Recent Developments in Forecast Quality Assessment

Timothy DelSole

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

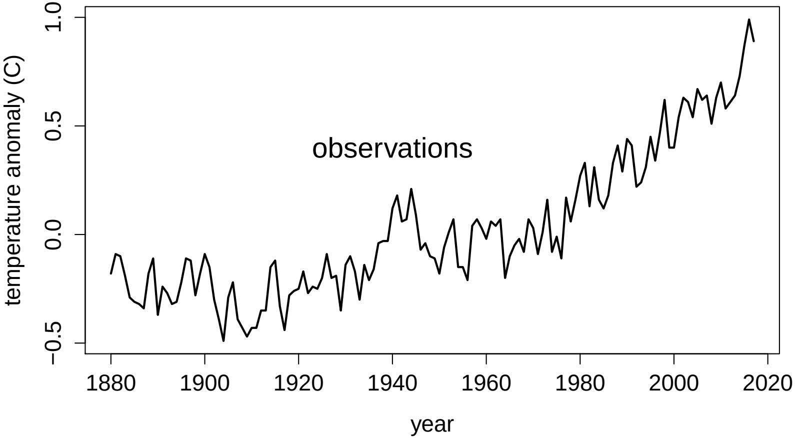
September 16, 2018

collaborator: Michael Tippett

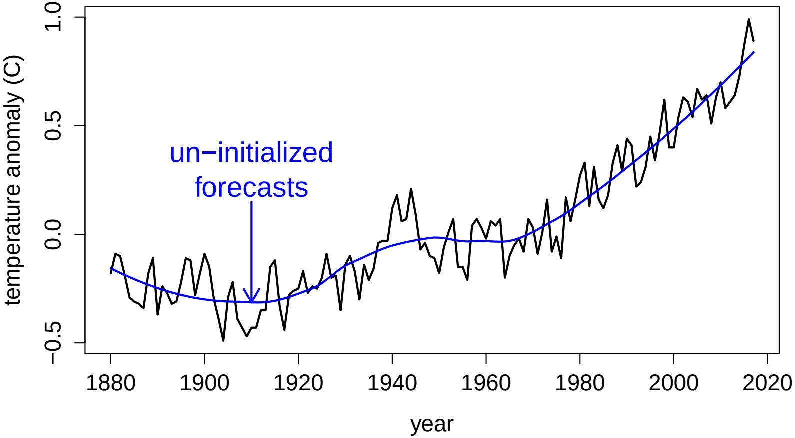
Is one forecast better than another?

- Operational forecasters: when to switch to new prediction system?
- Modelers: did the change in model improve skill?
- Scientists: why? resolution? initialization? physics?

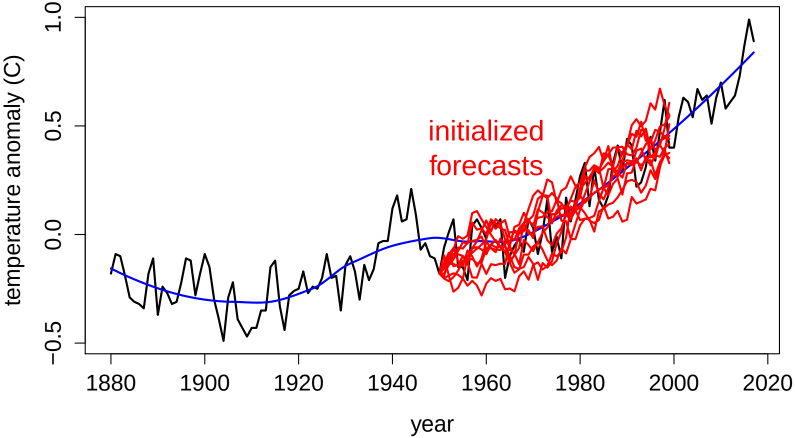
Initialized vs. Unitialized Forecasts



Initialized vs. Unitialized Forecasts

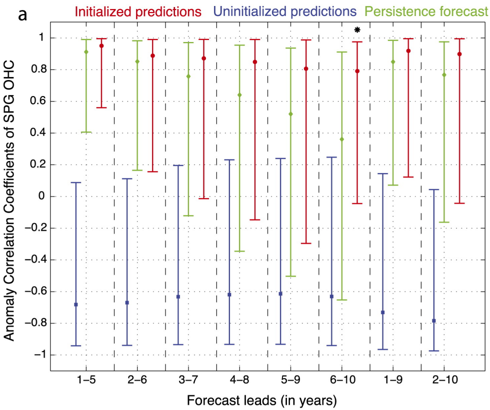


Initialized vs. Unitialized Forecasts



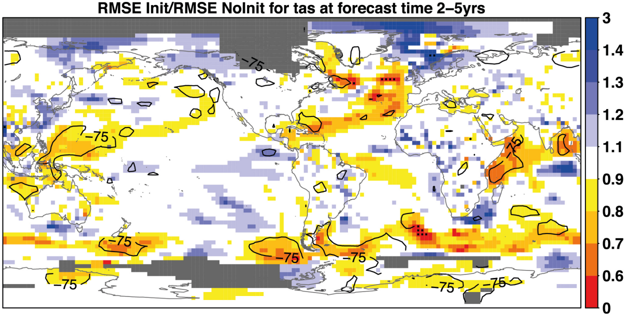
Deterministic Skill Measures for a Time Series

- correlation coefficient
- mean square error



Anomaly correlations of the North Atlantic Subpolar Gyre OHC anomalies (circle). The bar indicates the two-sided 90% confidence interval using Fishers z transform.

Msadek et al., 2014, J. Climate



Ratio of root mean square error of initialized over uninitialized decadal hindcasts. Dots indicate where the ratio is significantly above or below 1 with 90% confidence using a two-sided F-test.

IPCC AR5 WG1 fig. 11.4

Test Equality of Variance ($\sigma_1^2 = \sigma_2^2$)

Statistic: Let s_1^2 and s_2^2 be the sample variances:

$$F = \frac{s_1^2}{s_2^2}.$$

Theorem: If samples are independent and identically distributed as a Gaussian, then

$$F \sim F_{\nu_1, \nu_2}.$$

where ν_1 and ν_2 are the appropriate degrees of freedom.

Test Equality of Variance ($\sigma_1^2 = \sigma_2^2$)

Statistic: Let s_1^2 and s_2^2 be the sample variances:

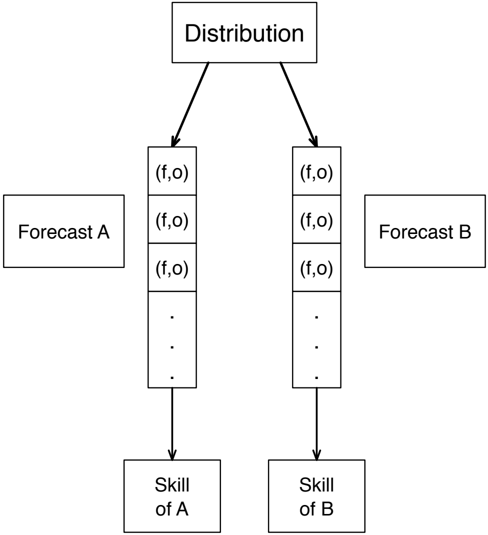
$$F = \frac{s_1^2}{s_2^2}.$$

Theorem: If samples are **independent** and identically distributed as a Gaussian, then

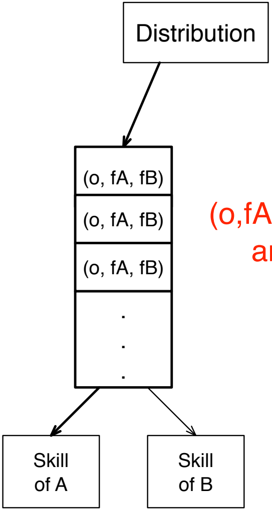
$$F \sim F_{\nu_1,\nu_2}.$$

where ν_1 and ν_2 are the appropriate degrees of freedom.

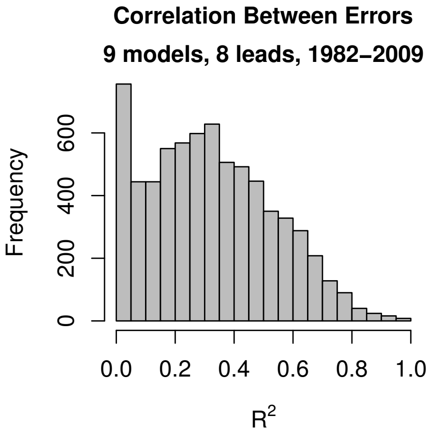
Standard Tests Assume Forecast-Verification Pairs are Independent



For Model Comparisons, Forecast-Verification Pairs are Dependent

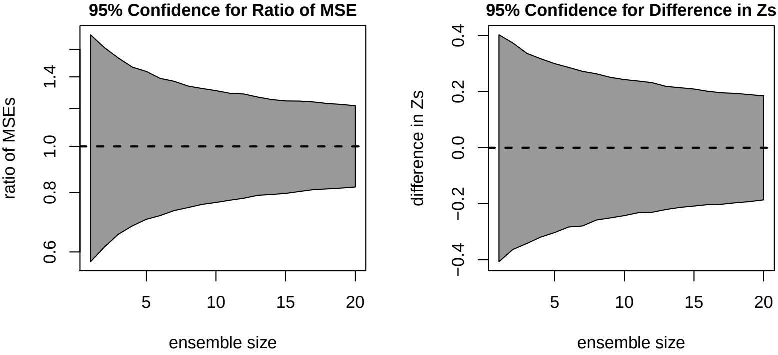


(o,fA) and (o,fB) pairs are dependent!



NMME skill estimates tend to be correlated in seasonal forecasting.

$$
\begin{aligned}
\text{observation} &= \text{signal} + \text{noise} \\
\text{forecast A} &= \text{signal} + \text{noise* } /\sqrt{E} \\
\text{forecast B} &= \text{signal} + \text{noise** } /\sqrt{E}
\end{aligned}
$$



Summary

1. Commonly used tests for skill differences are not valid if skills are computed using a common set of observations.
2. These tests do not account for correlated prediction errors.
3. Familiar tests wrongly judge differences in skill as insignificant.
4. The bias is not negligible for typical seasonal forecasts.

Some legitimate model improvements may have gone undetected using standard tests.

What **IS** the proper way to compare forecast skill?

Journal of Business & Economic Statistics, July 1995, Vol. 13, No. 3

Comparing Predictive Accuracy

Francis X. Diebold
Department of Economics, University of Pennsylvania, Philadelphia, PA 19104-6297, and
National Bureau of Economic Research, Cambridge, MA 02138

Roberto S. Mariano
Department of Economics, University of Pennsylvania, Philadelphia, PA 19104-6297

We propose and evaluate explicit tests of the null hypothesis of no difference in the accuracy of two competing forecasts. In contrast to previously developed tests, a wide variety of accuracy measures can be used (in particular, the loss function need not be quadratic and need not even be symmetric), and forecast errors can be non-Gaussian, nonzero mean, serially correlated, and contemporaneously correlated. Asymptotic and exact finite-sample tests are proposed, evaluated, and illustrated.

KEY WORDS: Economic loss function; Exchange rates; Forecast evaluation; Forecasting; Nonparametric tests; Sign test.

Similar Approaches in Weather Prediction

- Thomas Hamill, 1999: *Hypothesis Tests for Evaluating Numerical Precipitation Forecasts*, Mon. Wea. Rev.

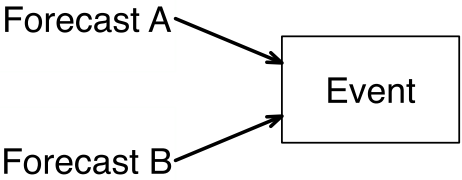
Similar Approaches in Weather Prediction

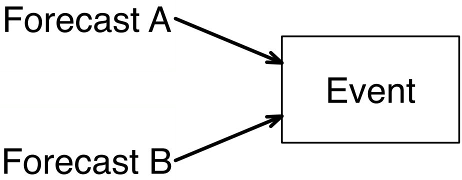
- Thomas Hamill, 1999: *Hypothesis Tests for Evaluating Numerical Precipitation Forecasts*, Mon. Wea. Rev.

Hering, A. S. and M. G. Genton, 2011: Comparing spatial predictions. *Technometrics*, **53 (4)**, 414–425.

DelSole, T. and M. K. Tippett, 2014: Comparing forecast skill. *Mon. Wea. Rev.*, **142**, 4658–4678.

Gilleland, E., A. S. Hering, T. L. Fowler, and B. G. Brown, 2018: Testing the tests... *Monthly Weather Review*, **146 (6)**, 1685–1703.

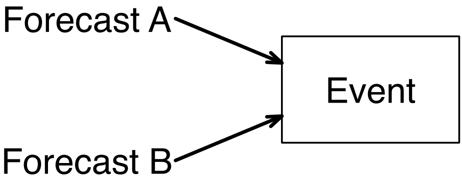

Forecast A
Forecast B
Event



If forecasts are equally skillful, then probability of

$$\text{skill of A} > \text{skill of B}$$

is 50%.



If forecasts are equally skillful, then probability of

$$\text{skill of A} > \text{skill of B}$$

is 50%. This is true:

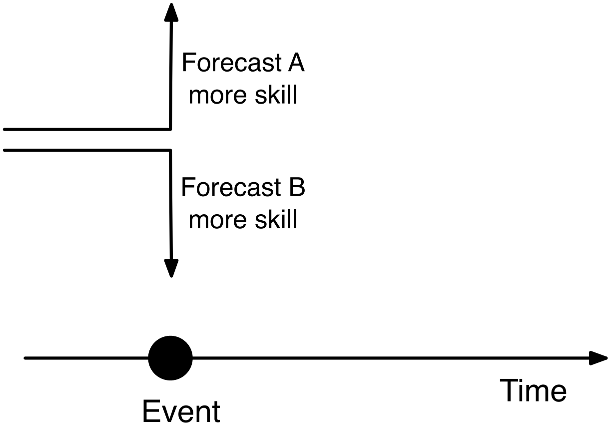
- regardless of the measure of skill.
- even if forecasts are highly correlated.
- regardless of forecast error distribution.

This test is exactly the test for deciding if a coin is fair.

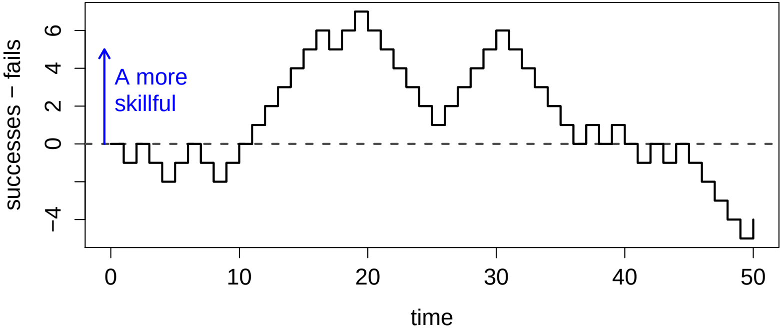
This test is exactly the test for deciding if a coin is fair.

- The number of heads follows a binomial distribution.
- The number of heads minus the number of tails is a **random walk**.

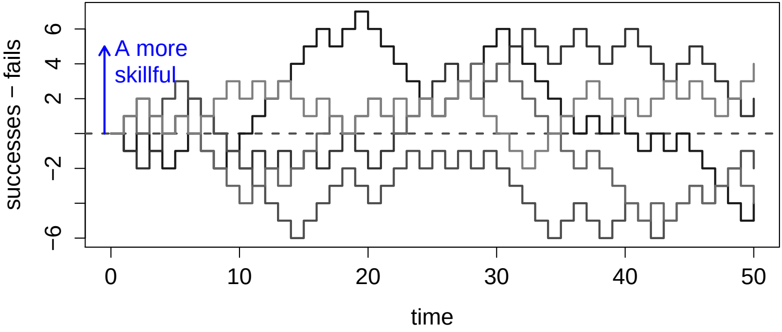
Random Walk Test



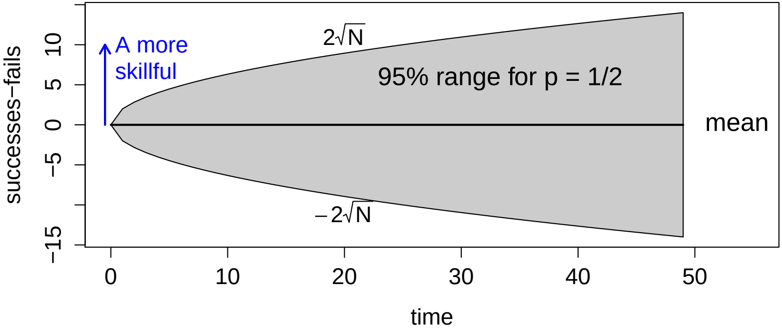
Random Walk Test



Random Walk Test



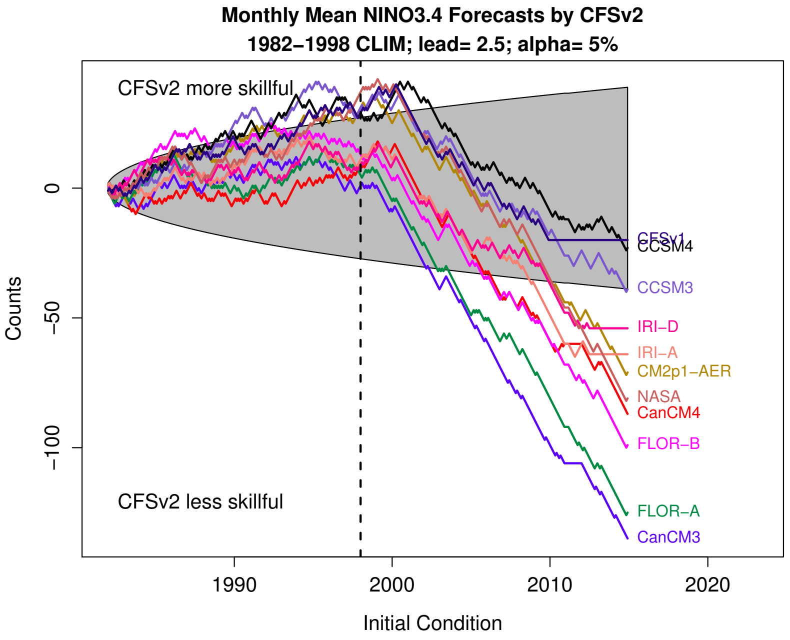
Random Walk Test



North American Multi-Model Ensemble

- Hindcasts initialized every month from 1982-2010 (29 years)
- Lead 2.5 months
- MSE of 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


Monthly Mean NINO3.4 Forecasts by CFSv2
1982–1998 CLIM; lead= 2.5; alpha= 5%
CFSv2 more skillful
CFSv2 less skillful
Counts
0
−50
−100
1990
2000
2010
2020
Initial Condition
CFSv1
CCSM4
CCSM3
IRI−D
IRI−A
CM2p1−AER
NASA
CanCM4
FLOR−B
FLOR−A
CanCM3

An Analysis of the Nonstationarity in the Bias of Sea Surface Temperature Forecasts for the NCEP Climate Forecast System (CFS) Version 2

A. Kumar and M. Chen

Climate Prediction Center, NOAA/NWS/NCEP, Camp Springs, Maryland

L. Zhang

Climate Prediction Center, NOAA/NWS/NCEP, Camp Springs, Maryland, and WYLE STE, McLean, Virginia

W. Wang and Y. Xue

Climate Prediction Center, NOAA/NWS/NCEP, Camp Springs, Maryland

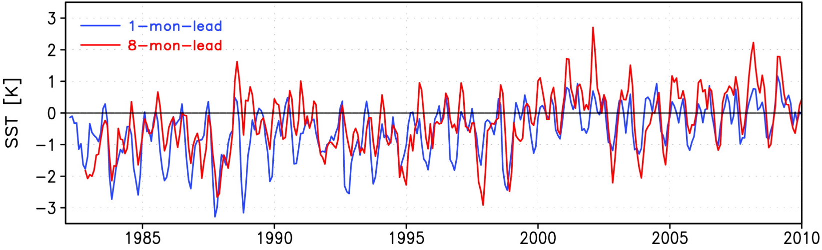
C. Wen

Climate Prediction Center, NOAA/NWS/NCEP, Camp Springs, Maryland, and WYLE STE, McLean, Virginia

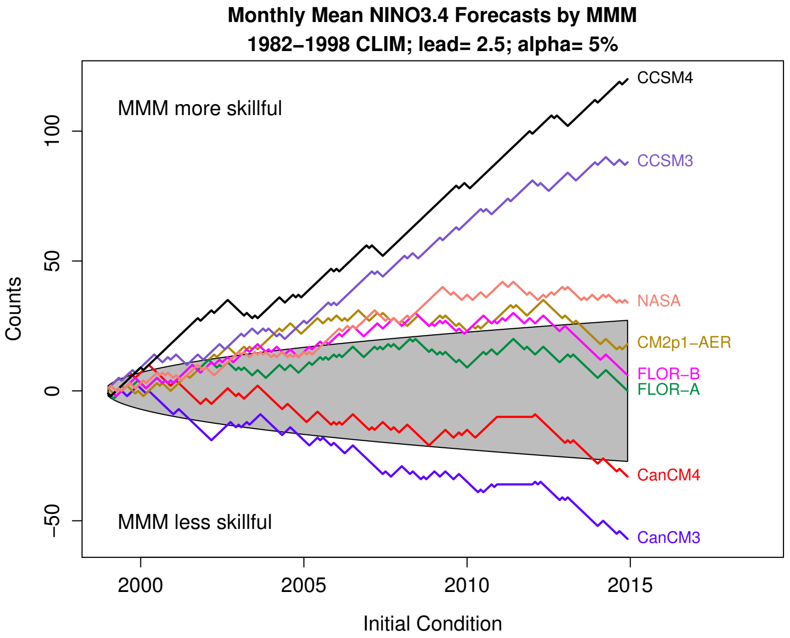
L. Marx and B. Huang

COLA, Calverton, Maryland

(Manuscript received 22 November 2011, in final form 31 January 2012)



Multimodel Mean

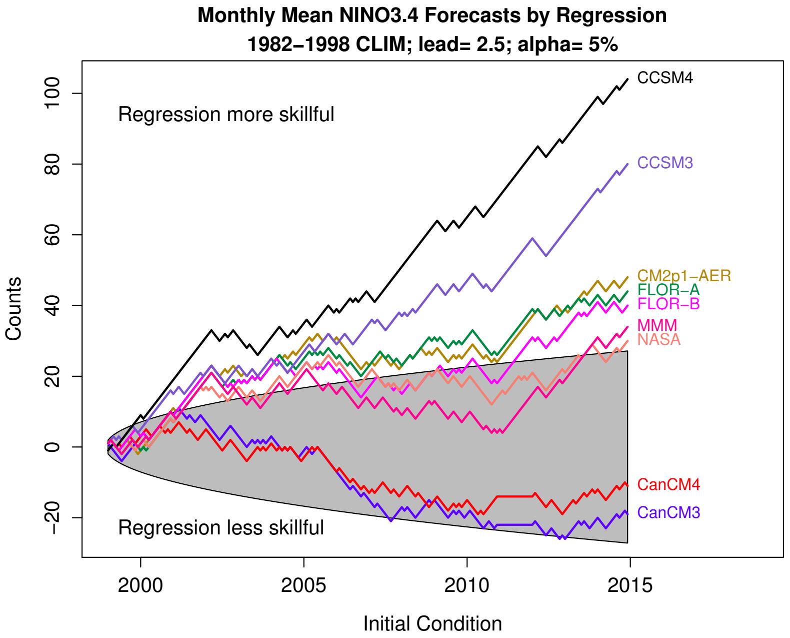


Monthly Mean NINO3.4 Forecasts by MMM
1982–1998 CLIM; lead= 2.5; alpha= 5%
MMM more skillful
MMM less skillful
Counts
100
50
0
−50
2000
2005
2010
2015
Initial Condition
CCSM4
CCSM3
NASA
CM2p1−AER
FLOR−B
FLOR−A
CanCM4
CanCM3

Statistical Prediction

$$\hat{T}_{m+\tau} = \hat{b}_{m,\tau} + \hat{a}_{m,\tau} T_m,$$

where $\hat{b}_{m,\tau}$ and $\hat{a}_{m,\tau}$ are least squares estimates of the slope and intercept estimated from independent data.

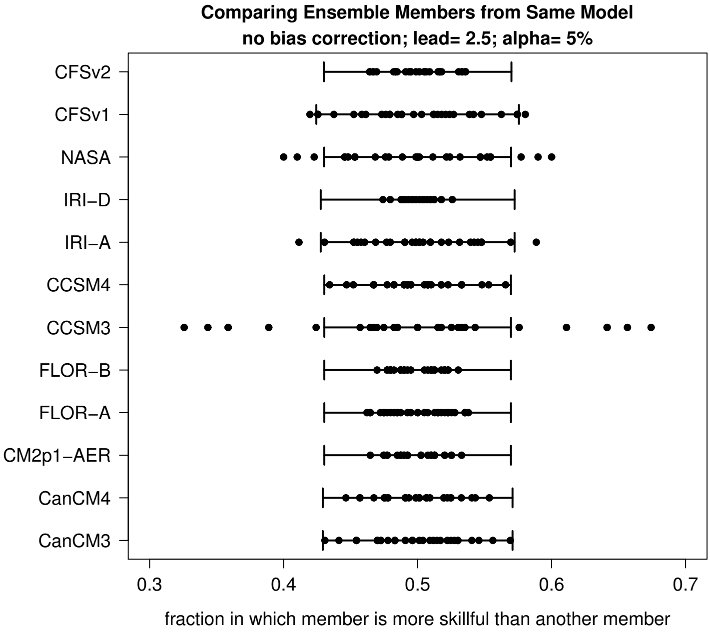


Monthly Mean NINO3.4 Forecasts by Regression
1982–1998 CLIM; lead= 2.5; alpha= 5%
Regression more skillful
Regression less skillful
CCSM4
CCSM3
CM2p1–AER
FLOR–A
FLOR–B
MMM
NASA
CanCM4
CanCM3
Counts
−20
0
20
40
60
80
100
2000
2005
2010
2015
Initial Condition

Exchangeability

Hypothesis: ensemble members exchangable.

Test: Compare skill of different ensemble members from same model.



Comparing Ensemble Members from Same Model
no bias correction; lead= 2.5; alpha= 5%
CFSv2
CFSv1
NASA
IRI−D
IRI−A
CCSM4
CCSM3
FLOR−B
FLOR−A
CM2p1−AER
CanCM4
CanCM3
0.3
0.4
0.5
0.6
0.7
fraction in which member is more skillful than another member

Strictly Exchangeable Not Strictly Exchangeable

CFSv1: Lagged ensemble for A (more widely spaced than CFSv2)

CFSv2: Lagged ensemble for A-L

NASA: some lagged ensemble, some breeding vectors

CCSM3: A-L-I initialized from different years in long control

CCSM4: Lagged ensemble for A, same I initialization as CCSM3

CanCM3: Different A-L-I-O initializations starting from different ICs

CanCM4: Different A-L-I-O initializations starting from different ICs

FLOR-A: Ensemble data assimilation

FLOR-B: Ensemble data assimilation

CM2p1-AER: Ensemble data assimilation

IRI-D: A-L initialized from AMIP runs

IRI-A: A-L initialized from AMIP runs

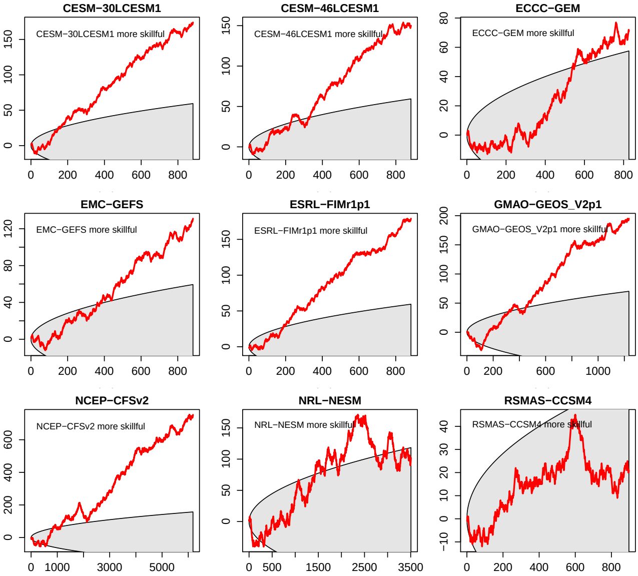
SubX Project

- 30+ day forecasts initialized each week.
- Hindcast Period: 1999-2015 (17 years).
- week 3-4 prediction (average from 15-28 day leads)
- contiguous U.S.
- pattern correlation

Model	Components	Ensemble Members	Length (Days)
NCEP-CFSv2	A,O,I,L	4	45
EMC-GEFS	A,L	11 [21]	35
ECCC-GEM	A,L	4 [21]	32
GMAO-GEOS5	A,O,I,L	4	45
NRL-NESM	A,O,I,L	4	45
RSMAS-CCSM4	A,O,I,L	3 [9]	45
ESRL-FIM	A,O,I,L	4	32

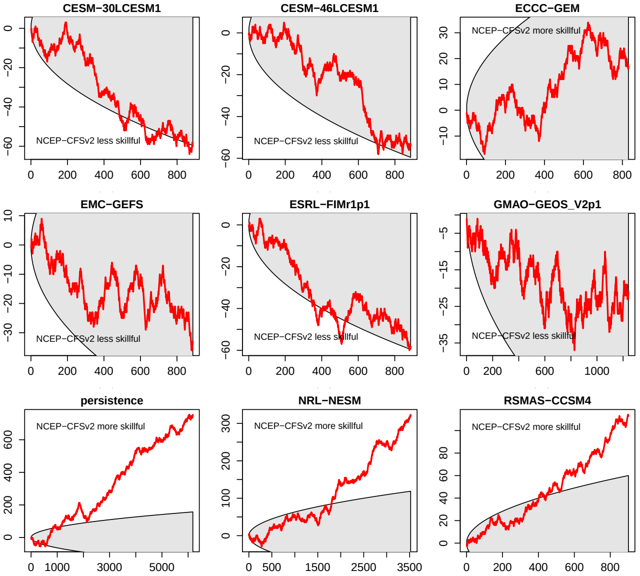
Compare to persistence forecast

Pattern Correlation of Week 3-4 Temperature Predictions



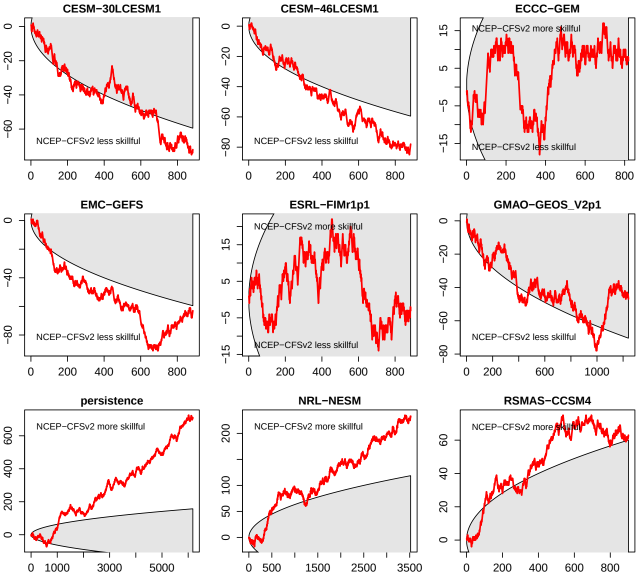
Compare to CFSv2 forecasts

Comparison to CFSv2



Precipitation

Comparison to CFSv2 Precipitation Forecasts



Summary

1. Skill measures computed on a common period or with a common set of observations are not independent.
2. Standard tests for differences in correlation or MSE are biased when evaluated over common period.
3. Random walk test avoids these problems and moreover applies to non-Gaussian distributions and arbitrary skill measures.
4. NMME: Canadian models are the most skillful dynamical models , even when compared to the multi-model mean.
5. NMME: A regression model is significantly more skillful than most other models.
6. NMME: There are significant skill differences between ensemble members from same model, reflecting differences from initialization.
7. SubX: Week 3-4 forecasts of Temp/Prec are more skillful than persistence forecasts.
8. SubX: CESM, EMC, ESRL, GMAO models more skillful than CFSv2 precipitation forecasts.