Ensemble Processing

David Unger
Climate Prediction Center
Schematic illustration
Schematic example
DEFINITIONS:

F = Forecast  C = Climatology  b = Observation

S = Standard deviation of forecast errors

\[ S^2 = (F - b)^2 \]

\[ \sigma_f^2 = (F - C)^2 \]

\[ \sigma_b^2 = (b - C)^2 \]

E = Ensemble Spread

\[ E^2 = \frac{\sum (F - F_m)^2}{N} \]
Regression = Bias Correction and Standardization

\[ Z = \frac{(F - \bar{F})}{\sigma_F} \]

\[ \hat{Z} = RZ \]

\[ \hat{F} = \hat{Z} \sigma_C + C \]
Regression relationships

\[ \sigma_{fm} = R_{fm} \sigma_c \quad \text{Standard deviation of forecasts} \]

\[ S_{fm} = \sigma_c \sqrt{1 - R_{fm}^2} \quad \text{Standard deviation of forecast errors} \]

\[ \sigma_{fm}^2 + S_{fm}^2 = R_{fm}^2 \sigma_{fm}^2 + \sigma_c^2 \left( 1 - R_{fm}^2 \right) \]

\[ \sigma_{fm}^2 + S_{fm}^2 = \sigma_c^2 \quad \text{Forecast and error variance sum to climatological variance} \]
$F = 0.5$

The graph shows two distributions with the following parameters:

- Red line: $R = 0.5, S = 0.866$
- Black line: Climo

The $Z$ value axis ranges from $-4$ to $4$. The peak of the red line is at $Z = 0$, indicating the mean of the distribution.
Schematic example
Kernel Math

- Ensemble Spread, E, and $S_z$, are constrained.
  - Depend on R and $R_m$
  - Also depends on the standard deviation of f and $f_m$. 
Variance Relationships

• Explained Variance = $R^2 \sigma_c^2$
• Unexplained Variance $S_{fm}^2$ divides into 2 parts:
  1) Ensemble Spread ($E^2$) (Variable)
  2) Residual Variance (Fixed) ($S_z^2$)

$$S_z^2 \leq S_{fm}^2 : S_{fm}^2 = S_z^2 + E^2$$
$$E^2 = \sum (f - fm)^2$$

$$S_z = \text{Conditional standard dev.}$$

$$S_f^2 = S_{fm}^2 + E^2$$

$$\sigma^2_{\text{Total}} = \sigma^2_c = \sigma^2_f + S_z^2$$

$$S_z^2 + E^2 = S_{fm}^2$$

$$S_z^2 = S_{fm}^2 - E^2$$
Optimum Spread

• Regression gives one candidate for an optimum spread value, $E^2$
• $E^2$ depends on the skill difference between the individual ensemble members and the ensemble mean.
• We then can calculate the unexplained variance not accounted for by the spread.
Optimum ensemble spread?

\[ S_f^2 = S_{fm}^2 + E^2 \]
\[ E^2 = S_f^2 - S_{fm}^2 \]
\[ S_f = \sigma_c \sqrt{(1 - R_f^2)} \]
\[ S_{fm} = \sigma_c \sqrt{(1 - R_{fm}^2)} \]
\[ E^2 = \sigma_c^2 (R_{fm}^2 - R_f^2) \]
Minimum Forecast Error Variance

\[ E^2 = \sigma_c^2 \left( R_{fm}^2 - R_f^2 \right) \]

\[ S_z^2 = S_{fm}^2 - E^2 \]

\[ S_z^2 = \sigma_c^2 \left( 1 - R_{fm}^2 \right) - \sigma_c^2 \left( R_{fm}^2 - R_f^2 \right) \]

\[ S_z^2 = \sigma_c^2 \left( 1 - 2 R_{fm}^2 + R_f^2 \right) \]
Effective Correlation

\[ S_z^2 = \sigma_c^2 \left( 1 - 2 R_{fm}^2 + R_f^2 \right) \]

\[ S_z^2 = \sigma_c^2 \left( 1 - R_z^2 \right) \]

\[ R_z^2 = 2 R_{fm}^2 - R_f^2 \]
Forecast computation

\[ Z_i = \frac{(F_i - \bar{F})}{\sigma_F} \]

\[ \hat{Z}_i = R_z Z \]

\[ F_i = Z_i \sigma_c + C \]

\[ \hat{F}_i = (F_i - F_m) \frac{E}{E} + F_m \]
Kernel vs. Mean
$R_z = 0.97$, $R_{fm} = 0.94$, $R_i = 0.90$
$R_z = 0.93, \ R_{fm} = 0.87, \ R_f = 0.30$
$R_z = .85, \ R_{fm} = .67, \ R_i = .41$
\( R_Z = 0.62, \ R_{fm} = 0.46, \ R_f = 0.20 \)
Weighting

\[ w_i = \frac{R_i}{(1 - R_i)} \]

\[ wt_i = \frac{w_i}{\left( \sum (w_i) \right)} \]

\[ R = .9: \quad 9 = \frac{.9}{(1 - .9)} \quad , \quad R = .8: \quad 5 = \frac{.8}{(1 - .8)} \]

\[ 9 + 5 = 14 \]

\[ .64 = \frac{9}{14} \quad .36 = \frac{5}{14} \]
3 members + 1 forecast
3 members half wt. + 1fcst
Nino 3.4 Consolidation
SST 3.4 Consolidation
6- Mo. Lead, Rz=.63,Rm=.54,R=.44
## Nino 3.4 SST

### 6-Mo. Lead, 1981-2003, All Start times

<table>
<thead>
<tr>
<th>Model</th>
<th>CRPSS</th>
<th>MAE (C)</th>
<th>Bias (C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFS</td>
<td>.293</td>
<td>.541</td>
<td>-.058</td>
</tr>
<tr>
<td>CFS Mean</td>
<td>.133</td>
<td>.677</td>
<td>.061</td>
</tr>
<tr>
<td>CA</td>
<td>.339</td>
<td>.526</td>
<td>.019</td>
</tr>
<tr>
<td>CA Mean</td>
<td>.296</td>
<td>.559</td>
<td>-.067</td>
</tr>
<tr>
<td>CON</td>
<td>.324</td>
<td>.380</td>
<td>-.016</td>
</tr>
<tr>
<td>CON Mean</td>
<td>.349</td>
<td>.366</td>
<td>.022</td>
</tr>
</tbody>
</table>
Reliability CFS
Reliability CA

A graph showing the relationship between Observed Frequency and Forecast Probability. The graph includes three lines:
- CA (red)
- CA Mean (green dotted)
- Perfect (black)

The axes are labeled:
- OBS FREQUENCY on the y-axis
- FORECAST PROBABILITY on the x-axis
Reliability Consolidation

![Reliability Consolidation](image)
Requirements

1. Forecast Mean
2. Standard deviation of individual members
3. Standard deviation of individual member errors
4. Standard deviation of ensemble mean
5. Standard deviation of ensemble mean errors
6. Observed mean
7. Standard deviation of observations
8. (Forecast anomaly * observed anomaly)
9. (Ensemble mean anomaly * obs anomaly)
10. Ensemble Spread
Advantages and disadvantages

Kernel Method

Advantages:
- Uses all ensemble information
- One equation set produces forecasts for many thresholds.
- Handles irregular distributions.

Disadvantages
- Often is very close to a simple ensemble mean
- Can’t handle regime dependent bias
Finish Line