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?
Initialized vs. Uninitialized Forecasts

observations
Initialized vs. Uninitialized Forecasts

The graph illustrates the temperature anomaly (°C) over the years from 1880 to 2020 for both initialized and uninitialized forecasts. The uninitialized forecasts show a trend of increasing temperature anomaly with time, indicating a warming pattern that is not constrained by initial conditions. The initialized forecasts, represented by a blue line, show a smoother trend, suggesting that initial conditions play a significant role in the temperature predictions. The graph highlights the importance of initialization in climate modeling and forecasting.
Initialized vs. Uninitialized 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 \( \nu_1 \) and \( \nu_2 \) 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 $\nu_1$ and $\nu_2$ are the appropriate degrees of freedom.
Standard Tests Assume Forecast-Verification Pairs are **Independent**

![Diagram of forecast verification pairs and skill distribution]

- Distribution
  - (f,o)
  - (f,o)
  - (f,o)
  - (f,o)
  - (f,o)
  - (f,o)
  - Forecast A
  - Forecast B
  - Skill of A
  - Skill of B
For Model Comparisons, Forecast-Verification Pairs are Dependent

(o, fA, fB) pairs are dependent!
NMME skill estimates tend to be correlated in seasonal forecasting.
observation = signal + noise

forecast A = signal + noise* / \sqrt{E}

forecast B = signal + noise** / \sqrt{E}
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?
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

Similar Approaches in Weather Prediction


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.
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

Forecast A
more skill

Forecast B
more skill

Event

Time
Random Walk Test

A more skillful
Random Walk Test

A more skillful
Random Walk Test

A more skillful

2√N

−2√N

95% range for p = 1/2

mean

successes-fails

−15 0 5 10

0 10 20 30 40 50

time
North American Multi-Model Ensemble

- Hindcasts initialized every month from 1982-2010 (29 years)
- Lead 2.5 months
- MSE of NINO3.4
- Verification: OISST

<table>
<thead>
<tr>
<th>model</th>
<th>ensemble size</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMC1-CanCM3</td>
<td>10</td>
</tr>
<tr>
<td>CMC2-CanCM4</td>
<td>10</td>
</tr>
<tr>
<td>COLA-RSMAS-CCSM3</td>
<td>6</td>
</tr>
<tr>
<td>GFDL-CM2p1</td>
<td>10</td>
</tr>
<tr>
<td>NASA-GMAO</td>
<td>10</td>
</tr>
<tr>
<td>NCEP-CFSv1</td>
<td>10</td>
</tr>
<tr>
<td>NCEP-CFSv2</td>
<td>10</td>
</tr>
</tbody>
</table>
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%
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

Counts

Initial Condition

Counts

Initial Condition
Exchangeability

Hypothesis: ensemble members exchangeable.

Test: Compare skill of different ensemble members from same model.
Comparing Ensemble Members from Same Model

no bias correction; lead = 2.5; alpha = 5%

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.
- week 3-4 prediction (average from 15-28 day leads)
- contiguous U.S.
- pattern correlation

<table>
<thead>
<tr>
<th>Model</th>
<th>Components</th>
<th>Ensemble Members</th>
<th>Length (Days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCEP-CFSv2</td>
<td>A,O,I,L</td>
<td>4</td>
<td>45</td>
</tr>
<tr>
<td>EMC-GEFS</td>
<td>A,L</td>
<td>11 [21]</td>
<td>35</td>
</tr>
<tr>
<td>ECCC-GEM</td>
<td>A,L</td>
<td>4 [21]</td>
<td>32</td>
</tr>
<tr>
<td>GMAO-GEOS5</td>
<td>A,O,I,L</td>
<td>4</td>
<td>45</td>
</tr>
<tr>
<td>NRL-NESM</td>
<td>A,O,I,L</td>
<td>4</td>
<td>45</td>
</tr>
<tr>
<td>ESRL-FIM</td>
<td>A,O,I,L</td>
<td>4</td>
<td>32</td>
</tr>
</tbody>
</table>
Compare to persistence forecast
Pattern Correlation of Week 3-4 Temperature Predictions
Compare to CFSv2 forecasts
Comparison to CFSv2

SubX Model vs. NCEP−CFSv2

Feof week34 USMexico tas2m ALL

 CESM–30LCESM1

 CESM–46LCESM1

 ECCC–GEM

 EMC–GEFS

 ESRL–FIMr1p1

 GMAO–GEOS_V2p1

 persistence

 NRL–NESM

 RSMAS–CCSM4

 NCEP–CFSv2 less skillful

 NCEP–CFSv2 more skillful

 NCEP–CFSv2 less skillful

 NCEP–CFSv2 more skillful

 NCEP–CFSv2 less skillful

 NCEP–CFSv2 more skillful

 NCEP–CFSv2 less skillful

 NCEP–CFSv2 more skillful
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.