Statistical transformation of dynamical model output

Henrik Feddersen
Danish Meteorological Institute

IRI Advanced Training Institute on Climate Variability and Food Security, 11-12 July 2002
Outline

- Introduction
- Methods for finding coupled climate patterns
  - EOF/principal components analysis
  - Singular value decomposition analysis
  - Canonical correlation analysis
- Transformation of model output
- Cross-validation
- Further postprocessing
- Transformation of ensemble forecasts
- Downscaling in space
Rainfall example

Simulated rainfall, JFM 1998

Arpege atmospheric GCM forced by observed SST; horizontal resolution approx. 2.8°×2.8°

Observed rainfall, JFM 1998

Gridded observations (Xie-Arkin); horizontal resolution 2.5°×2.5°
Rainfall anomalies

Simulated rainfall anomaly, JFM 1998

Observed rainfall anomaly, JFM 1998

Anomaly wrt. model climatology (1979-99)

Anomaly wrt. observed climatology (1979-99)
Rainfall in selected grid box
Skill of model simulated rainfall

Correlations between simulated and observed rainfall, JFM 1979–99
ENSO rainfall index

ENSO index = Rain[II] – Rain[I]
Model simulated ENSO index and observed rainfall anomaly
Correlation between ENSO index and observed rainfall
So...

- The dynamical model simulates large-scale variability (the ENSO rainfall index) well!
- Grid box rainfall can be specified accurately from the simulated large-scale index (~downscaling)
- In general, we need to find large-scale indices that are well simulated by the model and that correlate with grid box-scale observed rainfall...!
Empirical Orthogonal Functions (EOFs) and Principal Components

For a selected region find the ”best” index, \( u(t) \), from which the original field, \( x(t) \) can be reconstructed as \( x^*(t) = u(t) \cdot p \).

That is, find a pattern \( p \)
- such that the time series \( u(t) \) is given by the projection of the field onto the pattern, i.e. \( u(t) = x(t) \cdot p \)
- such that the sum of the squared difference between the reconstructed and the original fields
  \[ \sum_k \sum_t = [x_k(t) - x_k^*(t)]^2 \] (proportional to the ”unexplained” variance, \( \text{Var}[x - x^*] \) ) is minimized.
- That is, the explained variance 
  \[ (\text{Var}[x] - \text{Var}[x - x^*]) / \text{Var}[x] \] is maximized.
Empirical Orthogonal Functions (EOFs) and Principal Components

Solution:

- $\mathbf{p}$ (the first EOF) is the eigenvector associated with the largest eigenvalue of the covariance matrix of $\mathbf{x}(t)$.
- The largest eigenvalue is equal to the variance of the index (the first principal component) $u(t)$.
- The eigenvectors (EOFs) and principal components associated with the secondary eigenvalues ”explain” the maximal variance under the constraint that the EOFs be orthogonal to each other.
- The $n$’th largest eigenvalue is equal to the variance of the $n$’th principal component.
EOFs and principal components

Reconstruction:

\[ x^*(t) = \sum_j u_j(t) p_j \]

The terms that explain only small fractions of the total variance can be ignored without much loss of information.
First EOFs

Eigenvector scaled such that map shows correlations between principal component and simulated rainfall.
First principal components
Second EOFs

2nd EOF of simulated JFM rainfall, 1979–99

2nd EOF of observed JFM rainfall, 1979–99
Second principal components
Explained variance
Singular Value Decomposition (SVD) Analysis

- Find patterns and time series that maximize the fraction of explained covariance between the two fields.
- Patterns are obtained as the singular vectors of an SVD of the cross-covariance matrix.
- Time series are obtained by projection of the two fields onto the respective patterns.
First pair of SVD patterns

Map shows correlations between SVD time series (simulations) and simulated rainfall: Homogeneous correlation map.
First pair of SVD time series
Second pair of SVD patterns
Second pair of SVD time series
Squared covariance fraction
Canonical Correlation Analysis (CCA)

- Find patterns and time series that maximize the correlation between the time series.
- Prefiltering of data fields is necessary
  - Use first few principal components
  - Number of principal components << length of time series
  - Include principal components until a certain fraction of the total variance is explained
  - Include principal components until the associated eigenvalues no longer are significantly different
Canonical correlation analysis

1st CCA pattern, simulated rainfall [3 PCs]

1st CCA pattern, observed rainfall [3 PCs]
Canonical correlation analysis

1st CCA pattern, simulated rainfall [9 PCs]

1st CCA pattern, observed rainfall [9 PCs]

First CCA time series of JFM rainfall [9 PCs]

1st CCA time series, sim. 1st CCA time series, obs.
Canonical correlation analysis

2nd CCA pattern, simulated rainfall [3 PCs]

2nd CCA pattern, observed rainfall [3 PCs]

Second CCA time series of JFM rainfall [3 PCs]
Canonical correlation analysis

2nd CCA pattern, simulated rainfall [6 PCs]

2nd CCA pattern, observed rainfall [6 PCs]

Second CCA time series of JFM rainfall [6 PCs]

2nd CCA time series, sim.
2nd CCA time series, obs.
Canonical correlation analysis

2nd CCA pattern, simulated rainfall [9 PCs]

2nd CCA pattern, observed rainfall [9 PCs]

Second CCA time series of JFM rainfall [9 PCs]
## Canonical correlation analysis

### First CCA mode

<table>
<thead>
<tr>
<th>PCs</th>
<th>Expl. var (sim)</th>
<th>Expl. var (obs)</th>
<th>Canonical correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>31.6%</td>
<td>40.4%</td>
<td>0.95</td>
</tr>
<tr>
<td>4</td>
<td>31.1%</td>
<td>40.2%</td>
<td>0.95</td>
</tr>
<tr>
<td>5</td>
<td>28.8%</td>
<td>31.8%</td>
<td>0.98</td>
</tr>
<tr>
<td>6</td>
<td>29.3%</td>
<td>33.5%</td>
<td>0.98</td>
</tr>
<tr>
<td>7</td>
<td>24.7%</td>
<td>30.3%</td>
<td>0.99</td>
</tr>
<tr>
<td>8</td>
<td>28.6%</td>
<td>35.1%</td>
<td>0.99</td>
</tr>
<tr>
<td>9</td>
<td>23.4%</td>
<td>28.6%</td>
<td>1.00</td>
</tr>
</tbody>
</table>
### Canonical correlation analysis

#### Second CCA mode

<table>
<thead>
<tr>
<th>PCs</th>
<th>Expl. var (sim)</th>
<th>Expl. var (obs)</th>
<th>Canonical correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>16.1%</td>
<td>18.7%</td>
<td>0.80</td>
</tr>
<tr>
<td>4</td>
<td>14.8%</td>
<td>18.2%</td>
<td>0.87</td>
</tr>
<tr>
<td>5</td>
<td>15.7%</td>
<td>20.5%</td>
<td>0.92</td>
</tr>
<tr>
<td>6</td>
<td>15.1%</td>
<td>18.2%</td>
<td>0.95</td>
</tr>
<tr>
<td>7</td>
<td>16.3%</td>
<td>20.5%</td>
<td>0.97</td>
</tr>
<tr>
<td>8</td>
<td>12.3%</td>
<td>19.3%</td>
<td>0.97</td>
</tr>
<tr>
<td>9</td>
<td>6.8%</td>
<td>9.1%</td>
<td>0.99</td>
</tr>
</tbody>
</table>
From the “training” period we have

- EOF/SVD/CCA time series, \( u_j(t) \);
- EOF/SVD/CCA patterns/weight vectors, \( p_j \);
- linear regression coefficients, \( A_j \) for regression of time series on the observed field, i.e.

\[
y^*(t) = \sum_j A_j u_j(t).
\]

Transformation of model output in “forecast” period:
Transformation of model output

Transformation of model output, \( x \), in "forecast" period:

1. \( u_j(t_f) = x(t_f) \cdot p_j \)

2. \( y^*(t_f) = \sum_j A_j u_j(t_f) \).
Cross-validation

- Training and validation data must be independent!
- Need long training period!
- Need long validation period!

1. Withhold data from one point in time!
2. Use remaining data as training data!
3. Make "forecast" for the one point in time!
4. Repeat steps 1-3 until an independent forecast has been made for every point in time!
Cross-validated transformation of model-simulated rainfall based on first SVD mode.
Cross-validated skill of transformed model simulation

For comparison... the skill of the raw model output
...based on the first two SVD modes

Correlations between hindcast and observed rainfall, JFM 1979–99

Correlations between hindcast and observed rainfall, JFM 1979–99

Correlations between simulated and observed rainfall, JFM 1979–99

First SVD mode

Raw model output
...based on more SVD modes

Three modes

Seven modes

Nine modes
Further postprocessing

Transformation based on three SVD modes

\[ \frac{\text{Var[transformed sim.]} }{\text{Var[obs]} } = 85\% \]

Inflate transformed simulation...
Inflation of transformed model output

\[
\text{Var[transformed sim]} / \text{Var[obs]} = 10\%
\]

\[
\text{Var[inflated]} / \text{Var[obs]} = 100\%
\]
Transformation of dynamical model ensemble simulations

- Skill of nine-member ensemble mean (raw model output)

- Skill of the first member of the ensemble
Transformation of dynamical model ensemble simulations

- Cross-validated skill of SVD transformed ensemble mean (three SVD modes)

- Cross-validated skill of first SVD transformed ensemble member
Rainfall ensemble simulation in single grid box
Ensemble "Capture Rate"

Nine member ensemble:

Expected "Capture Rate" = 80%

≈ 17 / 21
Transformed ensemble simulation in single grid box

Capture Rate = 10 / 21
Ensemble inflation

\[ \text{Var}[\text{total}] = \text{Var}[\text{ensemble mean}] + \text{Var}[\text{ensemble}] \]

\[ \text{Var}[\text{observed}] = \text{Var}[\text{total}] + \text{Var}[\text{unexplained}] \]

\[ \text{Var}[\text{inflated ensemble}] = \text{Var}[\text{ensemble}] + \text{Var}[\text{unexplained}] \]

\[ \Rightarrow \text{Var}[\text{total, inflated}] = \text{Var}[\text{obs}] \]
Inflation factor, $a$

- **Inflated ensemble members**
  \[ y_{\text{inflated},j} = y_{\text{em}} + a (y_j - y_{\text{em}}) \]

- **Inflation factor**
  \[ a^2 = \frac{\text{Var}[\text{obs}] - \text{Var}[\text{ens. mean}]}{\text{Var}[\text{ensemble}]} \]
Ensemble inflated rainfall in single grid box

Capture Rate = 17 / 21
Transformed rainfall ensemble simulation in low-skill grid box

Capture Rate = 9 / 21
Transformed rainfall ensemble simulation in low-skill grid box

Capture Rate = 20 / 21
Downscaling

Statistical

Model Output Statistics (MOS)

Training: Model ◯ Obs

Application: Model ◯ Obs

+ Correction for model error

– Need long model hindcast period

– Downscaling related to *average* model skill
Downscaling

- Statistical
- Perfect prognosis

Training: Obs ○ Obs
Application: Model ○ Obs

+ No need for long model hindcast period
- No correction for model error
Downscaling

Dynamical

- High resolution regional climate model nested into global model
  - No need for training period
  - Can forecast extreme events
  - Model climate need not be stationary
    - Computationally expensive
    - Systematic errors
Summary

- Statistical transformation based on large-scale (EOF/SVD/CCA) patterns of dynamical model output can improve forecast skill.

- The same sort of transformation (MOS) can be applied to spatially downscale dynamical model output.

- Use of model ensembles further improves forecast skill and allows for probabilistic forecast information.

- Further postprocessing is required in order to improve the reliability of ensemble forecasts.