Statistical downscaling daily rainfall statistics from seasonal forecasts using canonical correlation analysis or a hidden Markov model?

Andrew W. Robertson  
*International Research Institute for Climate and Society (IRI), Palisades, NY*

Koen Verbist  
*Water Centre for Arid Zones in Latin America and the Caribbean, La Serena (CAZALAC), Chile*
Outline

- **Goal:** downscale retrospective GCM seasonal forecasts of precipitation to a network of rainfall stations over central Chile, to predict daily rainfall statistics pertinent to dry-land management.

- **Targeted statistics:** winter season rainfall total, rainfall frequency, mean daily intensity; drought indices: number of heavy rainfall days, (daily) accumulated precipitation deficit.

- **2 Methods:**
  1. Linear regression of *seasonal statistics* using CCA
  2. Hidden Markov model (HMM) of *daily rainfall sequences*
Experiment design

- **Obs data:** 42 daily rainfall stations in central Chile, May–August, 1981–2005
- **Seasonal forecasts:** CFS (T63) of May–August gridded precip [20°–40°S, 65°–85°W], initialized on April 1

**Rainfall seasonality**

- A) Average number of wet days
- B) Average rainfall intensity (mm/day)

**Rainfall vs ENSO**

- Niño 3.4 Index
- 1940 to 2000
Multiple linear regression via Canonical correlation analysis (CCA)

- Regress seasonal average observed rainfall fields $y$ onto GCM f’cast fields $x$, $y = Ax + \varepsilon$

- Expand $x$ and $y$ in truncated Principal Component time series $V_x$ and $V_y$, and standardize the PCs

- The singular value decomposition $V_y^T V_x = \text{RMS}^T$ identifies linear combinations of the observation and predictor PCs with maximum correlation and uncorrelated time series (Barnett and Preisendorfer, 1987)

- these new pattern-variables give a diagonal regression matrix whose coefficients are correlations: $(V_y R) = M (V_x S)$

- The CCA modes with low correlation are neglected
Leading canonical modes

with 4 X- and 5 Y-PCs
Pearson correlation skill using CCA

Leave-5-out cross-validation

\[
\text{seas. amount} = (\text{no. of wet days}) \times (\text{mean intensity on wet days})
\]
Station: LA_TORRE 30.62S, 71.

### Continuous measures:
- Pearson's correlation: 0.6640
- Spearman's correlation: 0.7123
- 2AFC score (continuous): 76.33%
- % variance: 0.44%
- Variance ratio: 0.6499
- Mean bias: 1.58
- Root mean squared error: 68.15
- Mean absolute error: 50.00

### Categorical measures:
- Hit score: 60.00%
- Hit skill score: 40.00%
- LEPS score: 50.63%
- Gerrity score: 49.31%
- 2AFC (forecast categories): 80.92%
- 2AFC (continuous forecasts): 83.57%
- ROC area (below-normal): 0.9444
- ROC area (above-normal): 0.7847
Hidden Markov model of daily rainfall

Simplify the joint probability distribution of all the daily data (42 stations, 123 days, 25 years) by introducing a latent discrete state variable $S_t$ with a Markovian daily time dependence:

$$p (R_{1:T}, S_{1:T}) = \left[ p(S_1) \prod_{t=2}^{T} p(S_t | S_{t-1}) \right] \left[ \prod_{t=1}^{T} p(R_t | S_t) \right]$$

Graphical model:
Non-homogenous HMM conditioned on GCM forecasts

The means and maps obtained from the two different methods were found to be quite similar. The fact that an alternative clustering methodology such as K-means, which uses no information about temporal ordering of the rainfall measurements, produces state descriptions that are qualitatively similar to those produced by the HMM, suggests that these states are an inherent property of the data and insensitive to the particular modeling methodology being used.

In summary, daily rainfall states \( s \) and \( t \) identified by the HMM are associated with well-known patterns of interannual variability in winds, OLR, and SST. These associations provide a basis for the downscaling of seasonal GCM predictions and this is pursued in the following section.

5 A Non-Homogeneous HMM Downscaling Prototype

The NHMM generalizes the homogeneous HMM in that the transition probabilities in Equation 1 are allowed to vary with time. In particular, for downscaling applications, the transition probabilities between states are allowed to vary as a function of external input variables. Hughes and Guttorp introduced this model in the context of modeling rainfall occurrence. The NHMM used in this paper is based on this original work of Hughes and Guttorp with some minor modifications.

In this section we illustrate the ability of an NHMM to downscale atmospheric GCM simulations over NE Brazil. It is found that introducing atmospheric input variables does not visibly change the appearance of the state composites, nor appreciably change the rainfall probabilities. Thus, a vostate model is chosen for consistency with the HMM in the previous section.

For demonstration purposes and for consistency with IRI’s current seasonal forecast scheme, we define the inputs to the NHMM from the GCM’s simulated seasonal mean rainfall anomaly. The daily values needed as inputs to the NHMM are derived by simply repeating the seasonal mean input value for each day within the FMA season.

5.1 The Non-Homogeneous Hidden Markov Model

Let \( X_t \) be a \( D \)-dimensional column vector of predictors for day \( t \) in derived, for example, from a GCM. By \( X_1:T \) we will denote the sequence \( X_1, \ldots, X_T \). We now replace Equation 1 in the homogeneous HMM with:

\[
P(i|S_{t-1} = j, X_1:T) = \frac{\exp(\sigma_{ji} + \rho_i' x)}{\sum_{k=1}^K \exp(\sigma_{jk} + \rho_k' x)}.
\]

- use seasonal mean PC-1 of GCM precip [5S–40S, 100W–50W] repeated daily
- 15 CFS members * 10 stochastic realizations = 150 simulations
- 4-state model
NHMM rainfall states
Daily state sequence

Sea level pressure anomaly composites

State 4:
2.7% days
56% total rainfall
0.52 ENSO cor.
Pearson correlation skill using NHMM

(a) Seasonal Rainfall Amount  (b) Rainfall Frequency  (c) Rainfall Intensity

Leave-3-out cross-validation
Summary comparison of correlation skill via CCA
via NHMM
Drought indices computed from NHMM daily rainfall sequences
CLIMATE PREDICTABILITY TOOL
Evaluating seasonal climate
Designed for MOS application

Copyright 2002-2009 International Research Institute for Climate and Society

http://iri.columbia.edu/climate/tools
Summary

- **canonical correlation analysis (CCA)** is employed to the time series of the targeted seasonal statistic, calculated from daily station rainfall observations, together with the GCM (here CFS) retrospective forecasts of gridded seasonal-mean precipitation.

- a **non-homogeneous hidden Markov model (NHMM)** is trained on the daily station observations, using the GCM's seasonal-mean precipitation as a predictor; the targeted seasonal statistic is then computed from a large ensemble of stochastic daily rainfall sequences generated by the NHMM.

- Both methods are shown to perform quite similarly under cross-validation. Although more complex, the NHMM is able to provide probabilistic information and the daily rainfall sequences required for water balance modeling directly.