Evaluating Ensemble Seasonal Forecasts Using Information Metrics

Hannah Aizenman, Nir Krakauer, Michael Grossberg, Irina Gladkova

NOAA-CREST at CCNY

haizenm00@ccny.cuny.edu
How do we evaluate the accuracy of ensembles?

Observed Temperature in New York (40.48°N, 285.74°E) in August 1985
NCEP Forecast System Version 2 (CFSv2)

- Monthly mean 2-m temperature (K)
- Reanalysis observation: $x_o$
- Ensemble members 1-9, 0.5 month lag forecasts: $x_1, x_2, \ldots, x_k$
- 01/1979- 12/2009
- 1° x 1° spatial grid
Mean Error: $\mu(x_{1...k}) - x_o$ at (t,lat,lon)

All spreads have an error score of 0
Mean Error: $\mu (x_{1\ldots k}) - x_0$ at (t,lat,lon)

All spreads have an error score of 0

- skill is based on the error of the ensemble mean
- $\text{MSE} = \frac{\sum_{t=1}^{N} E_t^2}{N}$
- $\text{SS} = 1 - \frac{\text{MSE}_{\text{forecast}}}{\text{MSE}_{\text{ref}}}$
Negative Log Likelihood

Forecast PDF

$p(x)$

unlikely observation

Forecast NLL

$-\log(p(x))$

surprising occurrence

$x$
Negative Log Likelihood

\[ p(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

\[ -\log(p(x)) = \frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \log \sigma \]
Converting from Observation to PDF

- $p^f$ estimates $f_{i,t,\text{lat},\text{lon}}, \ i \ 1,2,\ldots,9$
- $p^c$ estimates $x_{j,\text{lat},\text{lon}},\ j \ 0,1,\ldots,t-1,$
Information Gain: $-\log_2 (p^f (x_0)) - - \log_2 (p^c (x_0))$
Aggregate IG: Normal PDF from Ensemble Members

[Graph showing the distribution of IG (bits) with a peak near 0, and a global mean time series from 1987 to 2007.]
Bias Correction Estimated from Previous Hindcasts

\[
bias_{i,t_k,la,ln} = \sum_{k=m/y_0}^{k<S} \frac{f_{i,t_k,la,ln} - obs_{t_k,la,ln}}{S - 1}
\]

s.t. k increments in steps of 12
Bias Correction Estimated from Previous Hindcasts

\[
bias_{i,t_k,la,ln} = \sum_{k=m/y_0}^{k<S} \frac{f_{i,t_k,la,ln} - obs_{t_k,la,ln}}{S - 1}
\]

s.t. k increments in steps of 12

- \( i = f_i \in forecasts \ [1, 9]\)
- \( k = month \in [1, 12], year \in [y_0, ..., y_f]\)
- \( t, la, ln = \) time, latitude, longitude
- \( S = \) current observation’s time-step
Improve Forecast by Subtracting the Bias

Temporal Mean

Global Mean
A Few Grid Points (e.g. in Arctic) have Extreme NLLs

High certainty in wrong prediction = spike in negative IG
Most Grid Points have Normal NLLs

Low certainty
= no spike
IG Improves by Using Climatology $\sigma$ Instead of Forecast
Global Mean IG

<table>
<thead>
<tr>
<th></th>
<th>TMP</th>
<th>PRATE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>median</td>
</tr>
<tr>
<td><strong>Forecast</strong></td>
<td>-4.9134</td>
<td>-0.4207</td>
</tr>
<tr>
<td><strong>Bias Corrected</strong></td>
<td>-0.8877</td>
<td>0.0964</td>
</tr>
<tr>
<td><strong>Climatology SD</strong></td>
<td>0.0558</td>
<td>0.0147</td>
</tr>
</tbody>
</table>

- IG improves after bias correction and substituting climatology SD for ensemble SD
- T reforecasts now average slightly better than climatology, P still worse
How does a normal PDF fit to an ensemble differ from a kernel density estimator?

(a) Normal Density Estimation  
(b) Kernel Density Estimation
Aggregate IG: KDE Approximation

Temporal Mean

Global Mean
Aggregate IG: Climatology KDE vs. Normal

The graph on the left shows the distribution of IG (information gain) in bits, with a peak at IG = 0 and decreasing counts as IG values increase. The y-axis represents the number of grid cells, and the x-axis represents IG in bits, ranging from -40 to 40.

The map on the right illustrates the temporal mean of IG, with different colors indicating varying IG values.

The bottom graph depicts the global mean of IG over time, from 1990 to 2006, with a fluctuating trend.
Aggregate IG: Forecast KDE vs. Normal
<table>
<thead>
<tr>
<th></th>
<th>TMP mean</th>
<th>TMP median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forecast</td>
<td>-4.9134</td>
<td>-0.4207</td>
</tr>
<tr>
<td>Bias Corrected</td>
<td>-0.8877</td>
<td>0.0964</td>
</tr>
<tr>
<td>Climatology SD</td>
<td>0.0558</td>
<td>0.0147</td>
</tr>
<tr>
<td>KDE</td>
<td>0.3587</td>
<td>0.1326</td>
</tr>
</tbody>
</table>

- IG improves even more using KDE
- KDE is sensitive to spread-spikes
Conclusion

- Gaussian estimation vs. kernel density estimation for converting ensemble forecasts to a PDF
- Treat ensemble as probabilistic predictions
- IG measures probabilistic predictions well
- Forecast can be improved by using IG to diagnose problems (e.g. mean and SD offsets)
Future Work

- Fit other distributions to climatology and ensemble forecasts
- Include trend estimates in probabilistic forecasts
- Develop predictive model that incorporates climatology and multiple model forecasts based on skill at a given region and season
The authors gratefully acknowledge support from NOAA under Grants NA11SEC4810004 and NA12OAR4310084. All statements made are the views of the authors and not the opinions of the funding agency or the US government.
