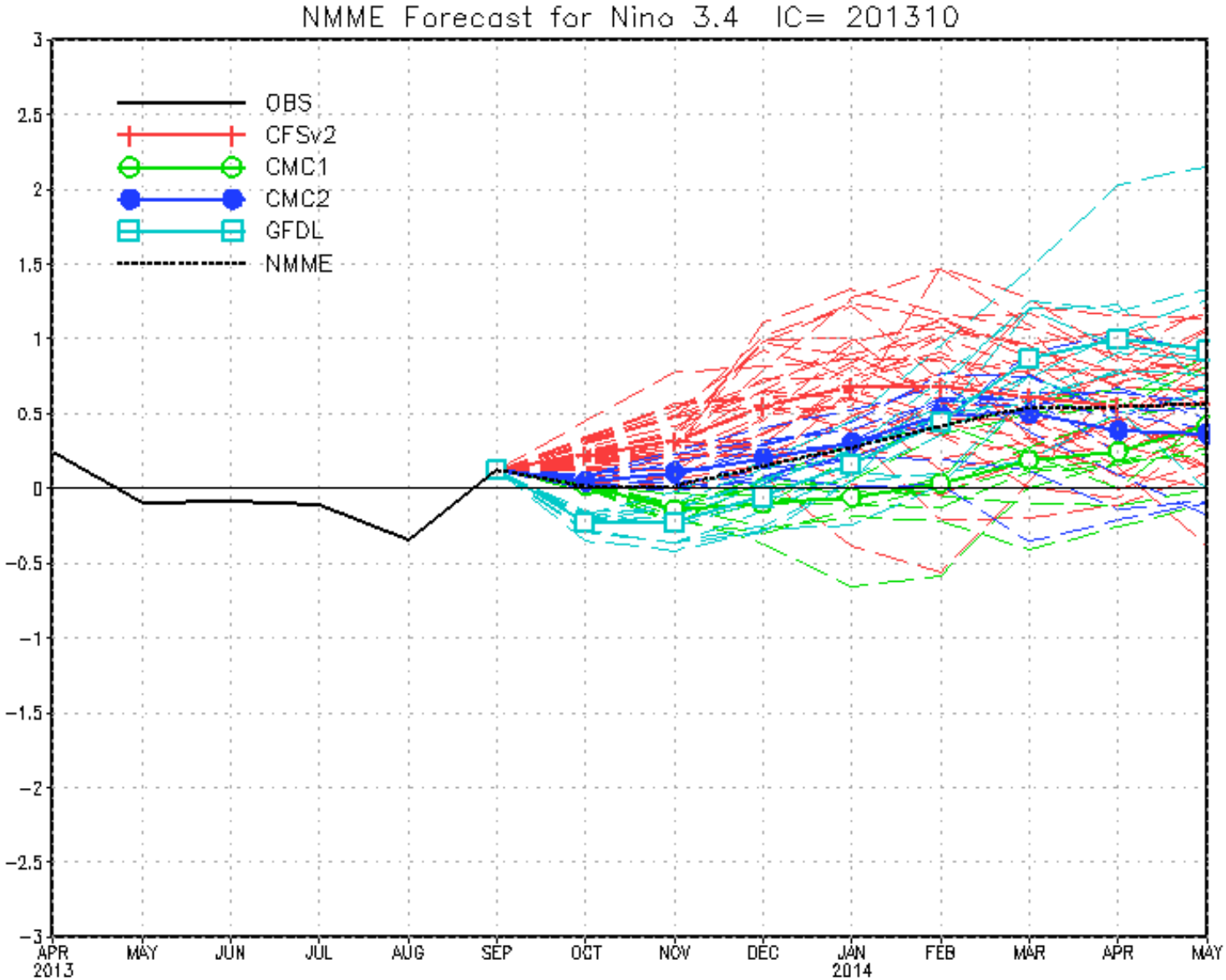
Does a Multi-Model Ensemble Enhance Skill?

Timothy DelSole

George Mason University, Fairfax, Va and
Center for Ocean-Land-Atmosphere Studies, Calverton, MD

October 22, 2013

Collaborators: Michael Tippett and Jyothi Natalla

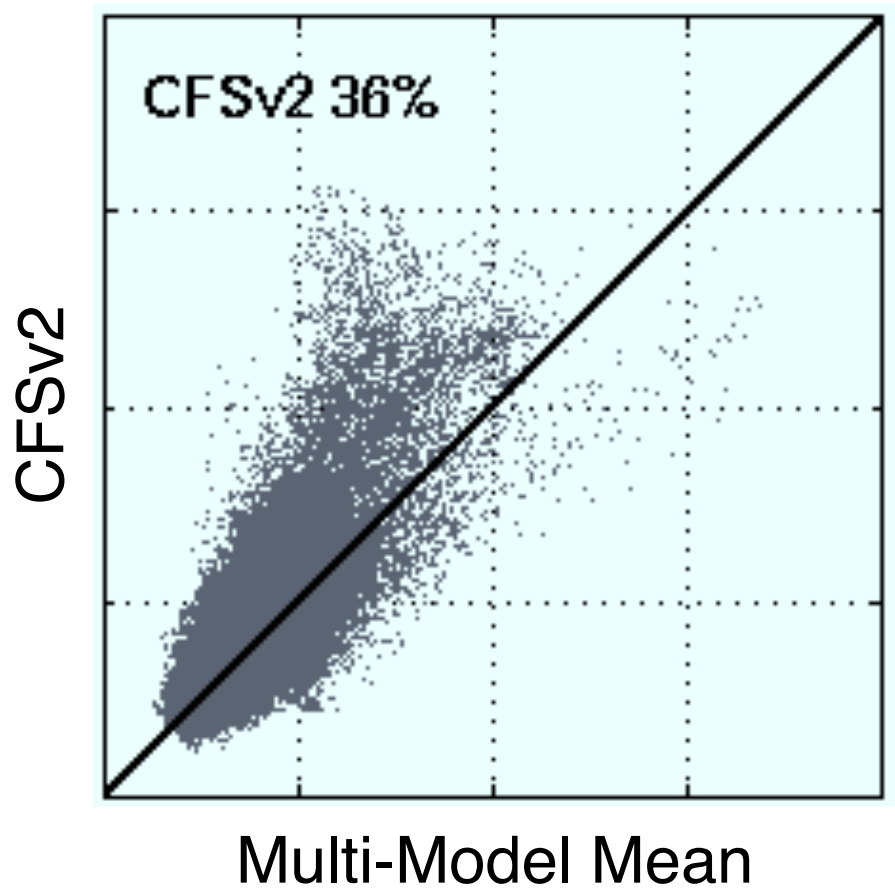

NMME Forecast for Nino 3.4 IC= 201310
OBS
CFSv2
CMC1
CMC2
GFDL
NMME
3
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
−3
APR
2013
MAY
JUN
JUL
AUG
SEP
OCT
NOV
DEC
JAN
2014
FEB
MAR
APR
MAY

Question about the NMME

Is the skill of a combined forecast significantly higher than that of a single forecast?

Compare Mean Square Error?

Root Mean Square Error SSTA
20S-20N, Sept. Starts, 1982-2009



Kirtman et al. 2013: The North American Multi-Model Ensemble... *BAMS*

Assertion

A rigorous test for the difference in MSEs does not exist when

- MSEs are calculated from the same verification.
- the prediction models are not nested.

Assertion

A rigorous test for the difference in MSEs does not exist when

- MSEs are calculated from the same verification.
- the prediction models are not nested.

Corollary: Confidence intervals cannot be attached to MSE to rigorously test differences in MSE.

Statement of the Question

o: observations

f: forecast from a model

c: combination of forecasts excluding f.

- The skill of a forecast f can be measured by correlation:

$$\rho = \text{cor}\,[o, f]$$

- The skill of the best linear combination of f and c can be measured by the *multiple correlation*

$$R = \max_{\beta_f, \beta_c} \text{cor}\,[o, \beta_f f + \beta_c c]$$

Statement of the Question

- o: observations
- f: forecast from a model
- c: combination of forecasts excluding f.

- The skill of a forecast f can be measured by correlation:

$$\rho = \text{cor}\,[o, f]$$

- The skill of the best linear combination of f and c can be measured by the *multiple correlation*

$$R = \max_{\beta_f, \beta_c} \text{cor}\,[o, \beta_f f + \beta_c c]$$

Question: Is $R > \rho$?

Statistical Test

$$\rho = \max_{\beta_f} \text{cor}\left[o, \beta_f f\right]$$

$$R = \max_{\beta_f, \beta_c} \text{cor}\left[o, \beta_f f + \beta_c c\right]$$

- The hypothesis $R = \rho$ is equivalent to the hypothesis $\beta_c = 0$.
- Testing the hypothesis $\beta_c = 0$ is standard and is based on

$$t = \frac{\hat{\beta}_c}{se_c}$$

- It can be shown that

$$t^2 = \frac{R^2 - \rho^2}{1 - R^2} \frac{N - 3}{1} = \frac{SSE_R - SSE_F}{SSE_F} \frac{N - 3}{1}$$

Equivalent Interpretation of the Test

A significant t-value means:
if the forecast f is regressed out of c, the residual c still has skill.

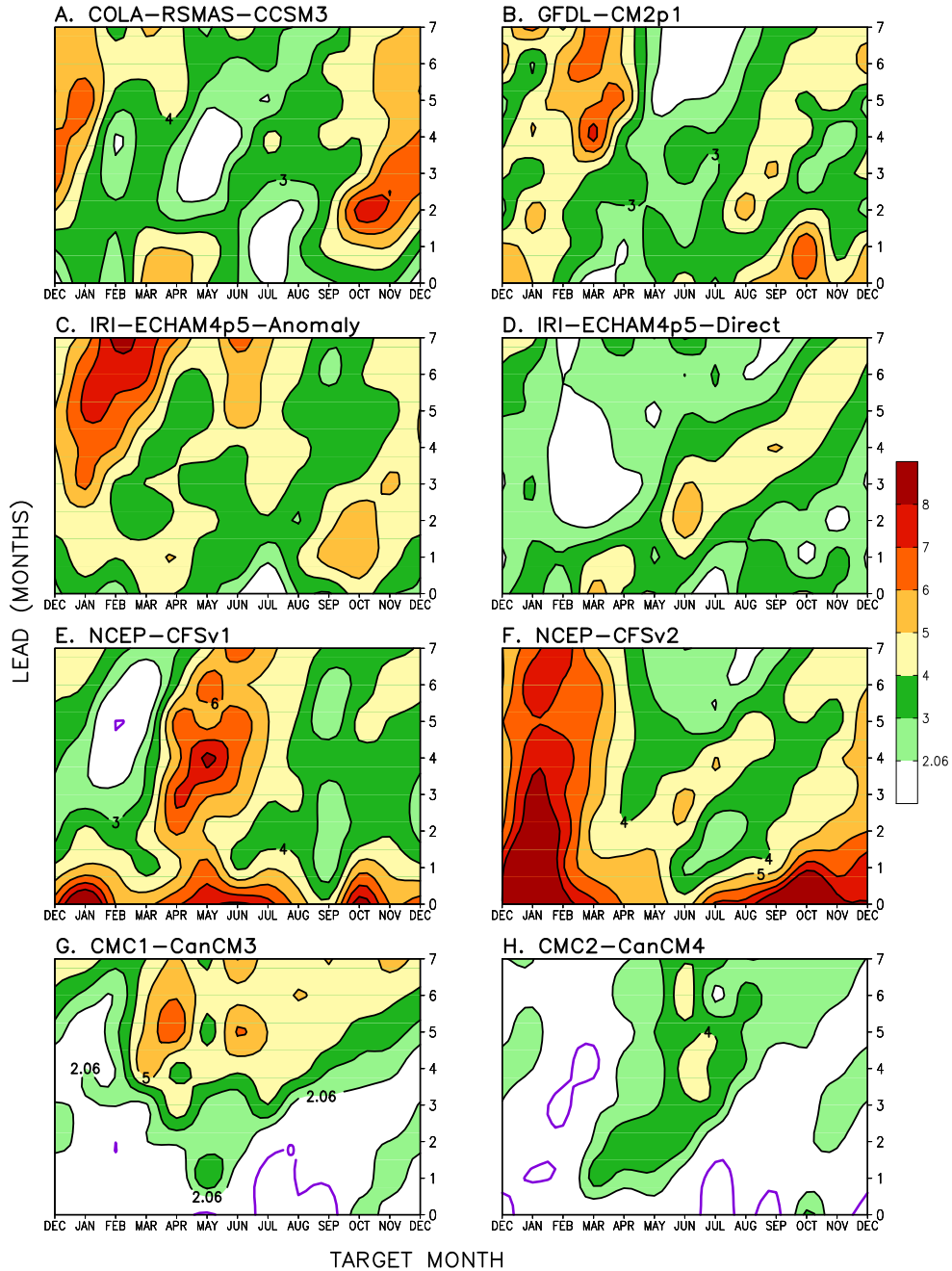
Combined Forecast

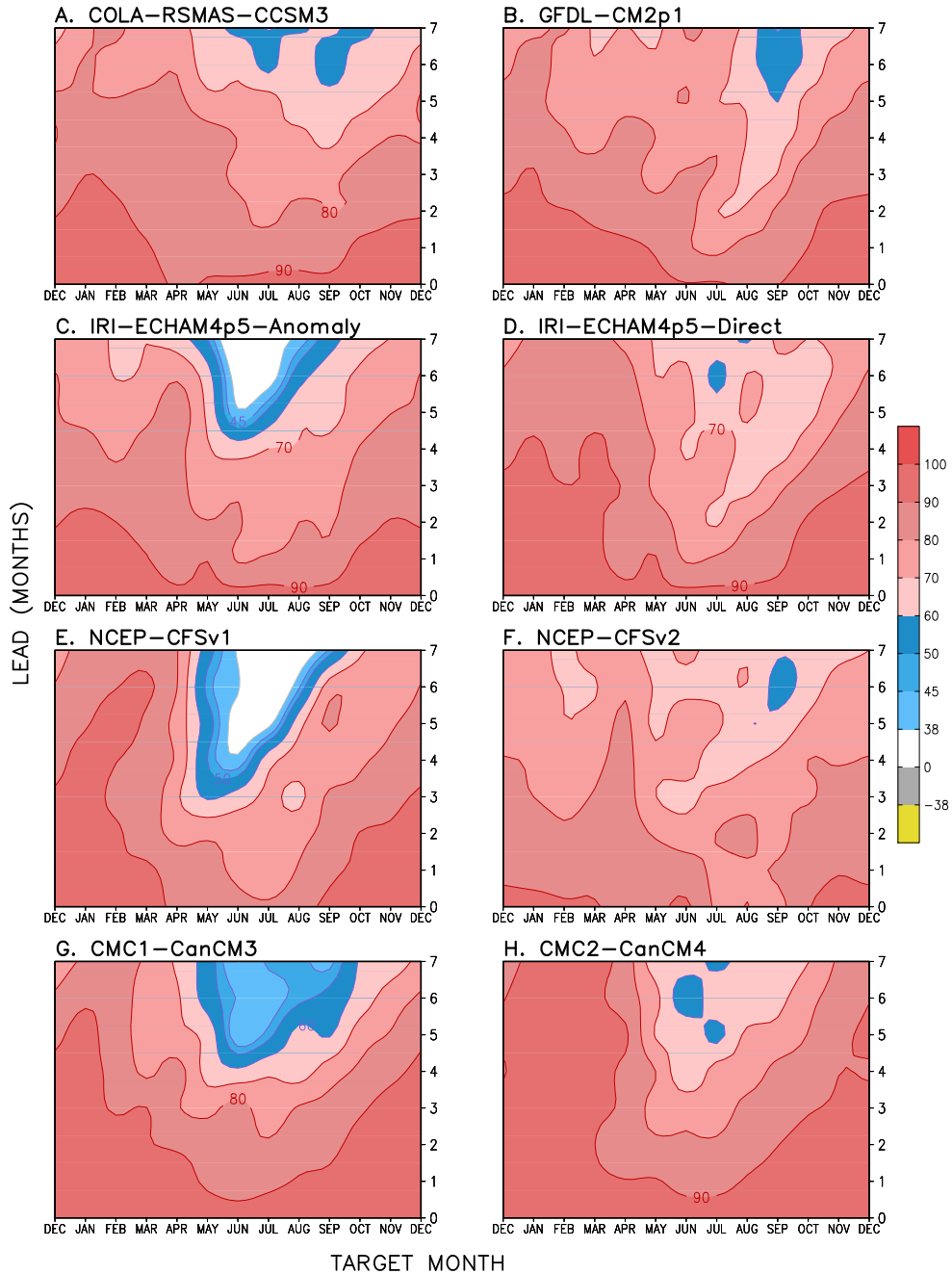
- We consider only equal weighting schemes.
- Equal weights is very competitive with more sophisticated schemes
  - Kharin and Zwiers, 2002, *J. Climate*
  - Hagedorn et al. 2005, *Tellus A*
  - Weigel et al. 2010, *J. Climate*
  - DelSole et al. 2012, *J. Climate*
  - Sansom et al. 2013, *J. Climate*

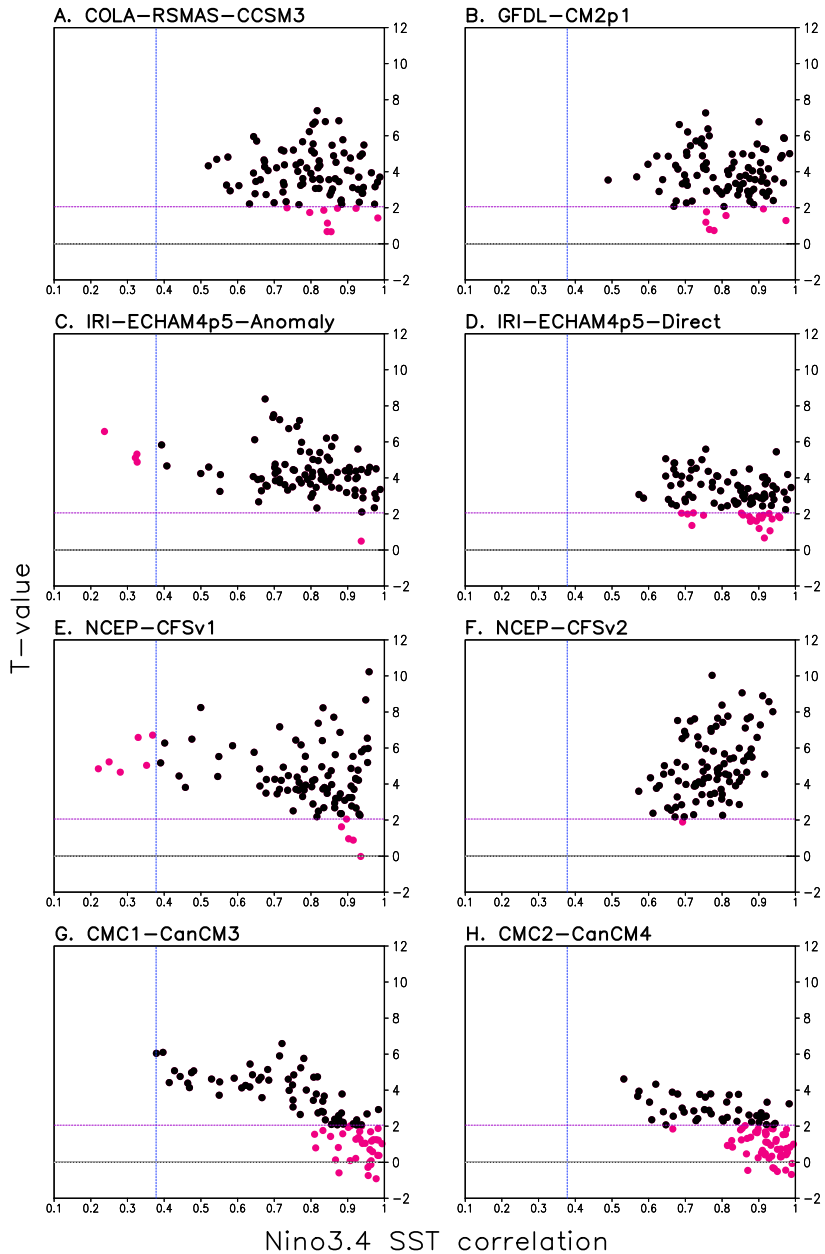
North American Multi-Model Ensemble

- Hindcasts initialized every month from 1982-2009
- At least 6 month lead
- Analyze NINO3.4
- Verification: OISST

model	ensemble size
CMC1-CanCM3	10
CMC2-CanCM4	10
COLA-RSMAS-CCSM3	6
GFDL-CM2p1	10
NASA-GMAO	10
NCEP-CFSv1	10
NCEP-CFSv2	10


A. COLA–RSMAS–CCSM3
B. GFDL–CM2p1
C. IRI–ECHAM4p5–Anomaly
D. IRI–ECHAM4p5–Direct
E. NCEP–CFSv1
F. NCEP–CFSv2
G. CMC1–CanCM3
H. CMC2–CanCM4
LEAD (MONTHS)
TARGET MONTH
DEC JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC
0 1 2 3 4 5 6 7
2.06
3
4
5
6
8


A. COLA−RSMAS−CCSM3
B. GFDL−CM2p1
C. IRI−ECHAM4p5−Anomaly
D. IRI−ECHAM4p5−Direct
E. NCEP−CFSv1
F. NCEP−CFSv2
G. CMC1−CanCM3
H. CMC2−CanCM4
LEAD (MONTHS)
TARGET MONTH
DEC JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC
0 1 2 3 4 5 6 7
80
90
70
100
60
50
45
38
−38


A. COLA−RSMAS−CCSM3
B. GFDL−CM2p1
C. IRI−ECHAM4p5−Anomaly
D. IRI−ECHAM4p5−Direct
E. NCEP−CFSv1
F. NCEP−CFSv2
G. CMC1−CanCM3
H. CMC2−CanCM4
T−value
Nino3.4 SST correlation

Summary

1. Proposed an objective procedure for deciding if the skill of a combined forecast is significantly higher than a single forecast.
2. Skill of each model in NMME is significantly enhanced by combining it with other models, at least for some lead time and target month.