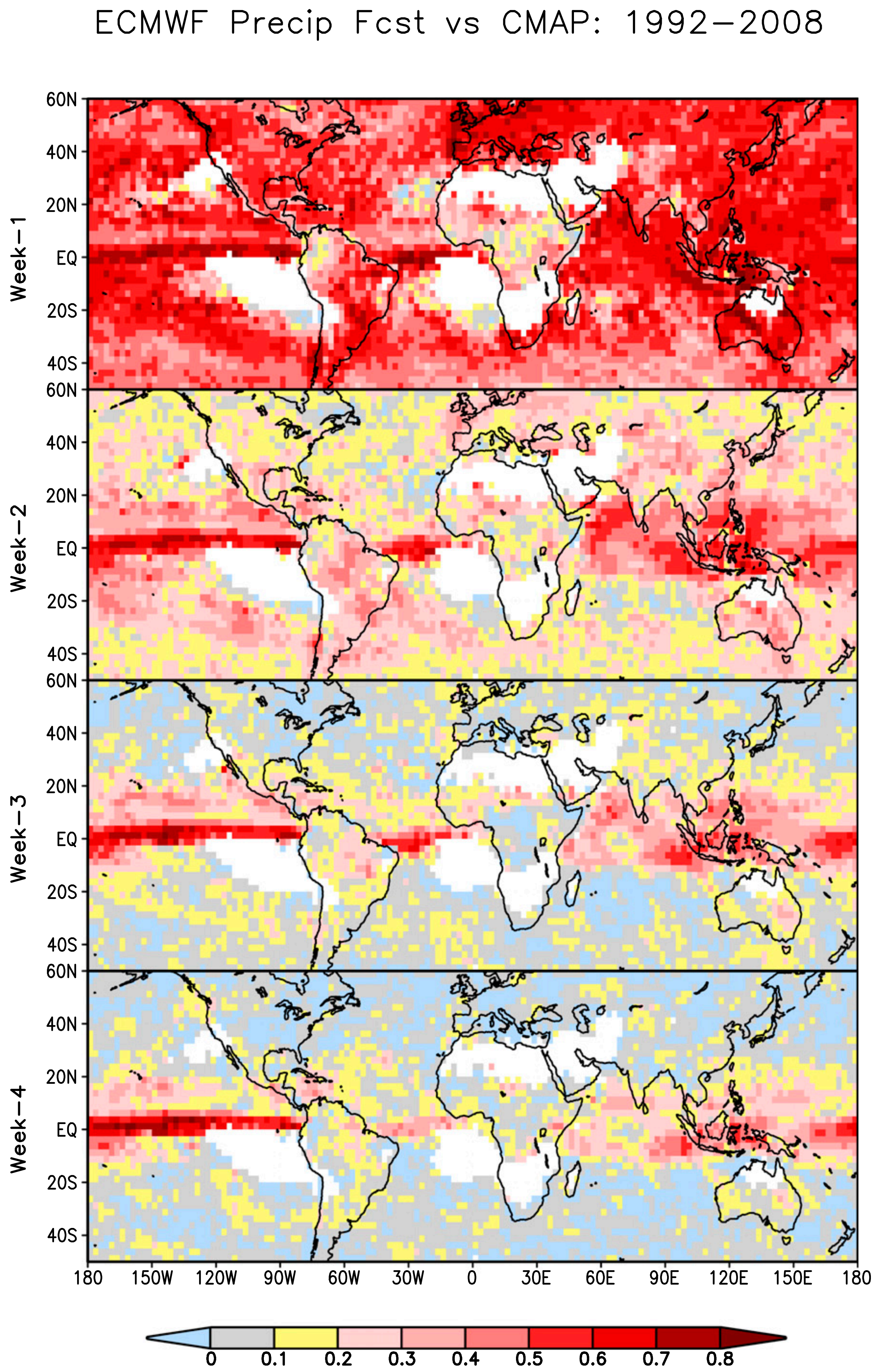
Lead-dependent climos subtracted



Li and Robertson (2015)

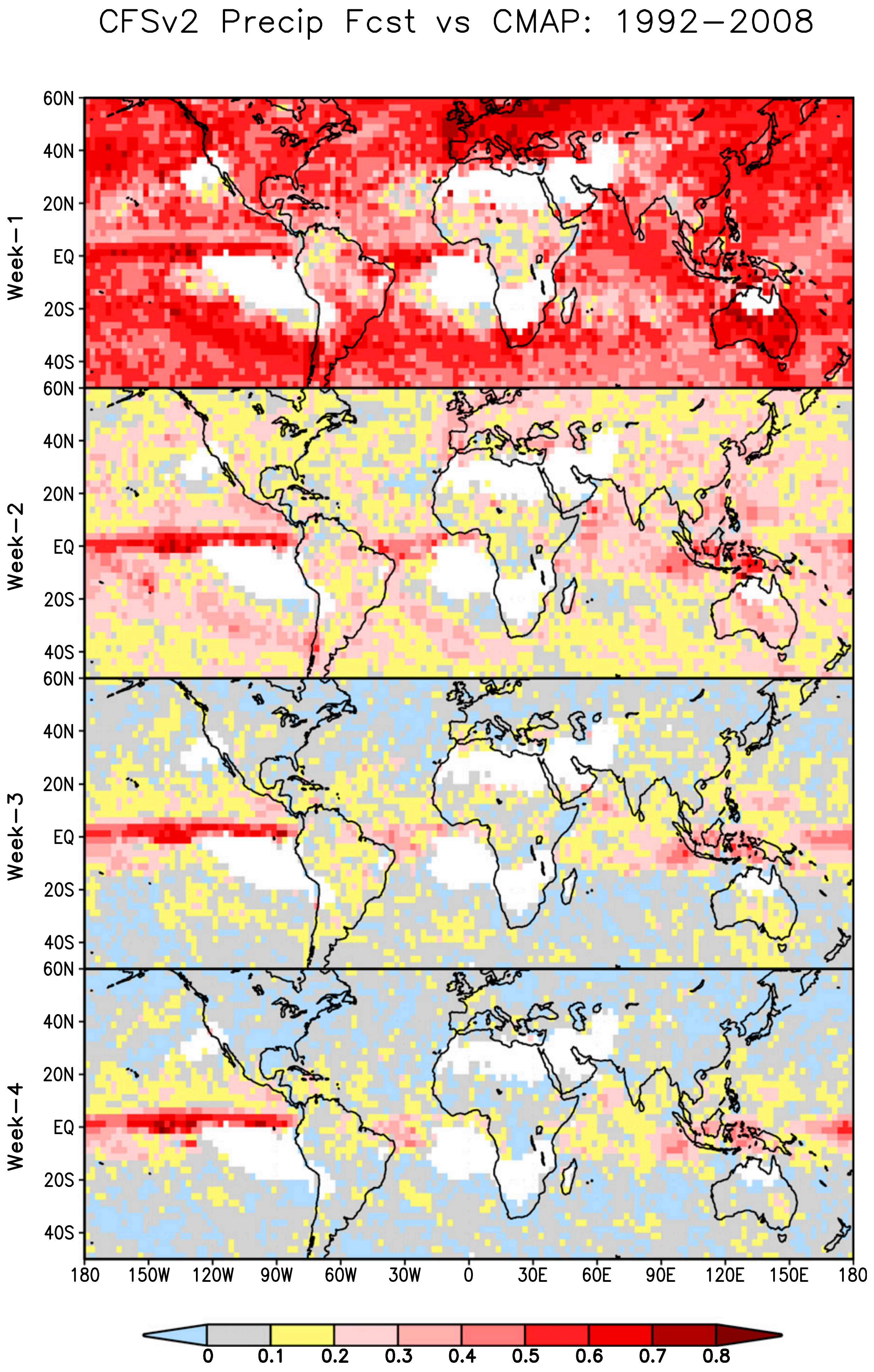
Anomaly correlation skill of weekly precipitation

ECMWF



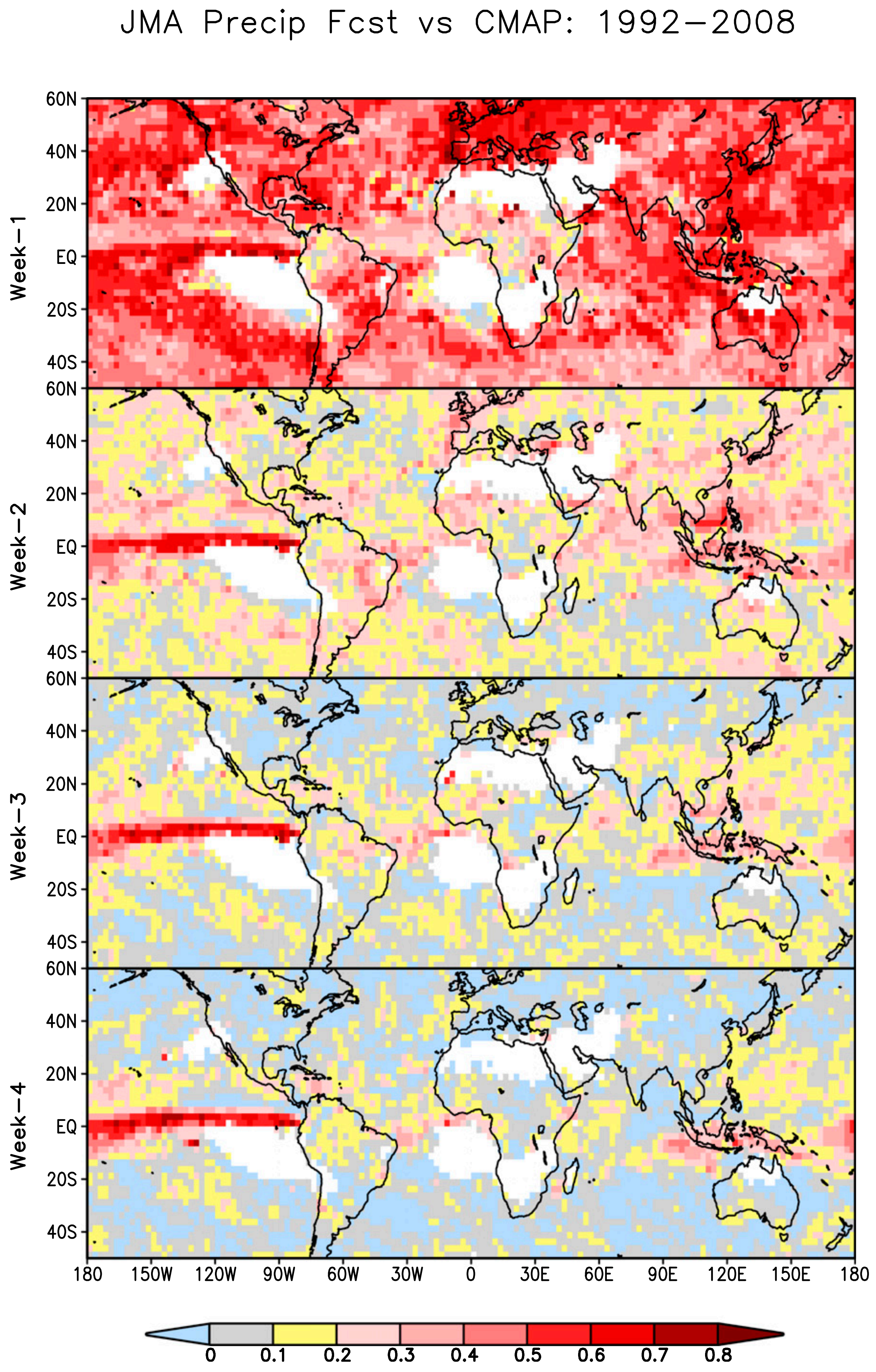
T399/255, coupled after day 10

CFSv2



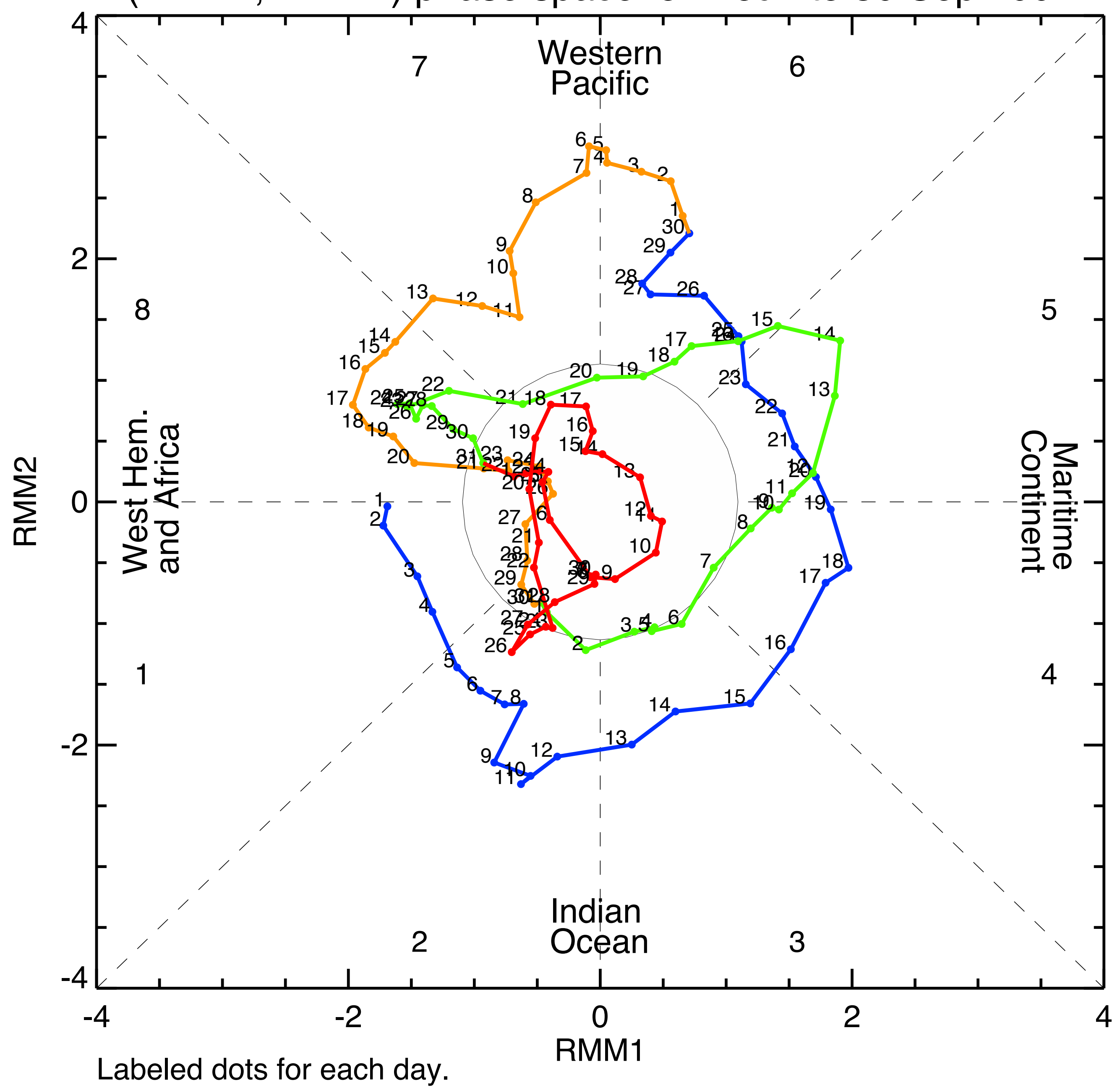
T126, coupled

JMA

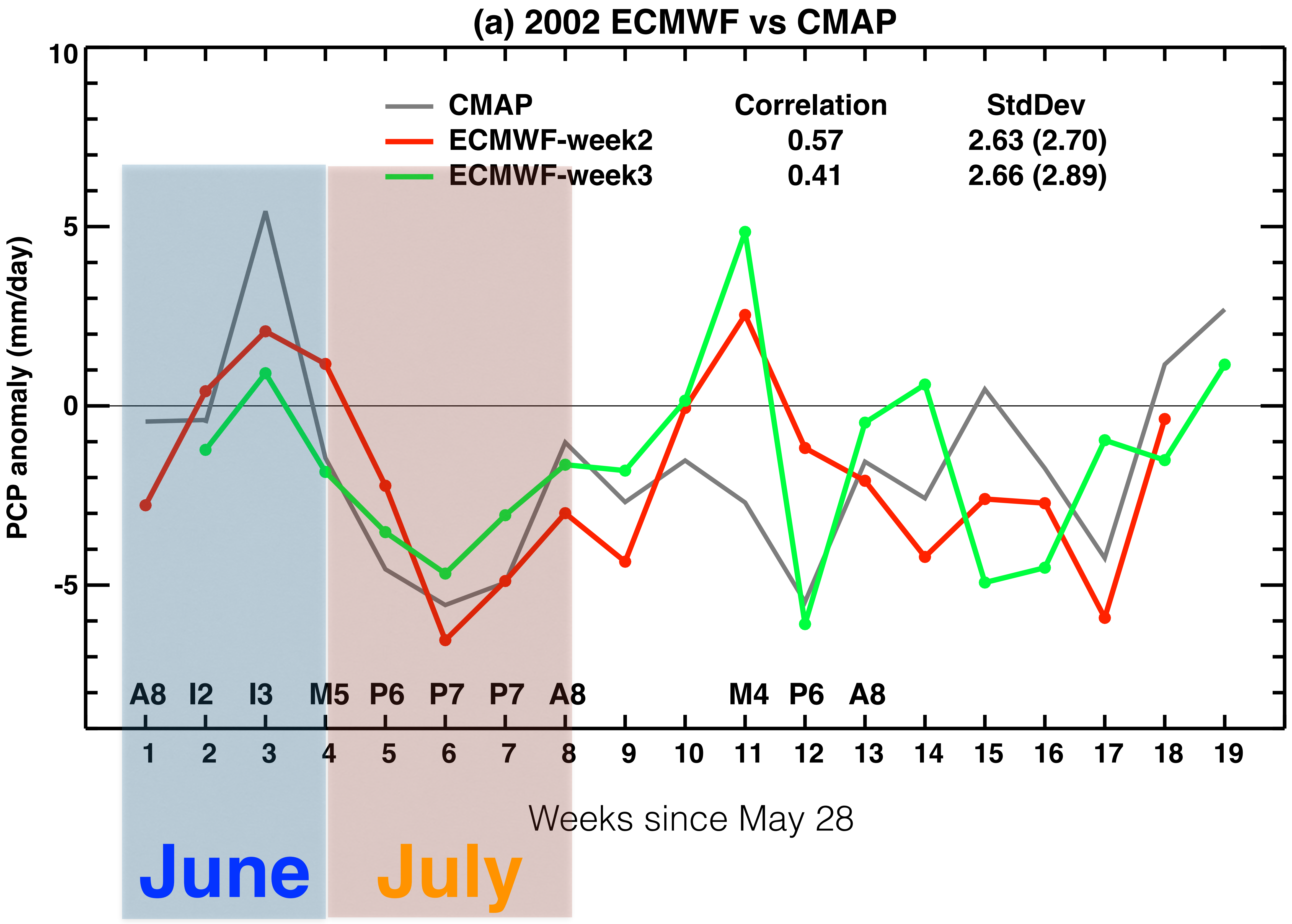


T159, persisted SST

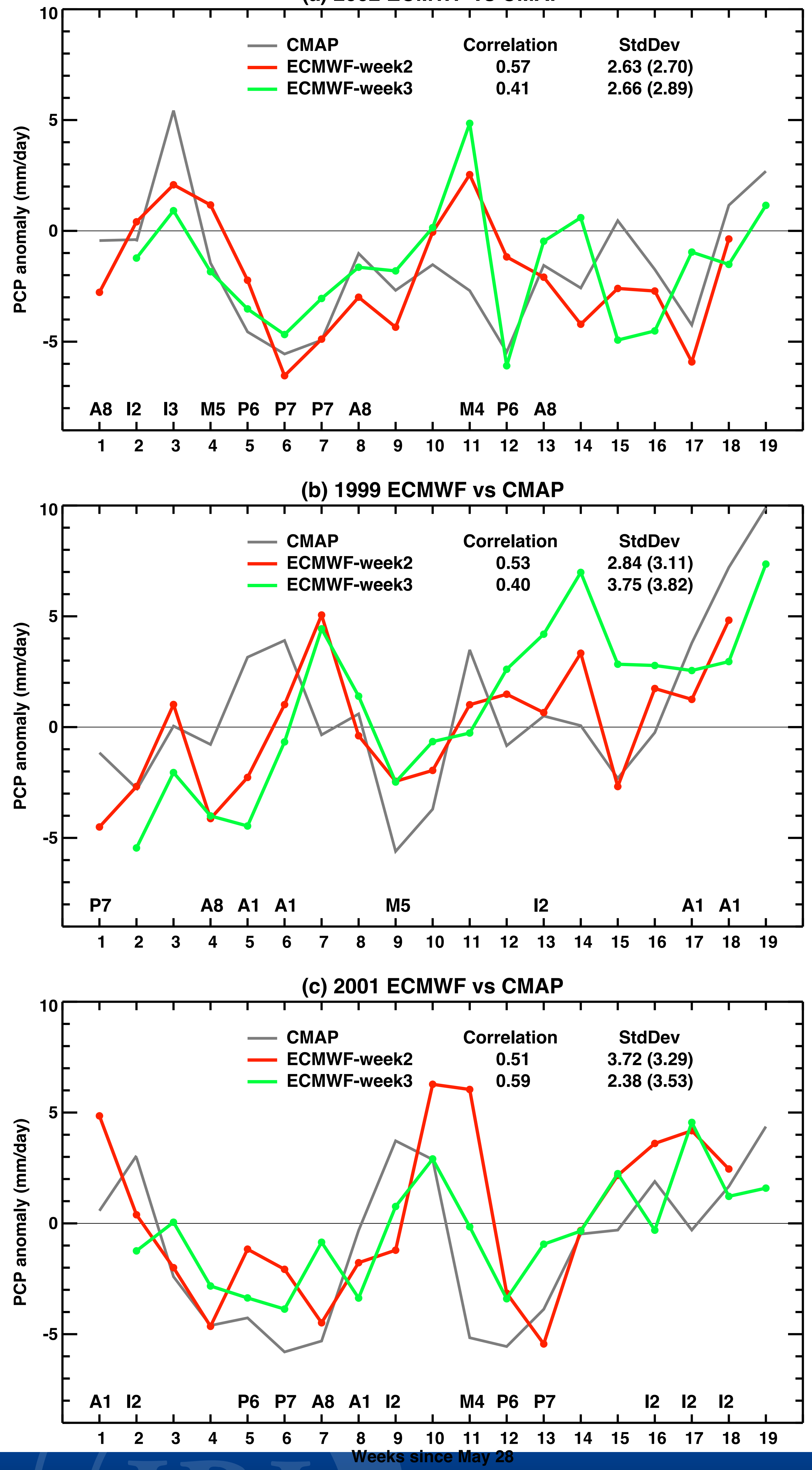
ECMWF Performance over Borneo

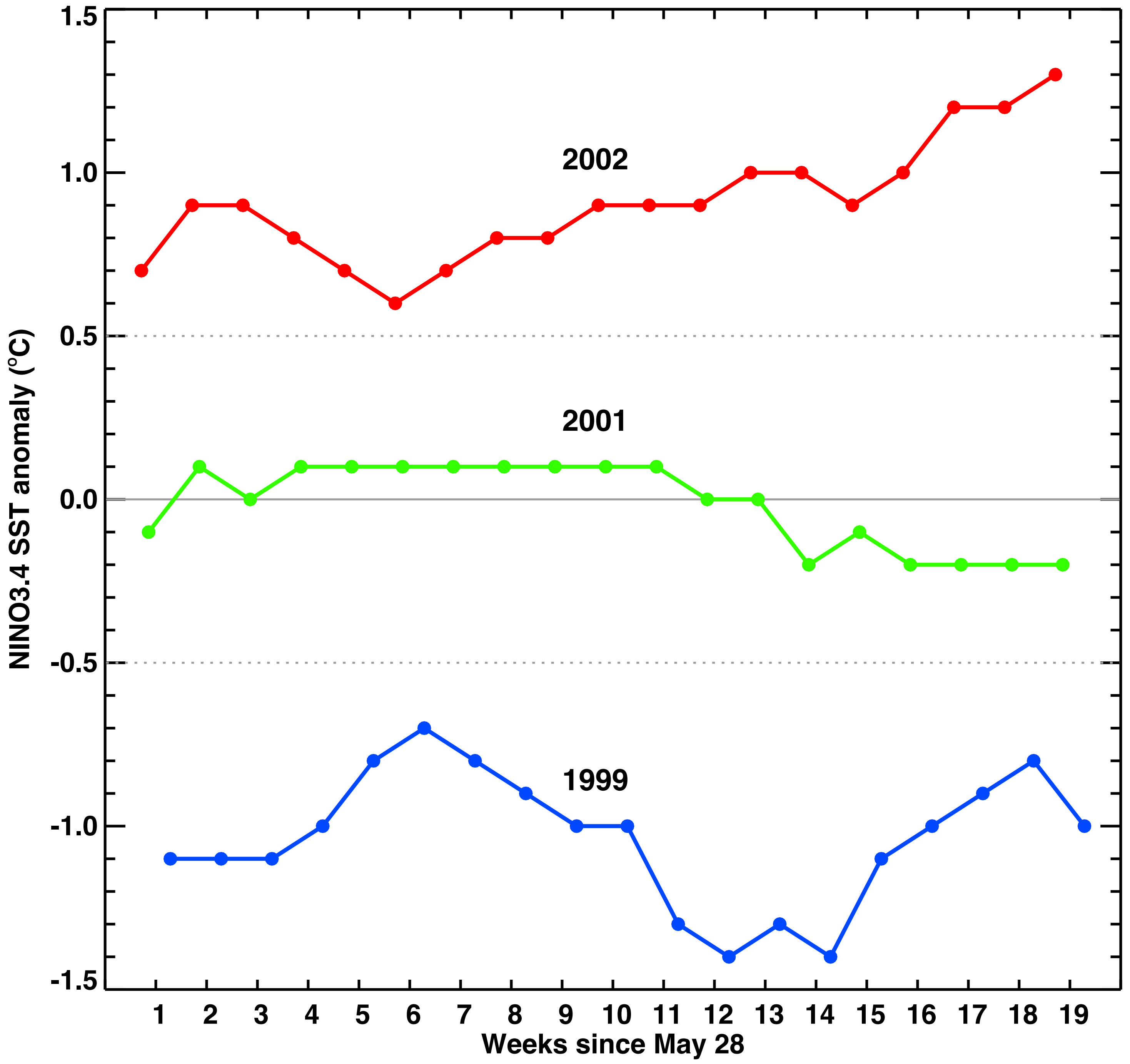


Labeled dots for each day.
Blue for Jun, orange for Jul, green for Aug, and red for Sep.



ECMWF Performance over Borneo

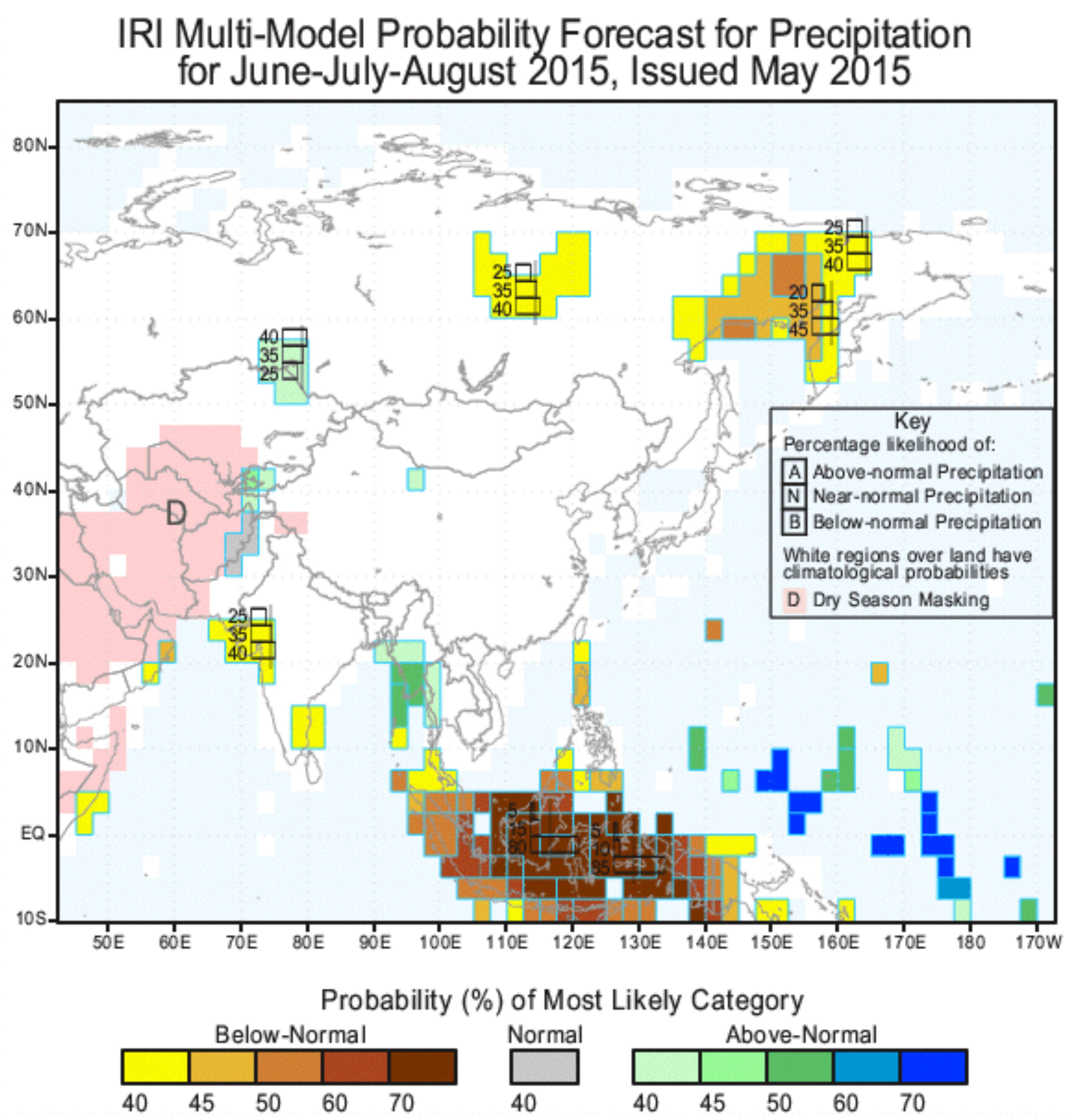




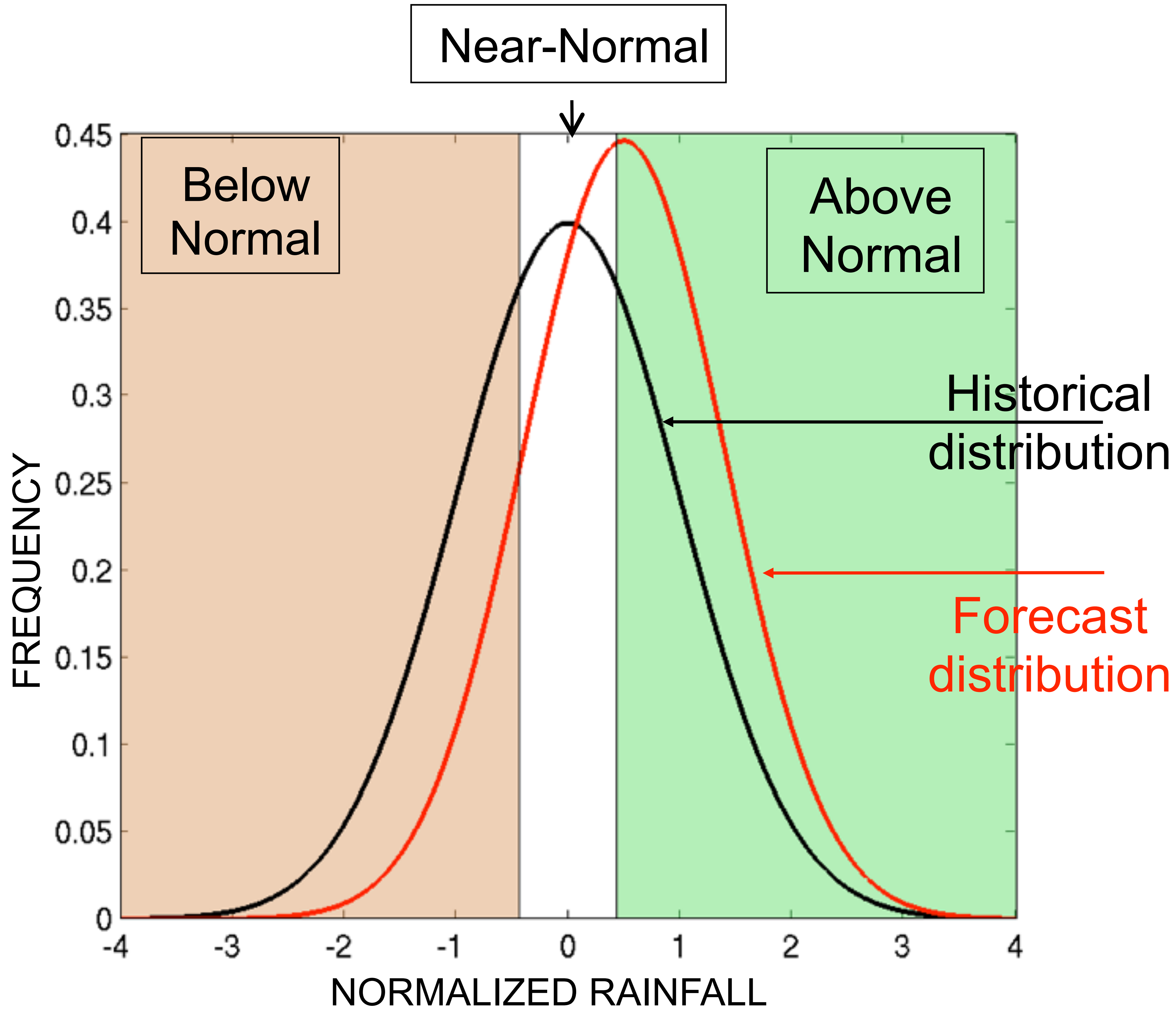
S2S partners

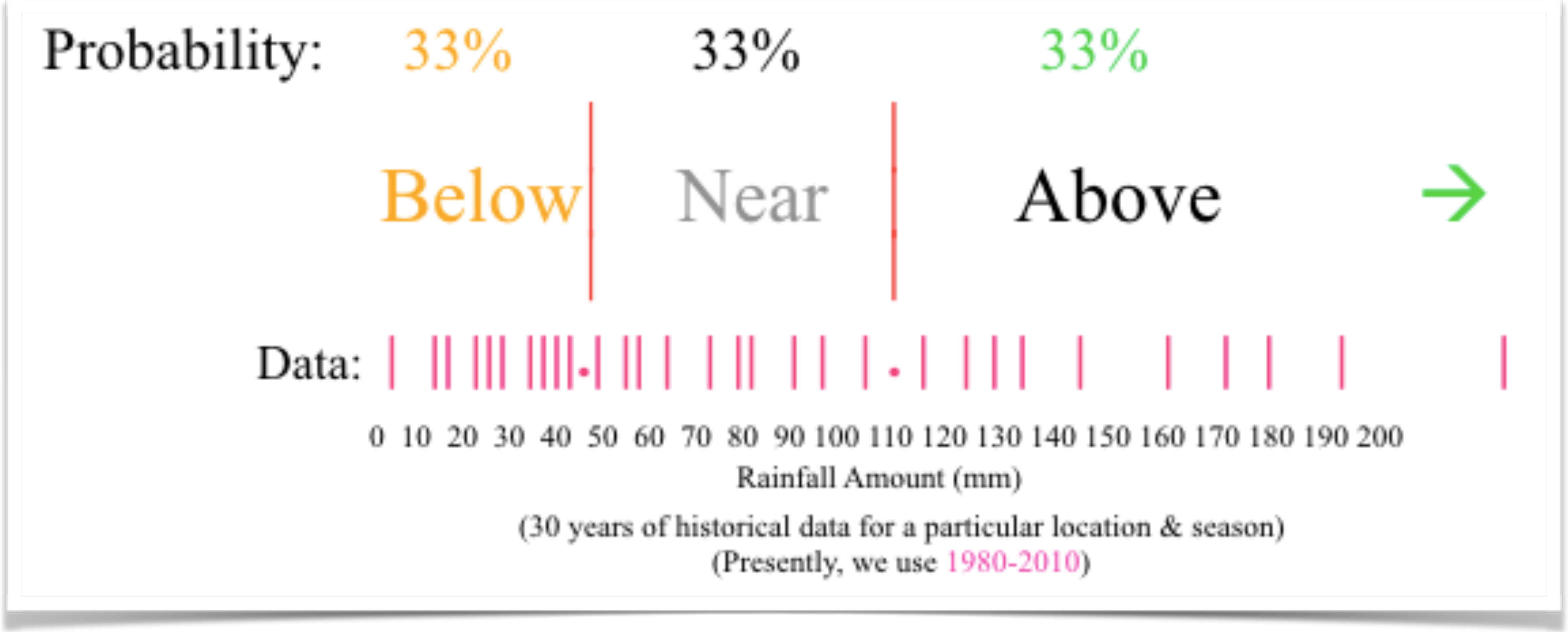
	Time-range	Resol.	Ens. Size	Freq.	Hcsts	Hcst length	Hcst Freq	Hcst Size
ECMWF	D 0-32	T639/319L91	51	2/week	On the fly	Past 18y	2/weekly	11
UKMO	D 0-60	N96L85	4	daily	On the fly	1989-2003	4/month	3
NCEP	D 0-45	N126L64	4	4/daily	Fix	1999-2010	4/daily	1
EC	D 0-35	0.6x0.6L40	21	weekly	On the fly	Past 15y	weekly	4
CAWCR	D 0-60	T47L17	33	weekly	Fix	1981-2013	6/month	33
JMA	D 0-34	T159L60	50	weekly	Fix	1979-2009	3/month	5
KMA	D 0-60	N216L85	4	daily	On the fly	1996-2009	4/month	3
CMA	D 0-45	T106L40	4	daily	Fix	1992-now	daily	4
Met.Fr	D 0-60	T127L31	51	monthly	Fix	1981-2005	monthly	11
CNR	D 0-32	0.75x0.56 L54	40	weekly	Fix	1981-2010	6/month	1
HMCR	D 0-63	1.1x1.4 L28	20	weekly	Fix	1981-2010	weekly	10

Probabilistic Verification: Was this a good forecast?



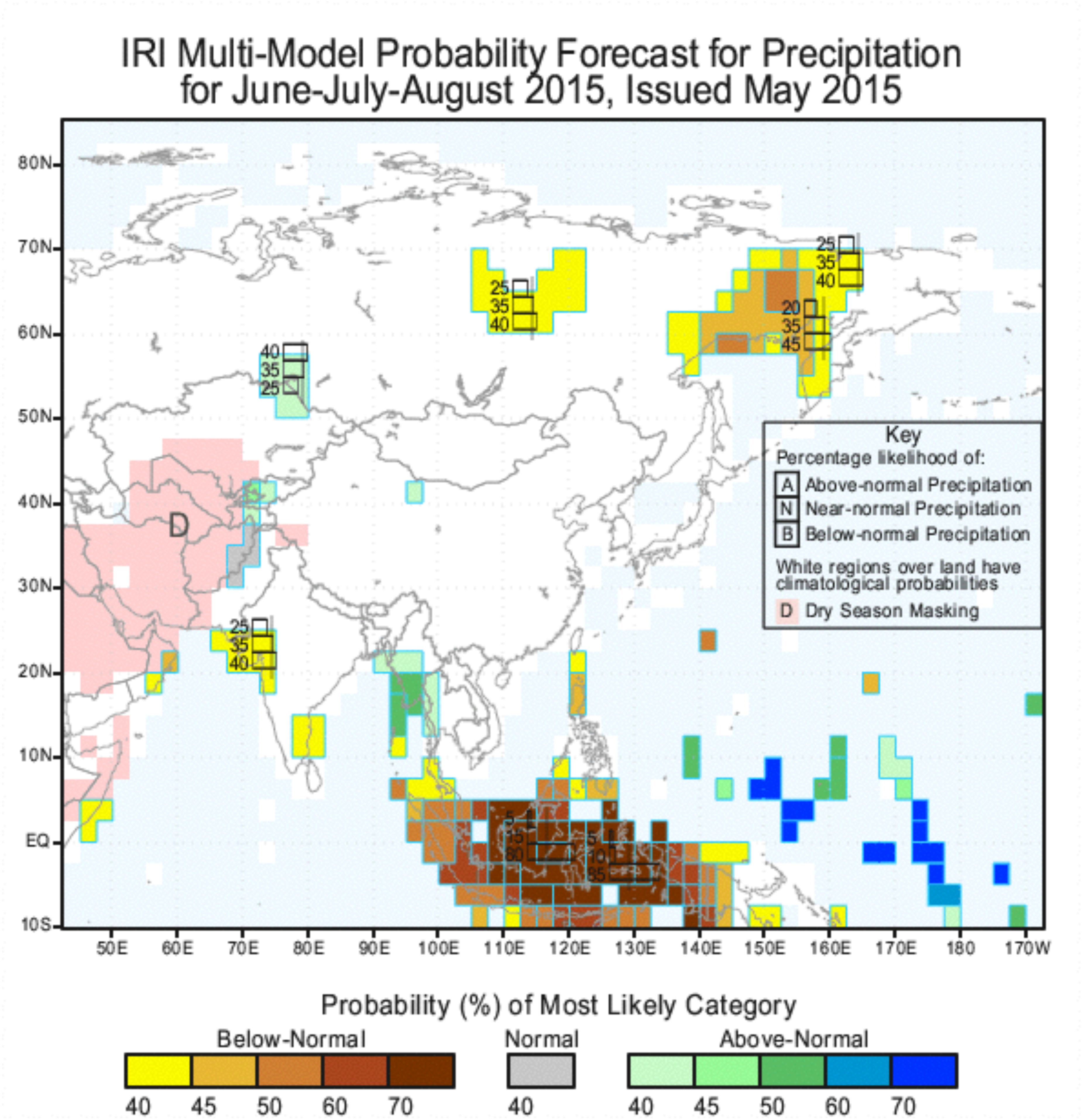
Displaying forecast probabilities



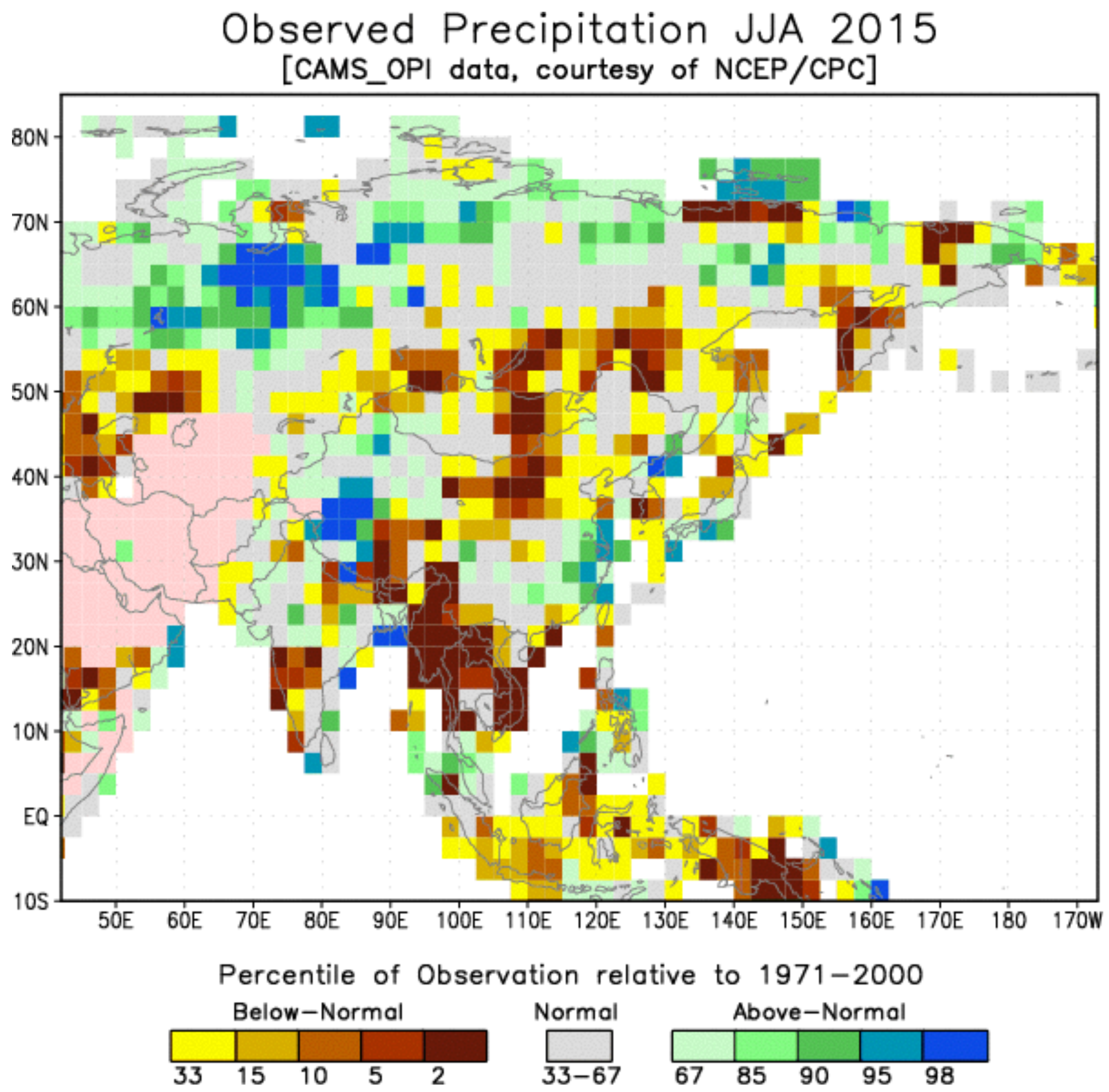


Historically, the probabilities of above and below are 0.33. Shifting the mean by half a standard-deviation and reducing the variance by 20% changes the probability of below to 0.15 and of above to 0.53.

Validation of a single probabilistic forecast



Forecast PDF



Verifying observation

Key Attributes of Probabilistic Forecasts

- **Sharpness:** refers to the concentration of the forecast distributions. The sharper, the better, provided the predictive distributions are calibrated.
- **Reliability:** Are the forecast probabilities correct on average, or is there some systematic bias toward under- or over-confidence?

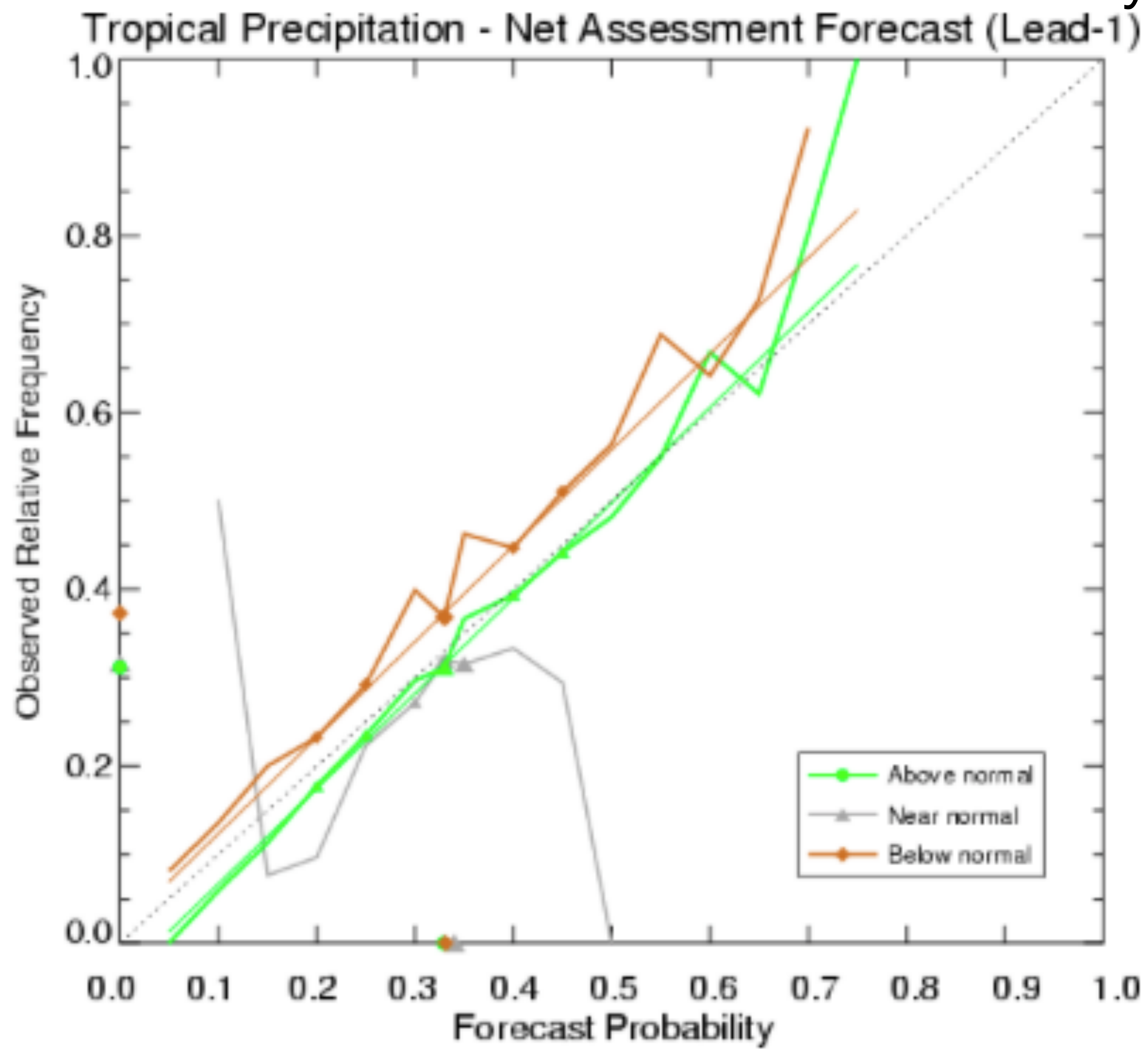
Reliability

- Did we correctly indicate the uncertainty in the forecast?

- Shows how well the forecast probabilities correspond to the subsequent observed relative frequencies of occurrence, across the full range of issued forecast probabilities

- The issued probabilities (from hindcasts) have to be binned, eg 0.45-0.55, 0.55-0.65, etc, so need long hindcast sets and pooling over space

IRI Real-time Forecast Reliability



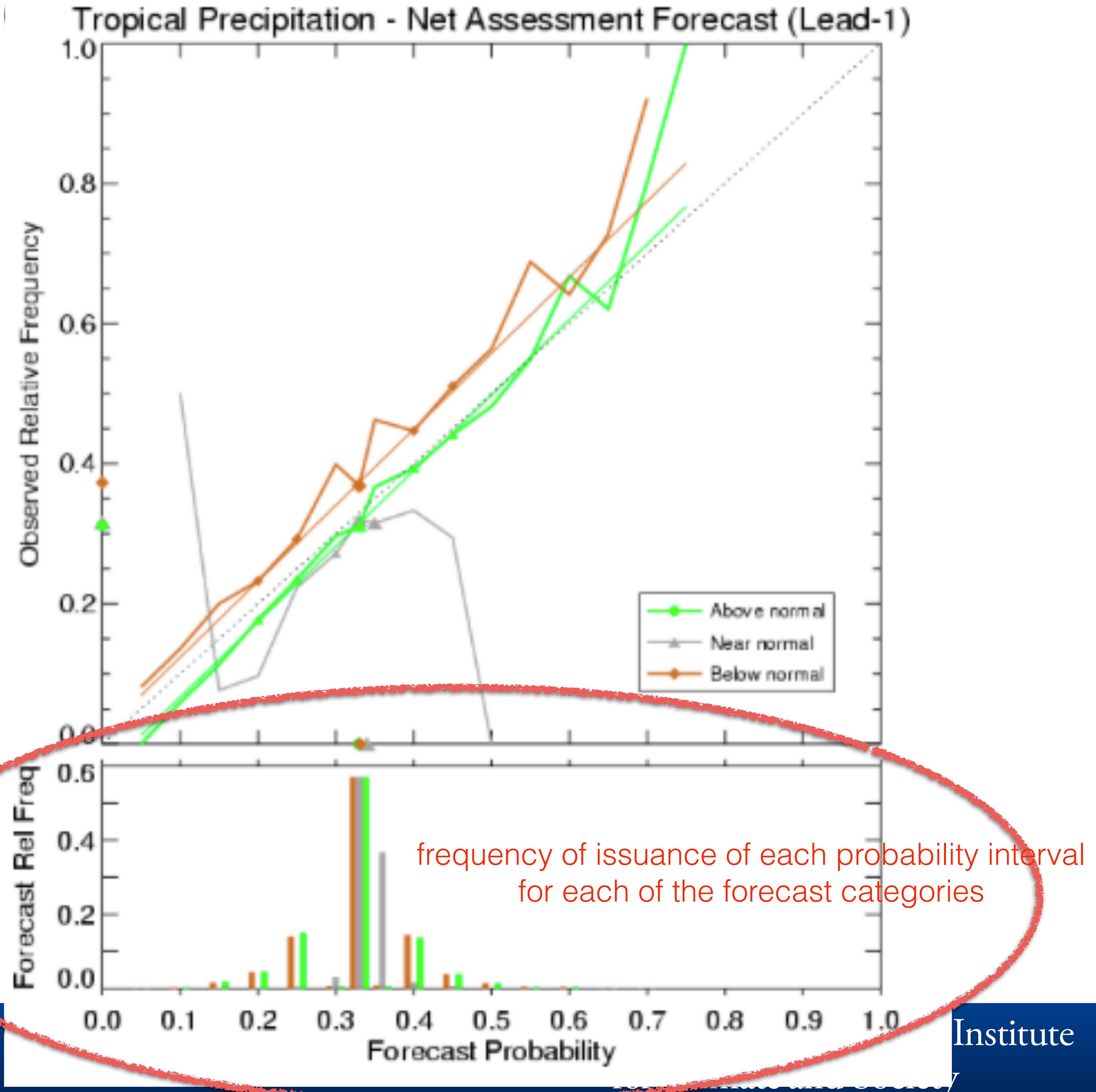
This set of forecasts is well calibrated!

Sharpness

- Sharpness measures whether the forecasts vary much from the climatological distribution.

- Most seasonal forecasts avoid being overly precise (3, or maybe 5, categories).

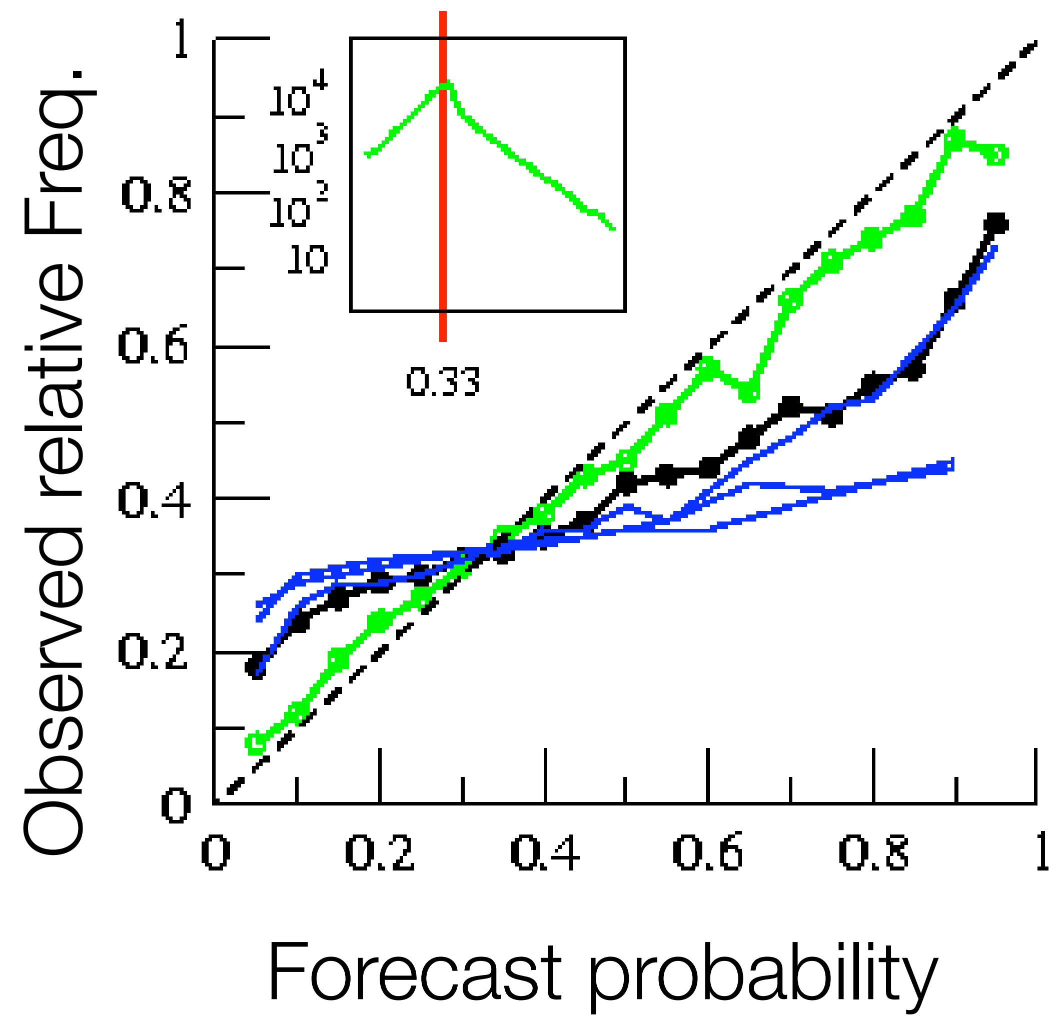
- If probabilities near 0 and 1 (100%) are used often, then the forecast is said to be sharp. If most of the forecast probabilities are in the range 40 to 60% then this forecast system would be said to be "smooth" or "not sharp" (as on right).



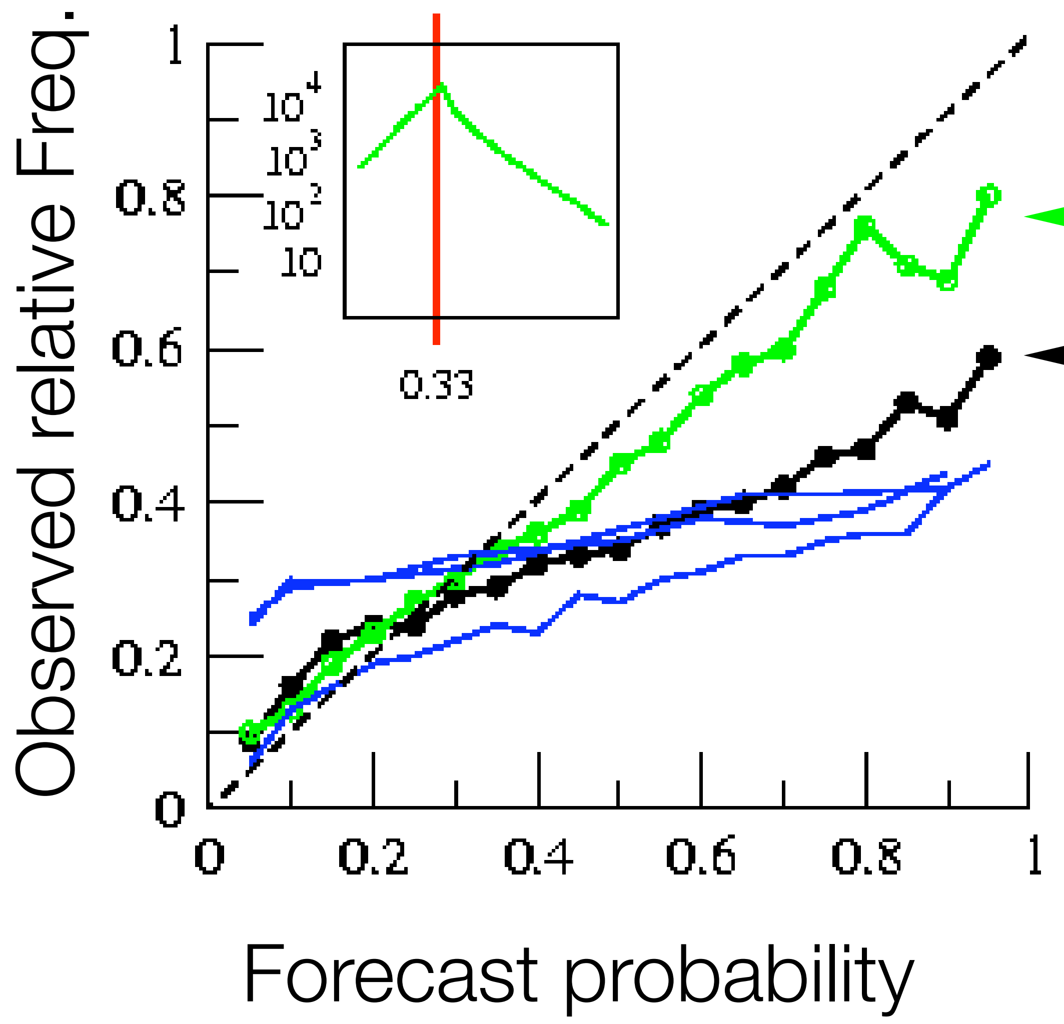
Examples of Seasonal Hindcast Reliability of 3 GCMs

JAS Precip., 30S-30N (3-model)

Above-Normal



Below-Normal



Combination of 3 GCMs

Pooling of 3 GCMs

S2S Sub-project on verification and products: Science questions

- What forecast quality attributes are important when verifying S2S forecasts and how they should be assessed?

  Which verification methods and forecast attributes are appropriate for reporting S2S forecast quality to users, and which provide added insight into forecast system development and improvement?

- How should issues of short hindcast period availability and reduced number of ensemble members in hindcasts compared to real-time forecasts be dealt with when constructing probabilistic skill measures?

- How can we best identify windows of forecast opportunity, including assessing the contributions of climate drivers, such as the MJO and ENSO, to S2S forecast skill (e.g. consider skill assessment conditioned on ENSO phases)?

- Which verification methods are most appropriate for the verification of extreme events, particularly given challenges associated with their rarity, small sample sizes and large uncertainties?

- How can we best verify active and break rainfall phases and wet/dry spells in current S2S forecast systems?

- How can we best address verification in a seamless manner, for comparing forecasts across timescales?

An S2S example

CFSv2 re-forecasts calibrated with extended logistic regression (Wilks 2009)

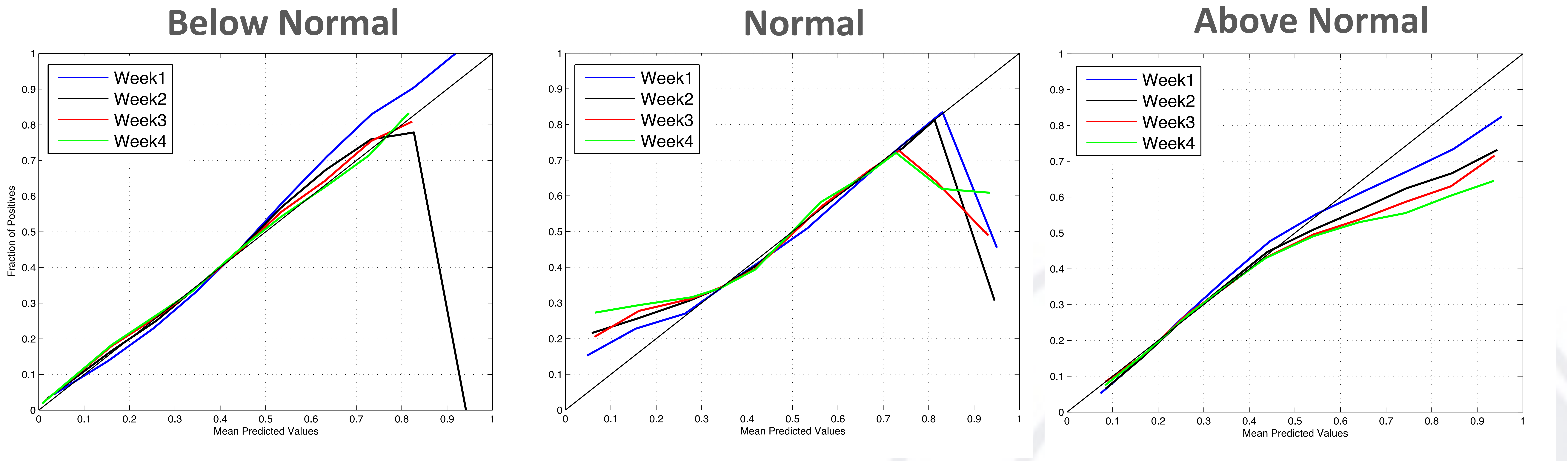


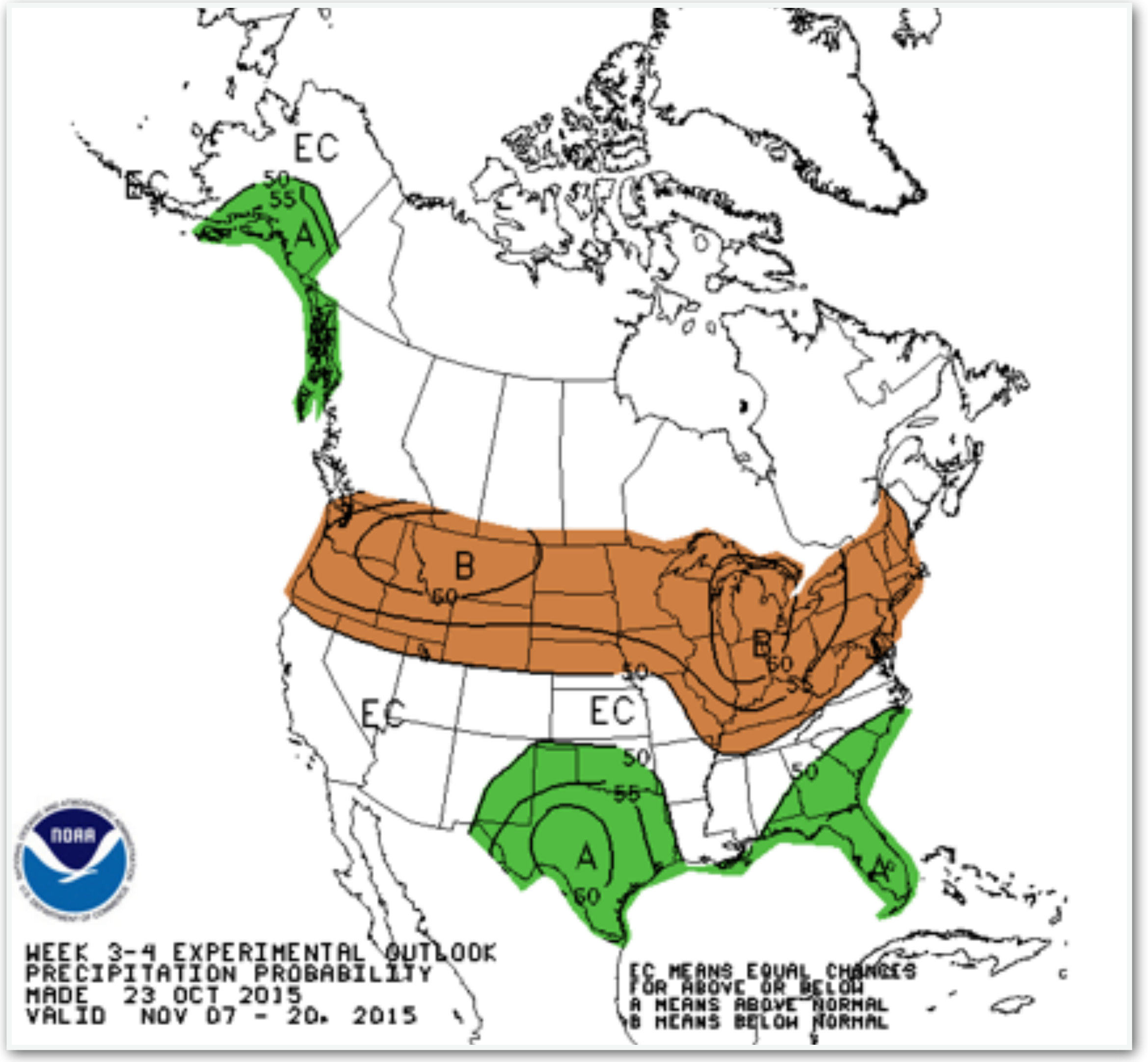
Figure 5: Reliability diagram for forecasts issued in JAS over the 1999-2010 period from one to four weeks lead over a US only window (land and ocean points)

courtesy of Nicolas Vigaud, IRI

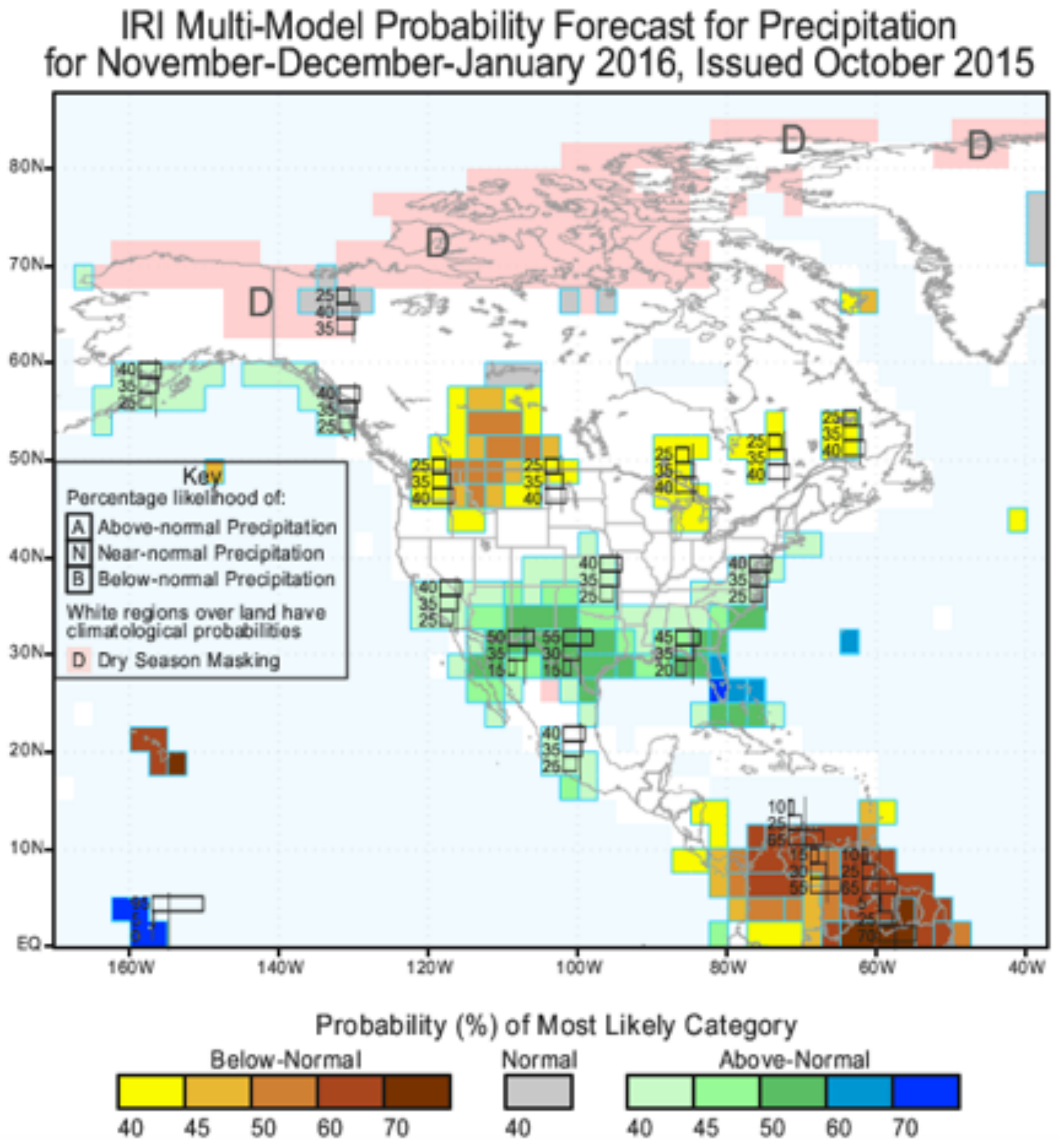
Which Forecast Format?

Daily weather Fcst

Week 3-4 Outlook



Seasonal Fcst



Summary of main points

- Forecast verification requires large sets of forecasts of reforecasts/ hindcasts. This poses questions for S2S where the hindcast sets are shorter than typically for seasonal forecasts, and the ensemble sizes are reduced

- Verification of probabilistic forecasts involves considering many attributes of forecast quality. Reliability and sharpness are important attributes.

- Calibration intimately involves verification because it seeks to maximize sharpness while maintaining reliability. But re-forecast data must not be used twice!