CONDITIONAL DEPENDENCE and SUFFICIENT STATISTICS of an ENSEMBLE

By

Roman Krzysztofowicz
University of Virginia

Presented at the
19th Conference on Probability and Statistics in the Atmospheric Sciences
New Orleans, Louisiana

20–24 January 2008

Acknowledgments:
Work supported by the National Science Foundation under Grant No. ATM–0641572.
Collaboration of Zoltan Toth, Environmental Modeling Center NOAA / NWS / NCEP.
DATA

Location: Savannah, GA

Predictand: 2m temperature

Forecast time: 00z

Lead time: 108h (12h, 156h)

Samples

• Climatic: January 1959 – December 1998 (40 years)
• Joint: January 2005 – December 2005 (1 year)
  forecast (HR control, LR control, 10 members)
  predictand
LIKELIHOOD FUNCTION

1. Standardization: stationarity, ergodicity
2. Marginal distributions of ensemble members summary statistics
3. Normal Quantile Transform (NQT)
4. Predictive regressions search for: sufficient statistics dependence structure
STANDARDIZATION

Purpose: to obtain time series that are
• stationary
• ergodic

1. Climatic sample of predictand
2. Joint sample of ensemble and predictand
Joint Sample

• January 2005 – December 2005

• For each day \( k \ (k=1,\ldots,365) \) given

\[ m_k \] – prior (climatic) mean of predictand
\[ s_k \] – prior (climatic) standard deviation of predictand

• Standardized ensemble member

\[
y_j'(k) = \frac{y_j(k) - m_k}{s_k} \quad j = 0, 1, \ldots, 10
\]
\[
k = 1, \ldots, 365
\]
Ensemble Members: Standardized Data
Savannah. 108h
MARGINAL DISTRIBUTIONS

• Catalog of 43 parametric families of distributions of continuous variates

• Uniform estimation method
Empirical and Parametric Distribution Functions: Savannah. 108h. Observed (First 6 Mo, N=176)
Empirical and Parametric Distribution Functions: Savannah. 108h. **Plus0** (First 6 Mo, N=176)

Log-Weibull
MAD: 0.0397
$\alpha = 1.8333$
$\beta = 11.7937$
$\eta = -5.0$
Empirical and Parametric Distribution Functions: Savannah. 108h. **Plus2** (First 6 Mo, N=176)

Log-Weibull

MAD: 0.0262

\(\alpha = 1.8507\)

\(\beta = 11.1741\)

\(\eta = -5.0\)
Empirical and Parametric Distribution Functions: Savannah. 108h. **Plus3** (First 6 Mo, N=176)

Log-Weibull
MAD: 0.0502
\( \alpha = 1.8440 \)
\( \beta = 10.8510 \)
\( \eta = -5.0 \)
Empirical and Parametric Distribution Functions: Savannah. 108h. **Minus2** (First 6 Mo, N=176)

Log-Weibull
MAD: 0.0314
$\alpha = 1.8424$
$\beta = 14.0569$
$\eta = -5.0$
Parametric Distribution Functions of 11 Standardized Ensemble Members: Savannah.108h. (First 6 Mo, N=176)
Normal Quantile Transform (NQT)

\[ v = Q^{-1}(G'(w')) \]

\[ z_j = Q^{-1}\left(\bar{K}_j'(y'_j)\right) \quad j = 0, 1, \ldots, 10 \]

- \( G' \) – prior (climatic) distribution of \( W' \)
- \( \bar{K}_j' \) – marginal (initial) distribution of \( Y'_j \)
- \( Q^{-1} \) – inverse of the standard normal distribution
**DEPENDENCE STRUCTURE**

- **Independence if**

  \[ \kappa(y_0, y_1, \ldots, y_J) = \prod_{j=0}^{J} \kappa_j(y_j) \]

  \[ \text{No: } 0.42 < \text{Rank Cor}(Y_i', Y_j') < .80 \quad i \neq j, \quad j = 0, 1, \ldots, 10 \]

  \[ \text{Warning: } [0.78 < \text{Rank Cor}(Y_i, Y_j) < .94] \text{ spurious (non-ergodicity)} \]

- **Conditional Independence if**

  \[ f(y_0, y_1, \ldots, y_J|w) = \prod_{j=0}^{J} f(y_j|w) \]

  **Factorization Theorem:**

  \[ f(y_0, y_1, \ldots, y_{10}|w) \approx f_8(y_8|y_0, w) f_2(y_2|y_0, w) f_0(y_0|w) \]

  \[ \text{No: } Y_8, Y_2 \text{ are independent, conditional on } W \text{ and } Y_0 \]

  \[ Y_1, Y_3, Y_4, Y_5, Y_6, Y_7, Y_9, Y_{10} \text{ are extraneous for } W, \text{ given } Y_0, Y_2, Y_8 \]
INFORMATIVENESS OF MEMBERS

<table>
<thead>
<tr>
<th></th>
<th>First 6 Months</th>
<th></th>
<th>Second 6 Months</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Day 3</td>
<td>Day 5</td>
<td>Day 7</td>
</tr>
<tr>
<td></td>
<td>$j$</td>
<td>IS</td>
<td>$j$</td>
</tr>
<tr>
<td><strong>Best</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>.72</td>
<td>.53</td>
<td>.33</td>
</tr>
<tr>
<td>2</td>
<td>.71</td>
<td>.45</td>
<td>.30</td>
</tr>
<tr>
<td>10</td>
<td>.69</td>
<td>.43</td>
<td>.30</td>
</tr>
<tr>
<td>4</td>
<td>.66</td>
<td>.43</td>
<td>.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Worst</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>.62</td>
<td>.29</td>
<td>.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Comb.</strong></td>
<td>0, 2, 3</td>
<td>.75</td>
<td>0, 2, 8, 4</td>
</tr>
</tbody>
</table>

Conclusions:  
(1) 3–4 ensemble members contain all information there is.  
(2) The membership in the best combination varies: lead time, season, sample.  
(3) The best combination method only for “sophisticated” users: adaptive BPE.
## INFORMATIVENESS OF STATISTICS

<table>
<thead>
<tr>
<th></th>
<th>First 6 Months</th>
<th>Second 6 Months</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Day 3</td>
<td>Day 5</td>
</tr>
<tr>
<td><strong>Comb.</strong></td>
<td>IS</td>
<td>IS</td>
</tr>
<tr>
<td>Mean</td>
<td>.75</td>
<td>.59</td>
</tr>
<tr>
<td>Mean / 0</td>
<td>.75</td>
<td>.56</td>
</tr>
<tr>
<td>HR</td>
<td>.77</td>
<td>.59</td>
</tr>
<tr>
<td>HR, Mean / 0</td>
<td>.77</td>
<td>.56</td>
</tr>
<tr>
<td>HR, Members</td>
<td>2, 7, 0</td>
<td>2, 1</td>
</tr>
</tbody>
</table>

Conclusions:  
1. The combination (HR, Mean / 0) is near optimal.  
2. The LR control is extraneous, given Mean / 0.  
3. The best combination method only for “sophisticated” users: adaptive BPE.