statistical methods for tailoring seasonal climate forecasts

Andrew W. Robertson, IRI

tailored seasonal forecasts

- why do we make *probabilistic* forecasts? – to reduce our uncertainty about the (unknown) future state of the seasonal climate

- how do we make *tailored* forecasts? – by conditioning our variable of interest on existing clues (forecasts, large-scale climatic fields)

- *it is the dependency between observations of the variable of interest (x) and forecasts (y) that makes the forecasts useful – we need to maximize this dependency (forecast calibration, correction, downscaling) and quantify it (forecast verification)*

- regression analysis provides the mathematical tools

two basic ingredients for constructing tailored forecasts

1. a set of predictors – observations with predictive power, or GCM forecasts

2. a set of one of more predictands – the variable we want to predict

- first we train the model using forecasts made for past years – called “hindcasts” or “retrospective forecasts” – for which we have observational data

- second, we verify the model, also using historical data

predictors and predictands

- **the predictors:** generally a *gridded field* of a large-scale climate variable with large spatial dimension (100+ gridpoints) over a relatively small set of years, typically 20-30 years

- note that the spatial dimension of the predictor field generally exceeds the number of years available – this has important implications for the regression approach

- **the predictands:** generally one or more station locations

- NB: the regression approach requires that we have the predictors and predictands for the *same* set of years!

two methods in CPT for tailored seasonal forecasts

- *observational predictor design* – based on recent climate observations of the slow components of seasonal climate variations; i.e. typically upper-ocean heat content given by SST

- *GCM-MOS design* – statistical post-processing of general circulation model forecasts

- both rely on regression analysis

- in building the regression models and verifying skill of the forecast scheme, we have to account for the important fact that our historical observational records and GCM hindcasts are always *finite*

    - *the finite length of the records introduces sampling errors into our models and verification statistics – it is critically important to be aware of these!*

observational predictor design

- regression models (CCA, PCR) can be used to make purely statistical forecasts, using lagged observed predictors like SST

- rationale: SST is a slowly-evolving variable (reflecting upper-ocean heat content) that contains key memory for the evolution of seasonal climate anomalies

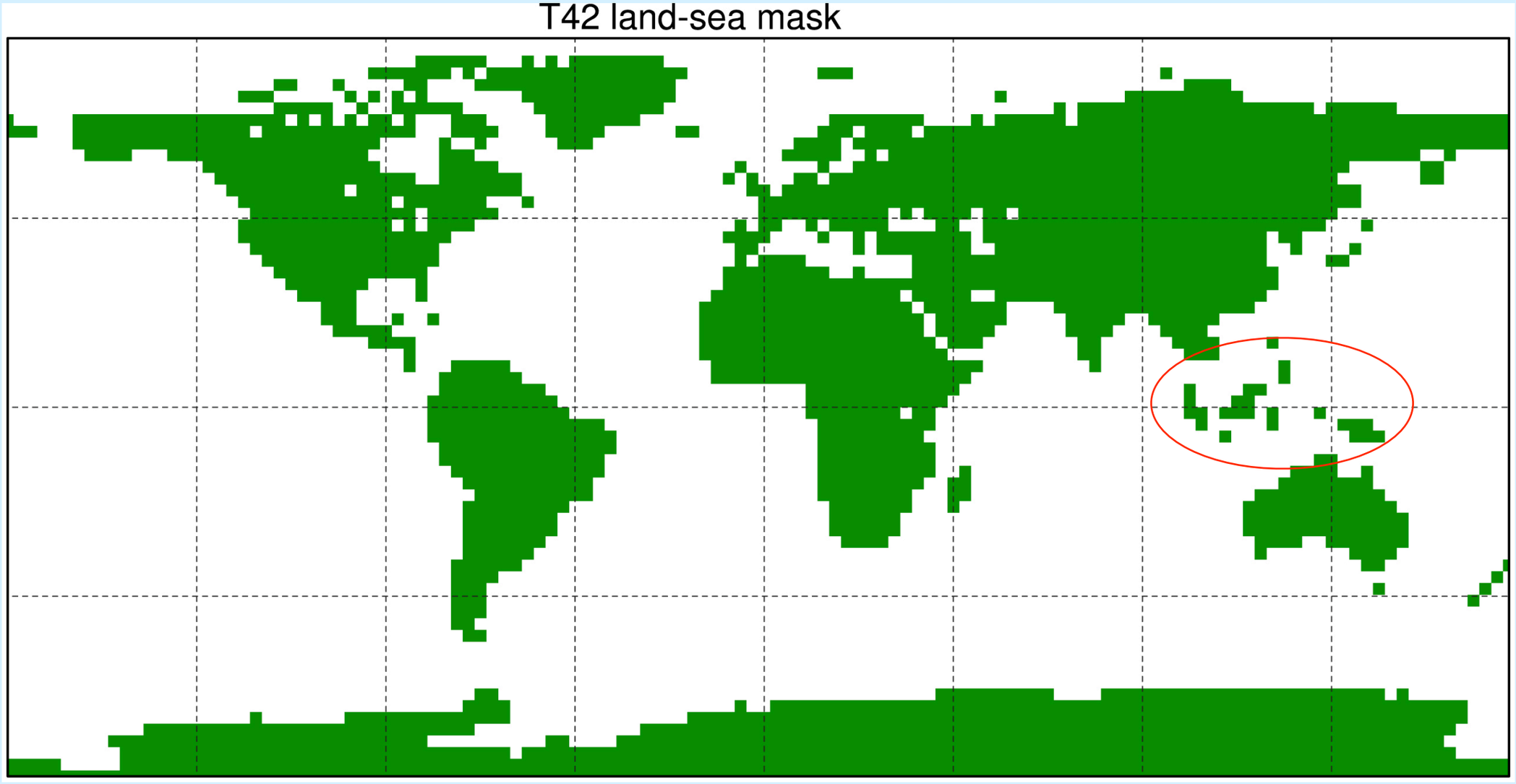
- can provide a valuable baseline for GCM predictions (often proves hard to beat!)

- we will illustrate this in our practical exercises

GCM-MOS design – statistical post-processing of general circulation model forecasts

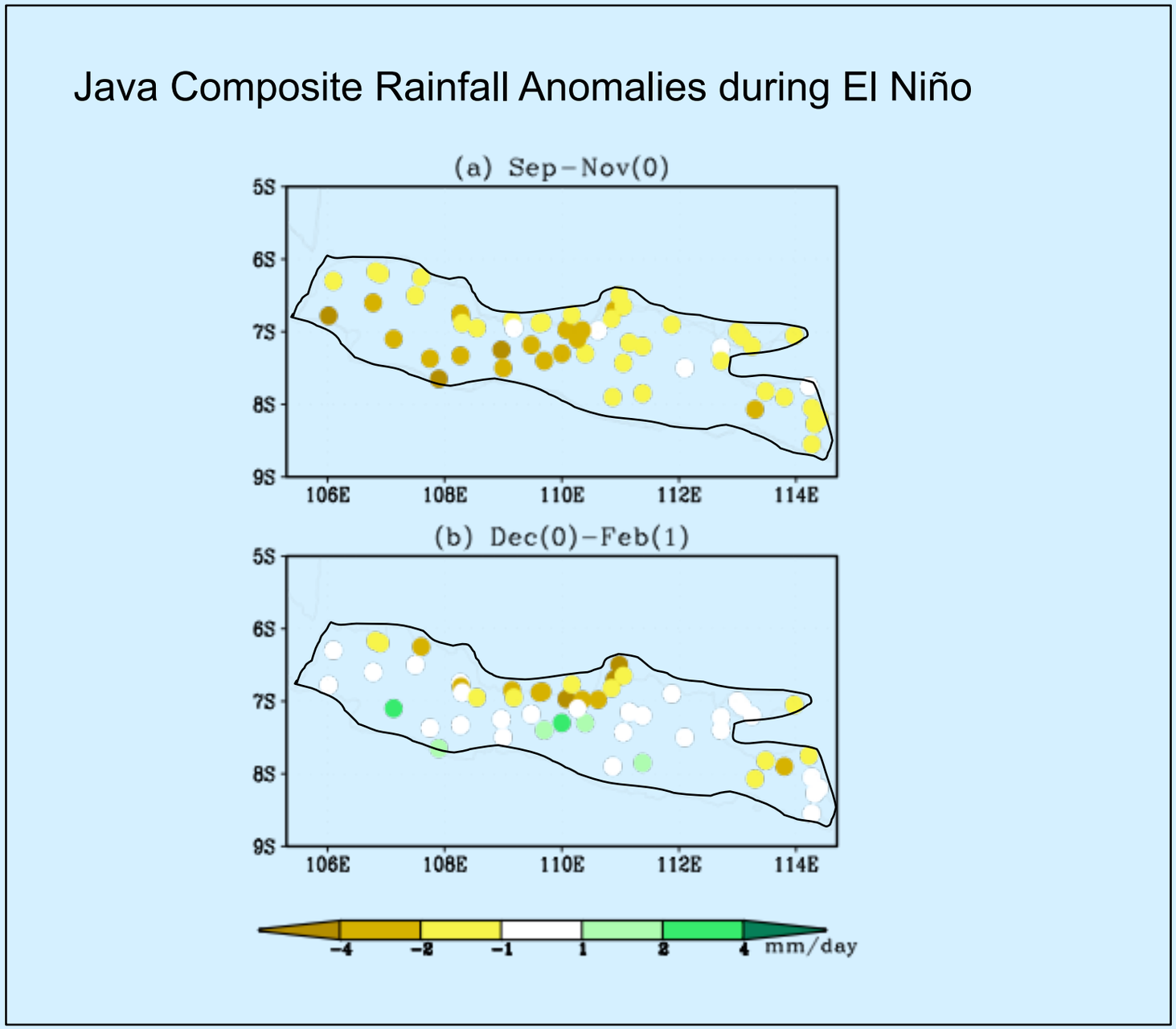
Motivation

- **GCM gridpoint forecasts are area-averages, and may not produce information at a scale that is useful for practical applications.**
- **A typical GCM grid represents about a 60,000 km^2 area**



Motivation

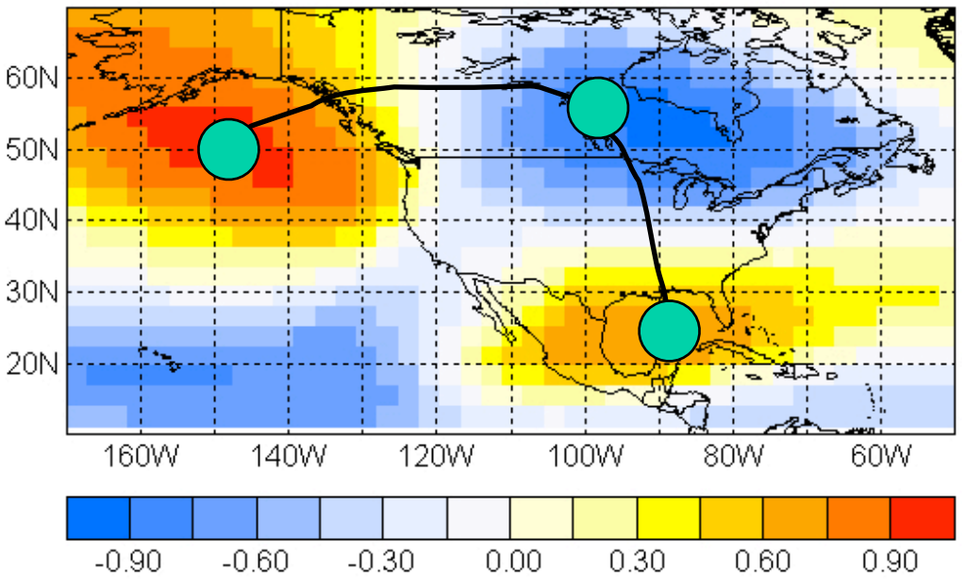
Climate can vary dramatically over short distances. Teleconnection patterns, like those associated with ENSO, can show similar fine scale structure.



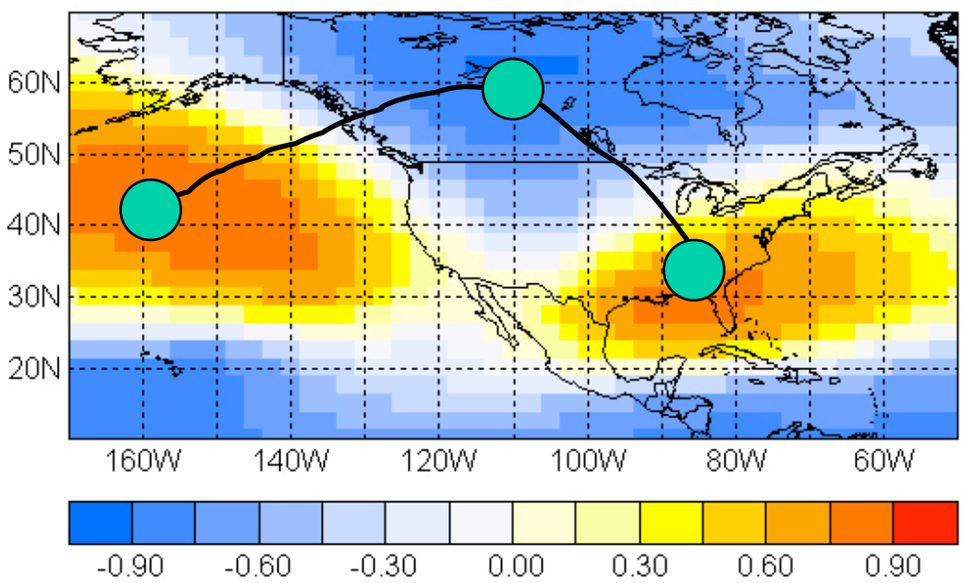
Motivation

Important climate features may be displaced in GCMs relative to observations: *Systematic spatial biases*

ECHAM 4.5 “PNA” Pattern

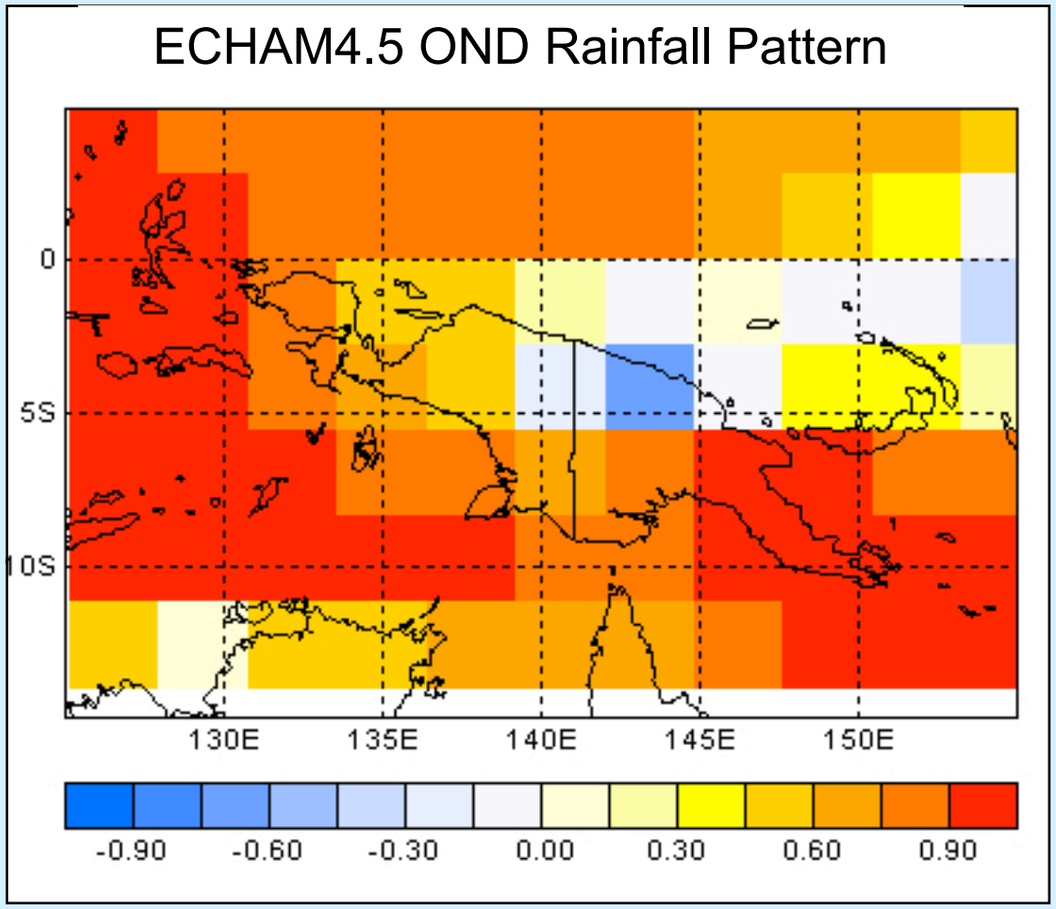


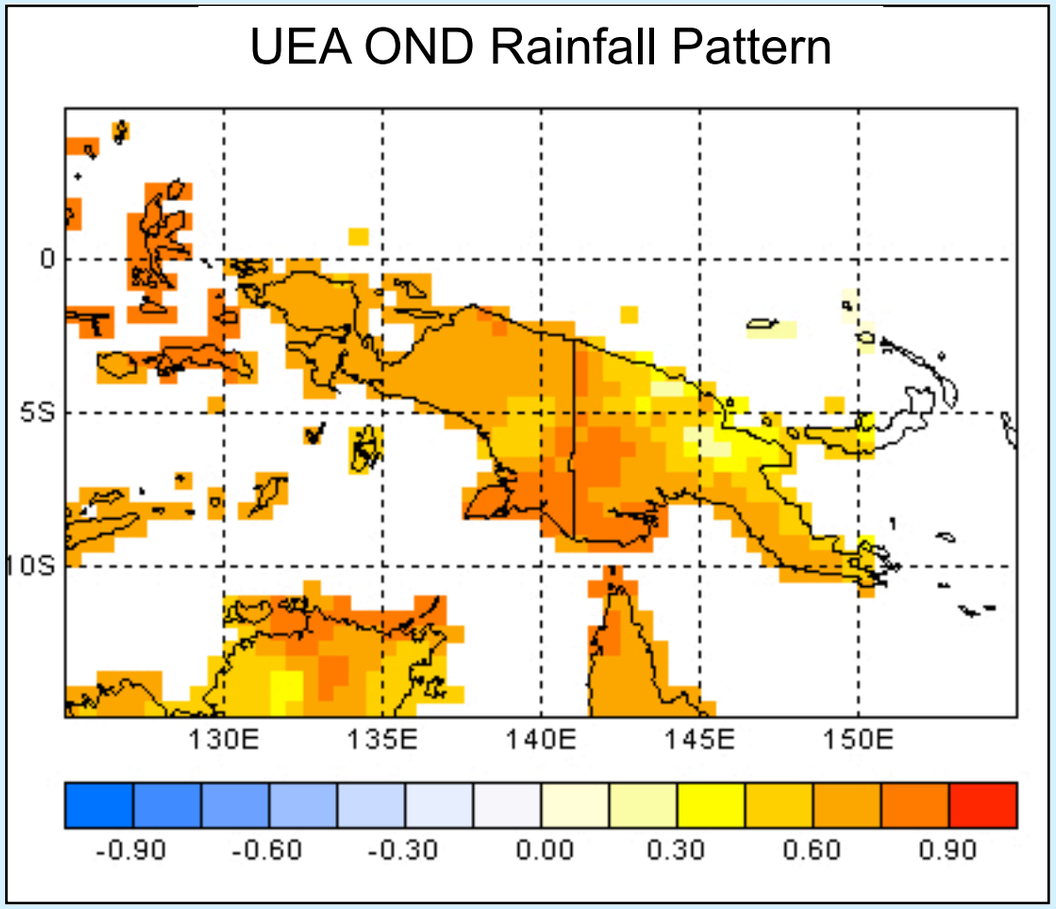
NCEP Reanalysis PNA



Motivation

Important climate features may be displaced in GCMs relative to observations: *Systematic spatial biases*

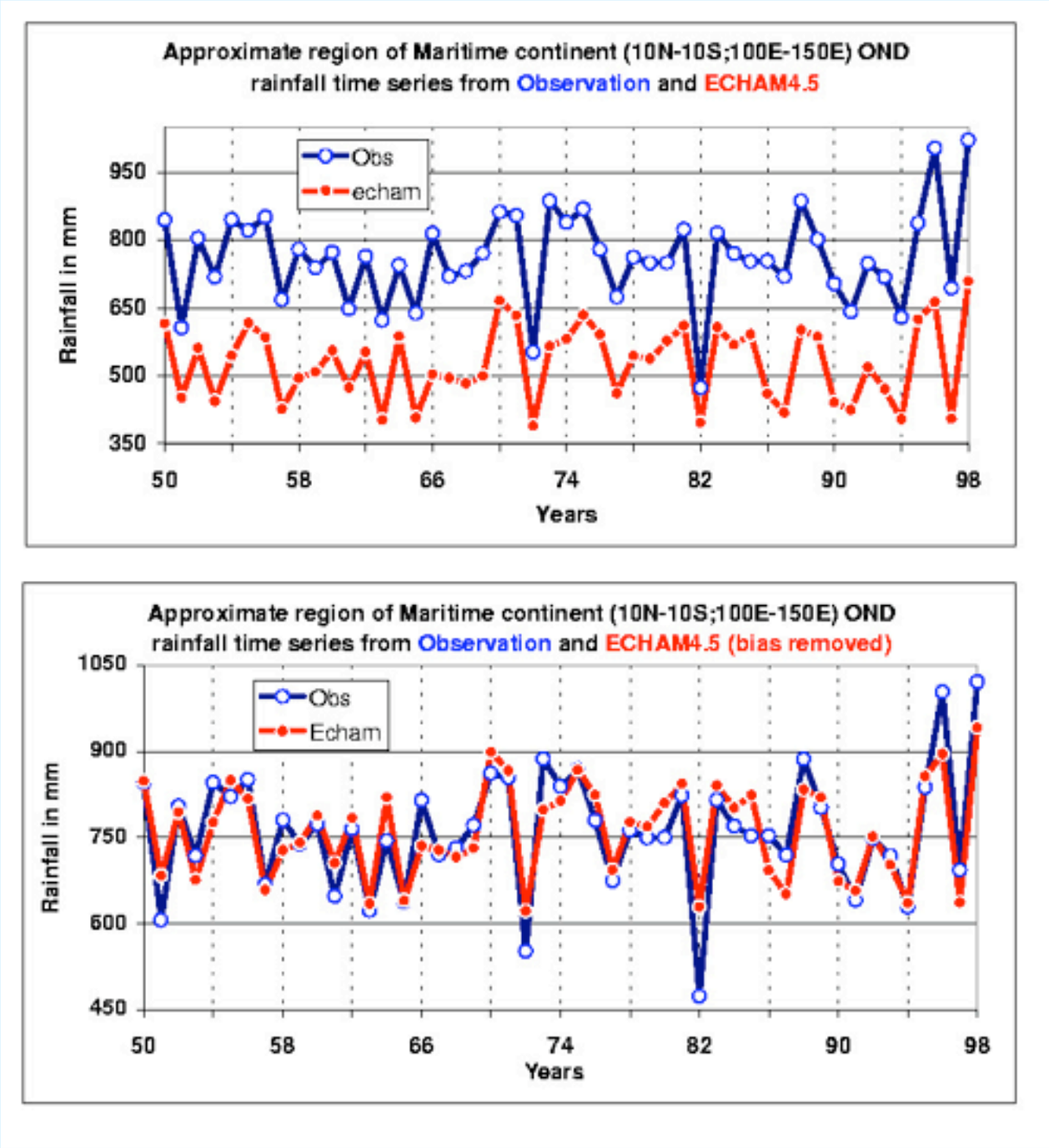




Motivation

GCM predictions may contain other systematic biases

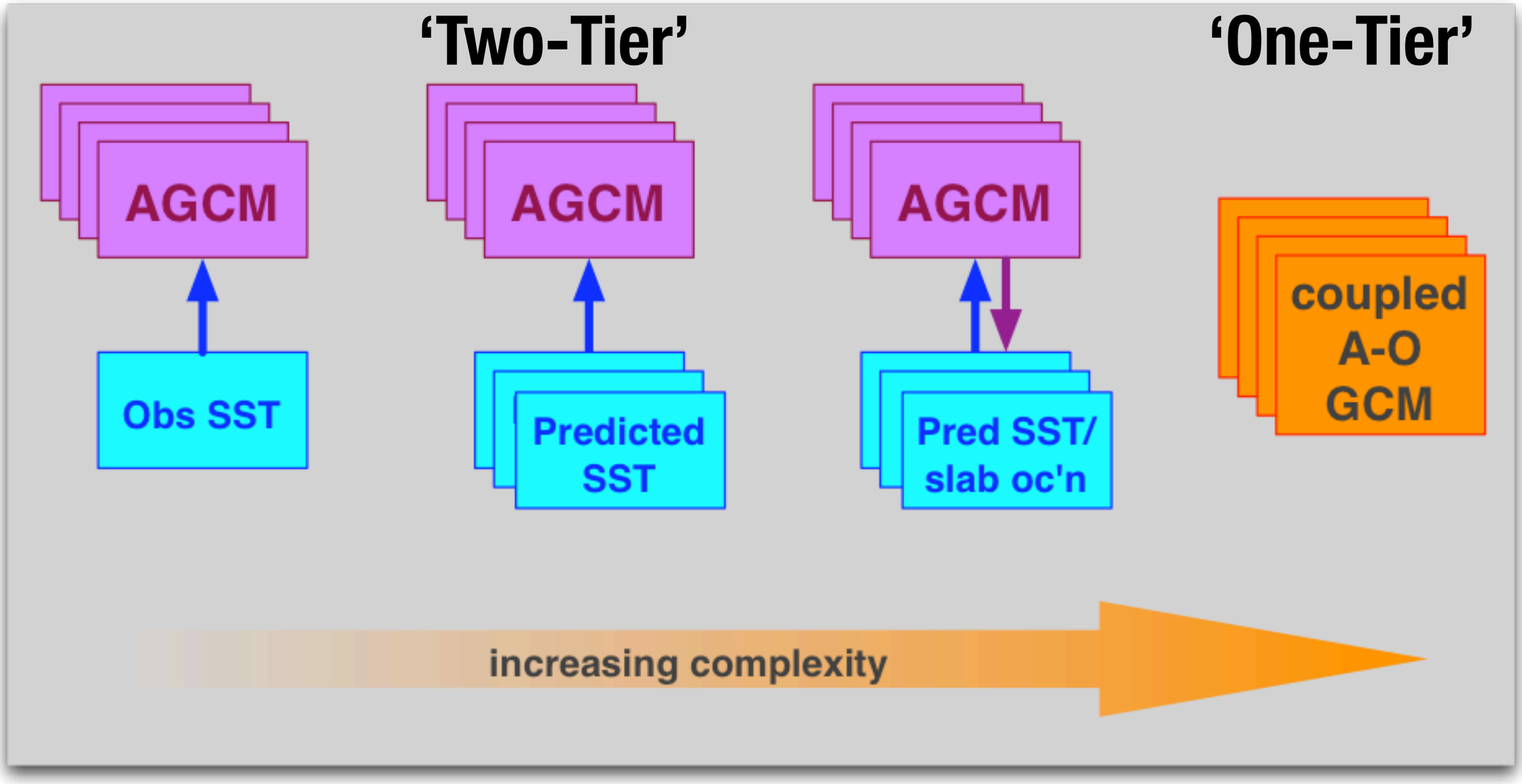
Systematic Mean Bias
(model too dry)



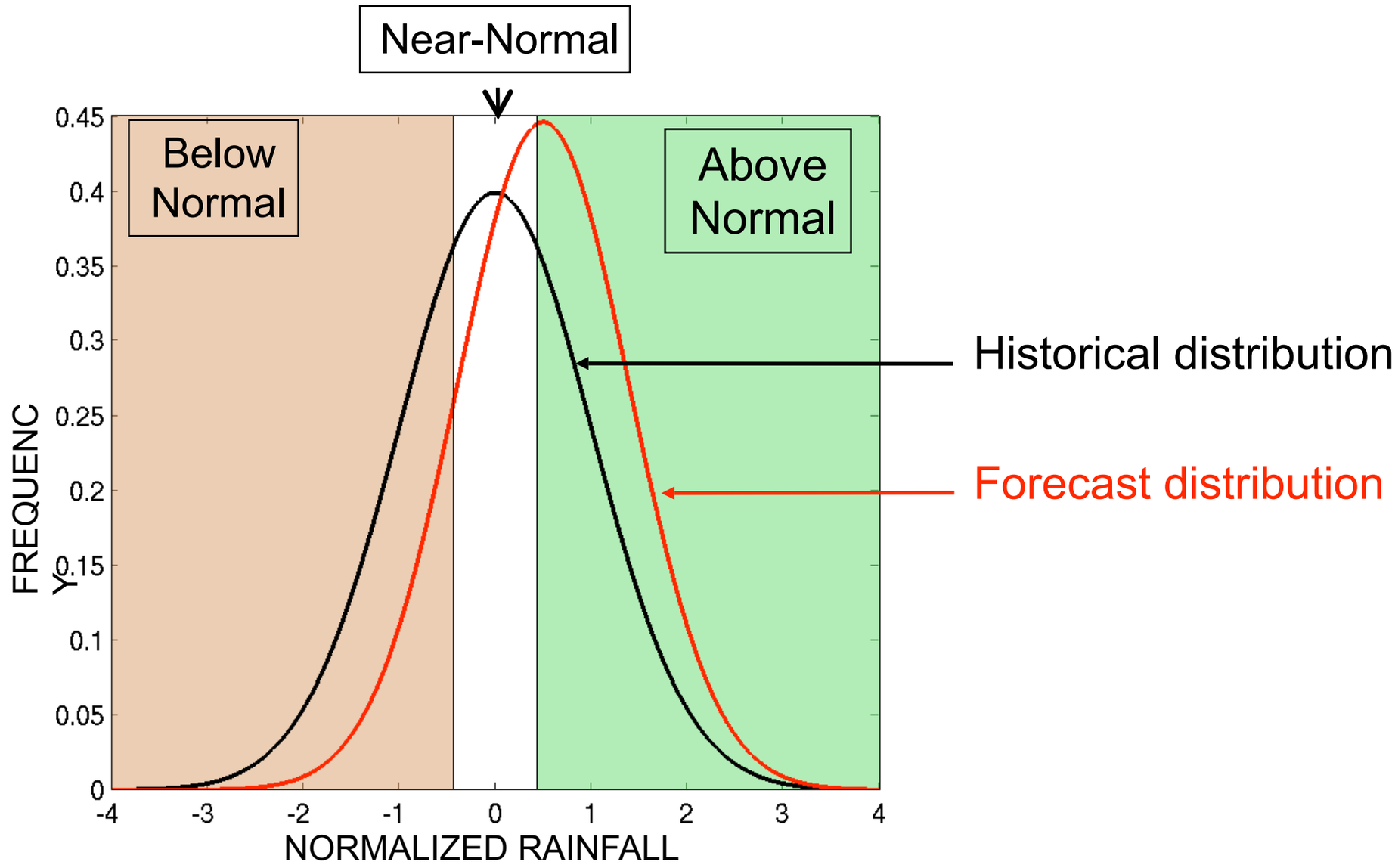
Mean Bias Correction

types of GCM forecast ...

GCMs for Seasonal Prediction



What probabilistic forecasts represent



Historically, the probabilities of above and below are 0.33. Shifting the mean by half a standard-deviation and reducing the variance by 20% changes the probability of below to 0.15 and of above to 0.53.

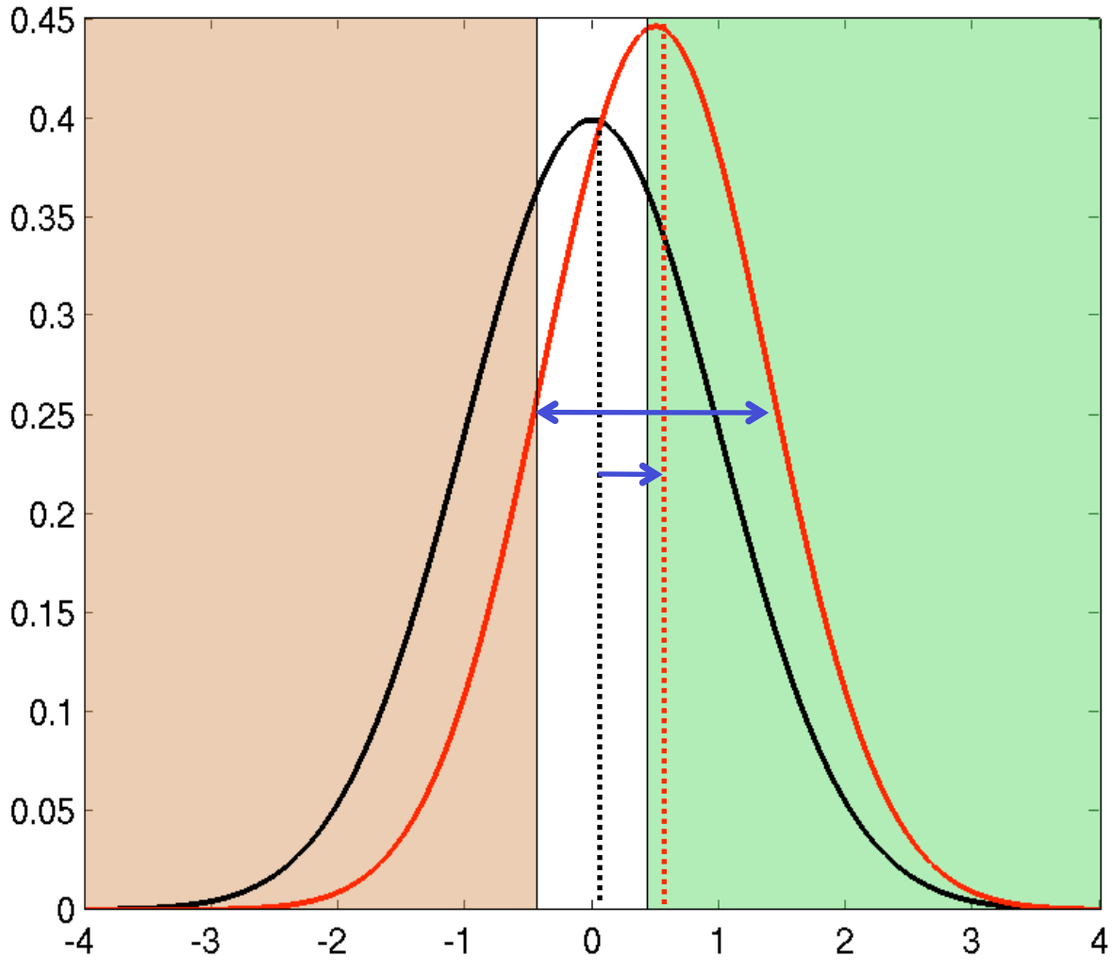
(Courtesy Mike Tippett)

L. Goddard

What probabilistic forecasts represent

..... Climatological Average

..... Forecast Mean



→ "SIGNAL

⟷ "NOISE

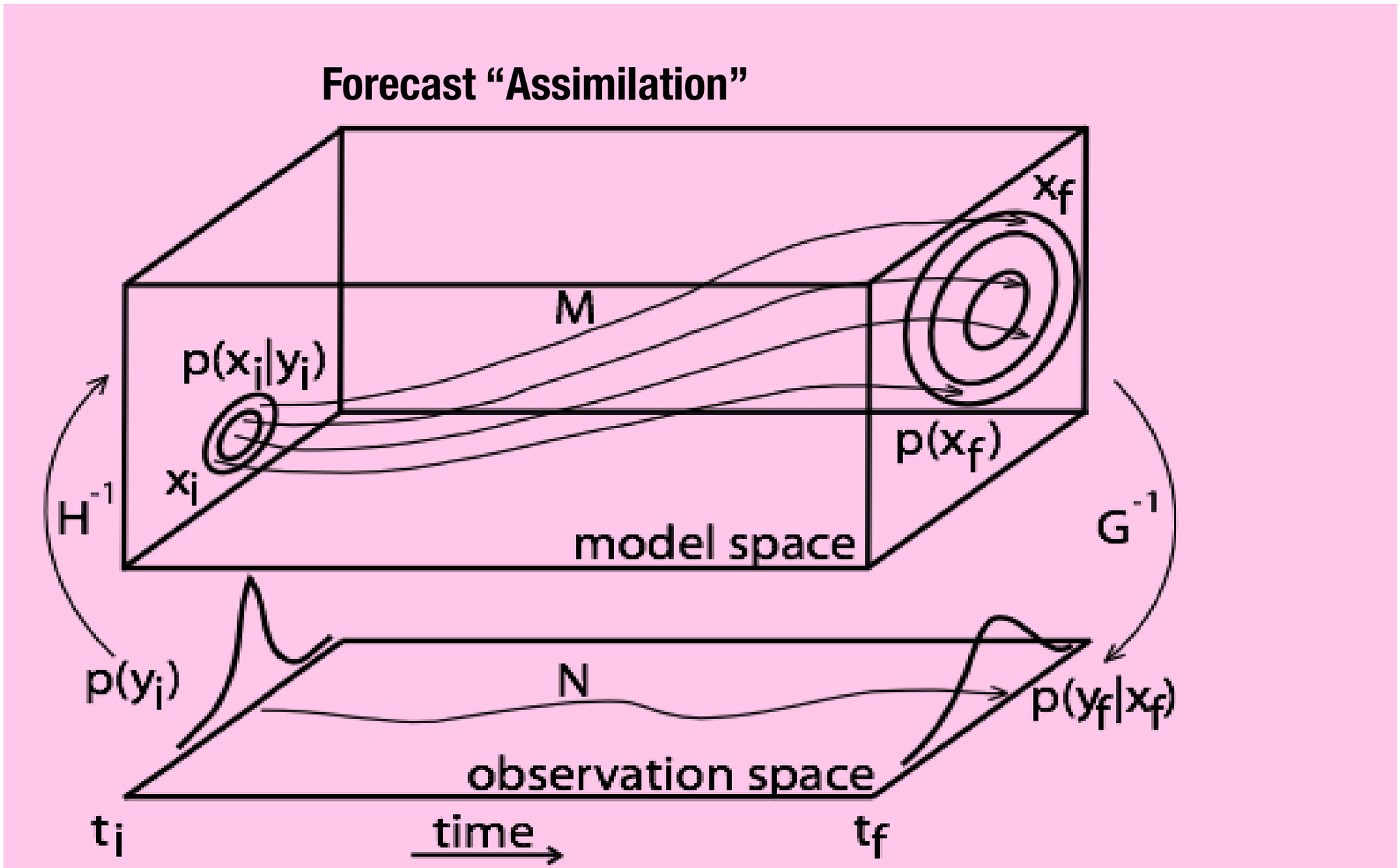
The **SIGNAL** represents the 'most likely' outcome, but quantifying the **NOISE** is an important part of the forecast. The **NOISE** represents internal atmospheric chaos, uncertainties in the boundary conditions, and random errors in the models.

L. Goddard

local data – the predictand

- meteorological station observations, e.g. rainfall
    - seasonal rainfall totals
    - seasonal counts of the number of rainy days
- could also be applications-relevant data such as streamflow records or satellite vegetation measures such as NDVI
- for regression analysis, long data records are essential in order to minimize sampling variable and to produce statistically significant results!

GCM predictions and regression models



Stephenson et al (2005)

recap so far ...

- we have a **predictor field of large spatial dimension** (from large-scale climate observations or GCM forecasts)

- we would like to make tailored forecasts for **one or more predictands**

- how do we use regression techniques to do this intelligently?

varieties of linear regression

- simple regression: a **single** predictor and a **single** predictand:
  $y = ax + b$

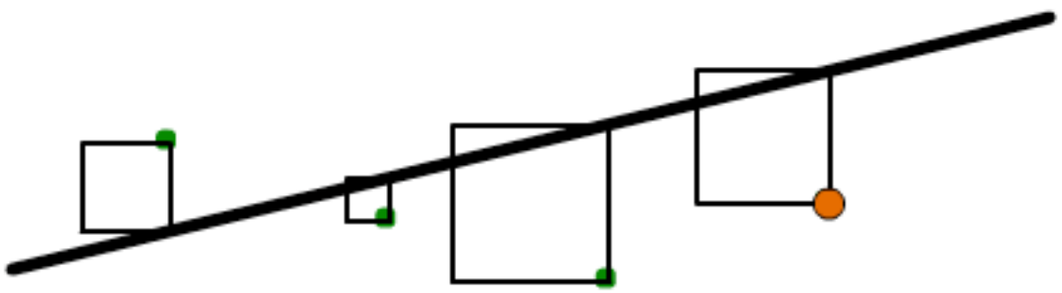
- multiple regression: **two or more** predictors, and a **single** predictand
  $y = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n$ (case of n predictors)
  --The PCR tool in CPT is in this category of regression

- multivariate (pattern) regression: **two or more** predictors, **two or more** predictands
  y = Ax (matrix **A**)

  -- The CCA tool in CPT is in this category of regression

linear regression - the method of least squares

Visual Least Squares



Residual SS: 8217

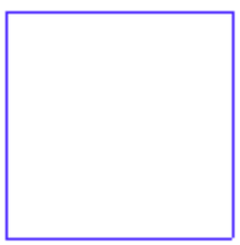
Total SS: 1537

R-square: --%

Residual SS: 1076

Total SS: 1537

R-square: 29.9%



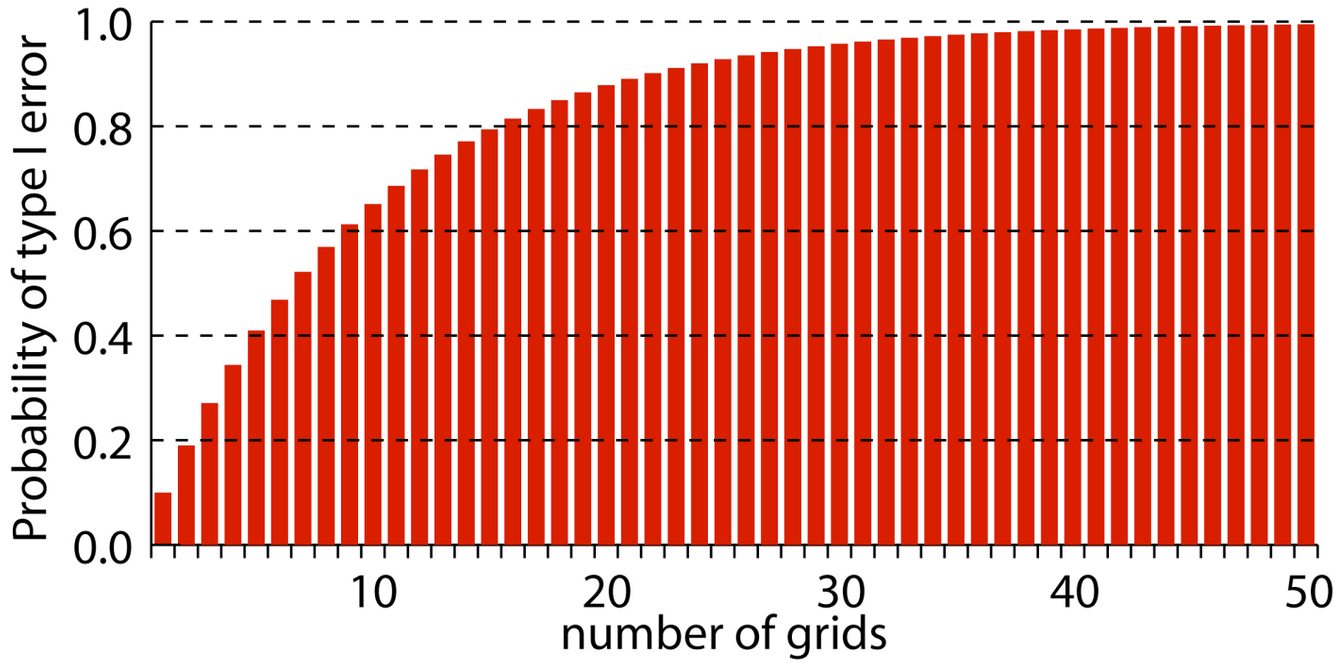
Fit & Lock

Unlock

Visual Least Squares. Press the keyboard arrow keys and buttons to see the squared deviations fluctuate. Fit & Lock finds and recomputes the least squares line as the orange point is moved. The residual sum of squared errors is shown in blue, turning green when minimized. Daniel G. Goldstein & Stephen M. Stigler.

Why not multiple Regression?

Multiplicity - Too many grids from which to choose.



Multicolinearity - Grids are strongly correlated.

$$\text{Nino3.4}_{\text{Mar}} = \beta_0 + 0.761 \times \text{Nino3.4}_{\text{Feb}} + \varepsilon$$

$$\text{Nino3.4}_{\text{Mar}} = \beta_0 + 0.628 \times \text{Nino3.4}_{\text{Jan}} + \varepsilon$$

$$\text{Nino3.4}_{\text{Mar}} = \beta_0 + 1.216 \times \text{Nino3.4}_{\text{Feb}} - 0.395 \times \text{Nino3.4}_{\text{Jan}} + \varepsilon$$

principal components regression

- multiple regression: two or more predictors, and a single predictand
  $y = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n$

- principal components regression (PCR) uses principal components (PC) analysis to obtain a *small* set of *uncorrelated* predictors

- the leading PCs account for the largest fraction of the temporal variance in a field of predictors; they *compress* the field optimally in terms of variance

  - greatly reduces the risk of *multiplicity*

- the PCs are uncorrelated in time

  - eliminates the problem of *multiple co-linearity* in multiple regression

canonical correlation analysis (CCA)

- to predict the observed precip anomaly field **y** from a predictor anomaly field **x**, we assume the linear relationship $\mathbf{y} = \mathbf{A}\mathbf{x}$

- we minimize the regression error $\langle(\mathbf{y} - \mathbf{A}\mathbf{x})^T(\mathbf{y} - \mathbf{A}\mathbf{x})\rangle$ by choosing $\mathbf{A} = (\mathbf{y}\mathbf{x}^T)(\mathbf{x}\mathbf{x}^T)^{-1}$

- there is insufficient data to determine the regression matrix **A** when the length of the historical record is smaller than the spatial dimension of **x** and **y**

- some regularization method is necessary to invert the (singular) matrix $\langle\mathbf{x}\mathbf{x}^T\rangle$ and reduce the effect of sampling

- a simple regularization method is to expand the observed and predictor anomaly fields in truncated Empirical Orthogonal Function (EOF) series using Principal Component Analysis (PCA).

- then, using singular value decomposition, we can factor the regression matrix as $\mathbf{A} = \mathbf{C_y M C_x^{-1}}$

- $\mathbf{C_y}$ and $\mathbf{C_x}$ are linear combinations of observation and predictor EOFs with maximum correlation, known as the observation and predictor homogeneous covariance maps (diagonal matrix **M** contains the canonical correlations)

- Therefore, the determination of the regression matrix **y=Ax** requires specifying the number of observation and predictor EOFs used, together with the number of CCA modes (non-zero elements of **M**)

- This is the reduction of the spatial dimension of the problem that makes it tractable!

physical interpretation

- the PCR and CCA are methods for building multivariate regression models using singular value decomposition to reduce the spatial dimension of the problem

- but should not be used as black boxes!

- visual inspection of the EOF and CCA modes is essential for building a physically robust model, and for choosing the appropriate number of components to retain

- simple correlation maps with a station-index can be a very helpful aid (e.g. using the on-line tool http://www.cdc.noaa.gov/correlation)

geographical choice of predictor field

- the longitude/latitude limits of the predictor field should be large enough to contain the main predictive features of the field from physical considerations

- for SST (observational predictor design) this might include the tropical Pacific and Indian Ocean

- for GCM-MOS design this might be a box 20-30 degrees across, centered over the region of interest

candidate predictors

1. contemporaneous SST
   – is local rainfall related to slowly-varying, potentially predictable components of the climate system?

2. lagged SST: simplest forecast!

3. GCM simulations
   – GCM variables must be related to local variables -- test with AGCM+obs SSTs

4. GCM predictions
   – GCM variables must contain predictability -- test with retrospective forecasts

recap: experimental design

- target seasons: Oct-Dec and Feb-Apr

- choice of predictors
  1. contemporaneous SST (e.g. NOAA_ERSST_1950-2006OND_CPT.txt [60S-60N, 0-360E])
  2. SST from previous month (e.g. NOAA_lead1_ERSST_1950-2006OND_CPT.txt i.e. **Sept SST**, [60S-60N, 0-360E])
  3. GCM forecast from previous month (e.g. E4CA_PPT_SEAsia_L2-4OND_CPT.txt **made on Sept 1 using Aug SST** [20S-20N, 90E-180E] ensemble mean)

- choice of predictands
  1. seasonal rainfall total
  2. number of dry days

recap: two varieties of linear regression in CPT

PCR: multiple regression: **two or more** predictors, and a **single** predictand

$y = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n$

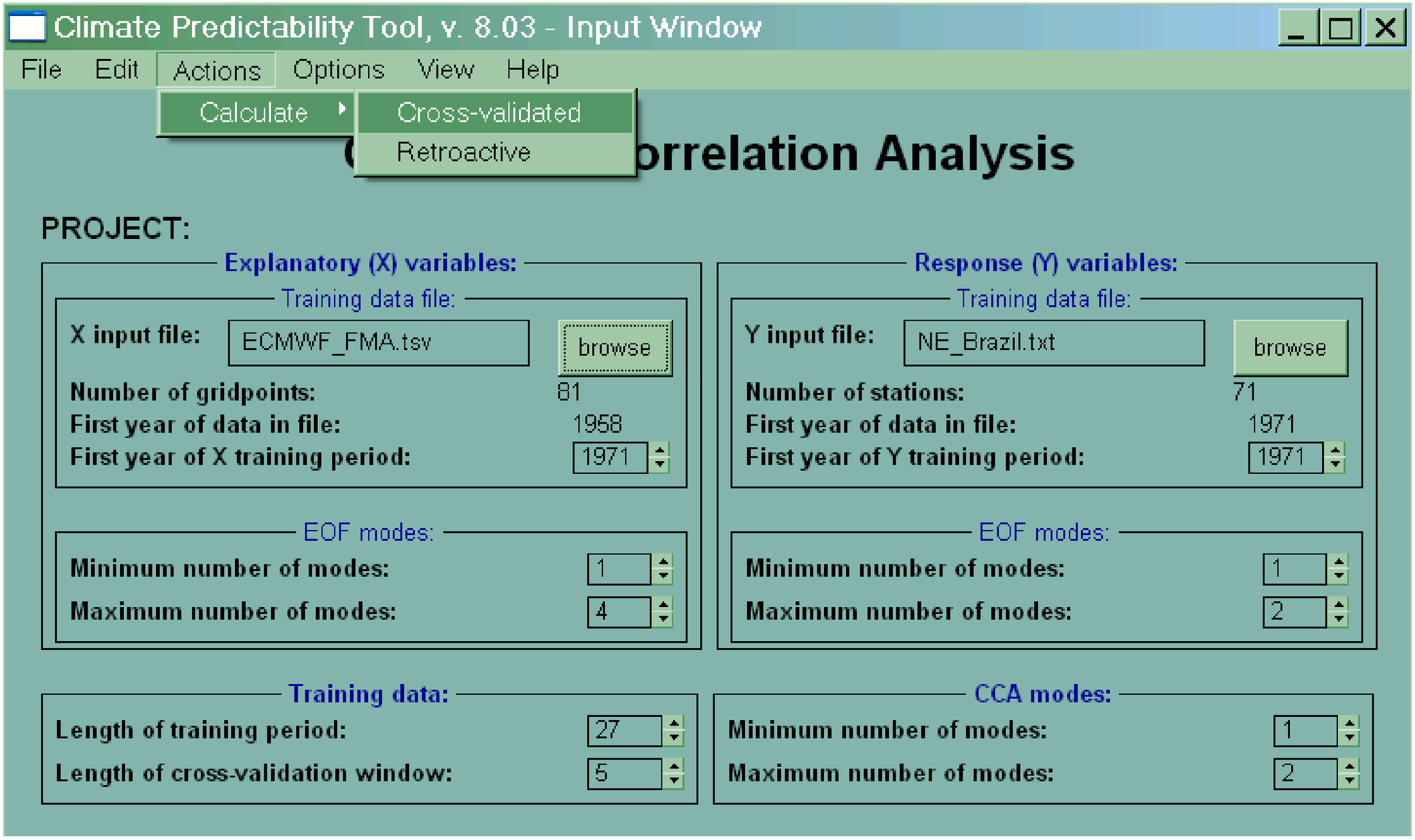
- problems of multiplicity and multicolinearity
- reduce large number of x's to just a few EOF time series

CCA: multivariate (pattern) regression: **two or more** predictors, **two or more** predictands **y = Ax** (matrix **A**)

- intractable for large spatial fields
- reduce large number of x's **and** y's to just a few EOF time series of each
- matrix **A** is thus reduced to the set of *patterns* of **x** and **y** that are best correlated in time

- in both cases the numbers of EOFs to include is important (neither "knows" about the variance explained)

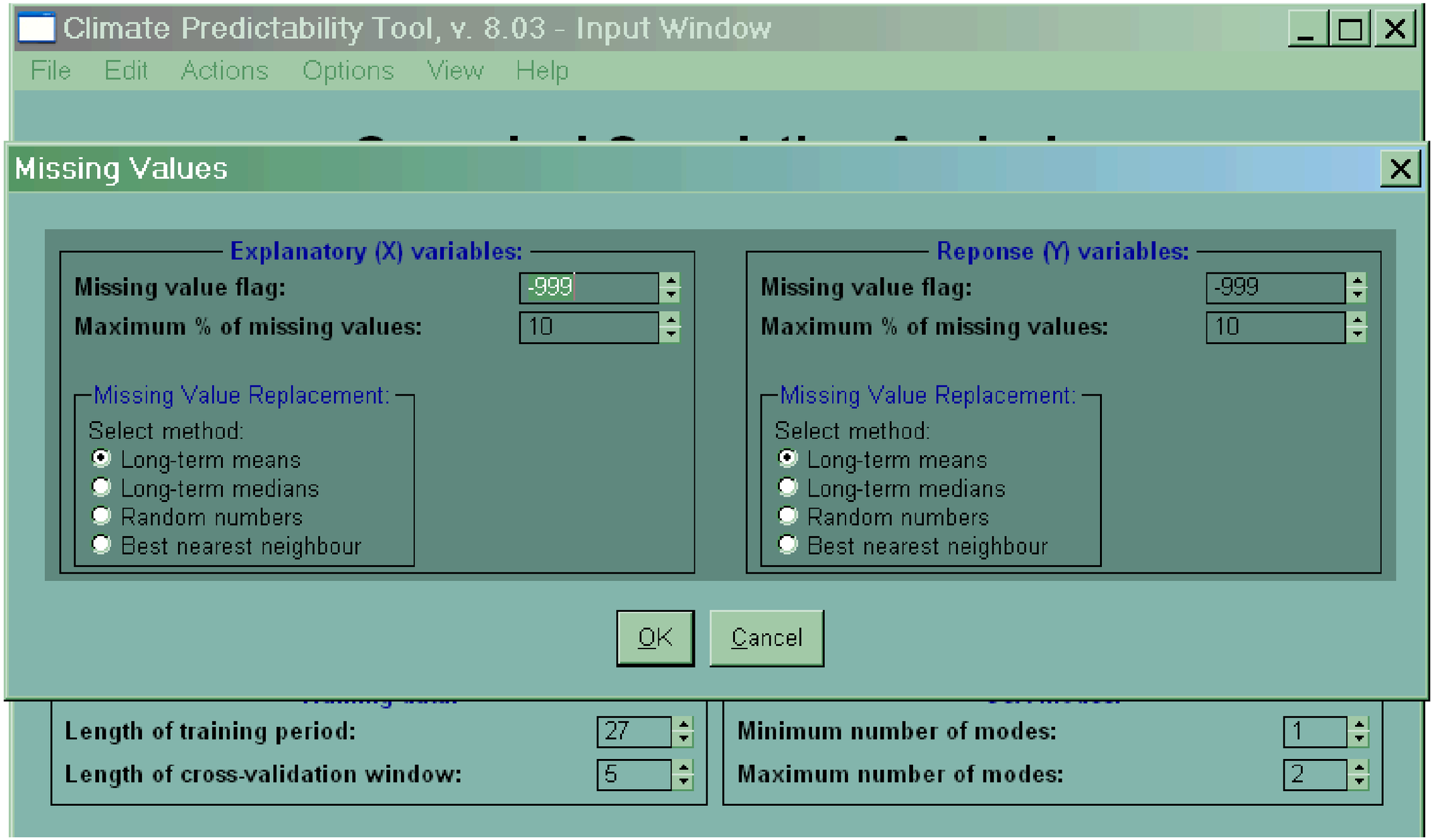
RUNNING CPT



Then you can run the analysis:

Actions => Calculate => Cross-validated

MISSING VALUES



Next to the Missing value flag box, you need to specify the number in your dataset that represents a missing value.

You can choose the maximum % of missing values. If a station has more than that percentage of missing values, CPT will not use that station in its model.

You can also choose which method you want CPT to use to replace the values.

DATA ANALYSIS

Climate Predictability Tool, v. 8.03 - Results Window

File Tools Customise Help

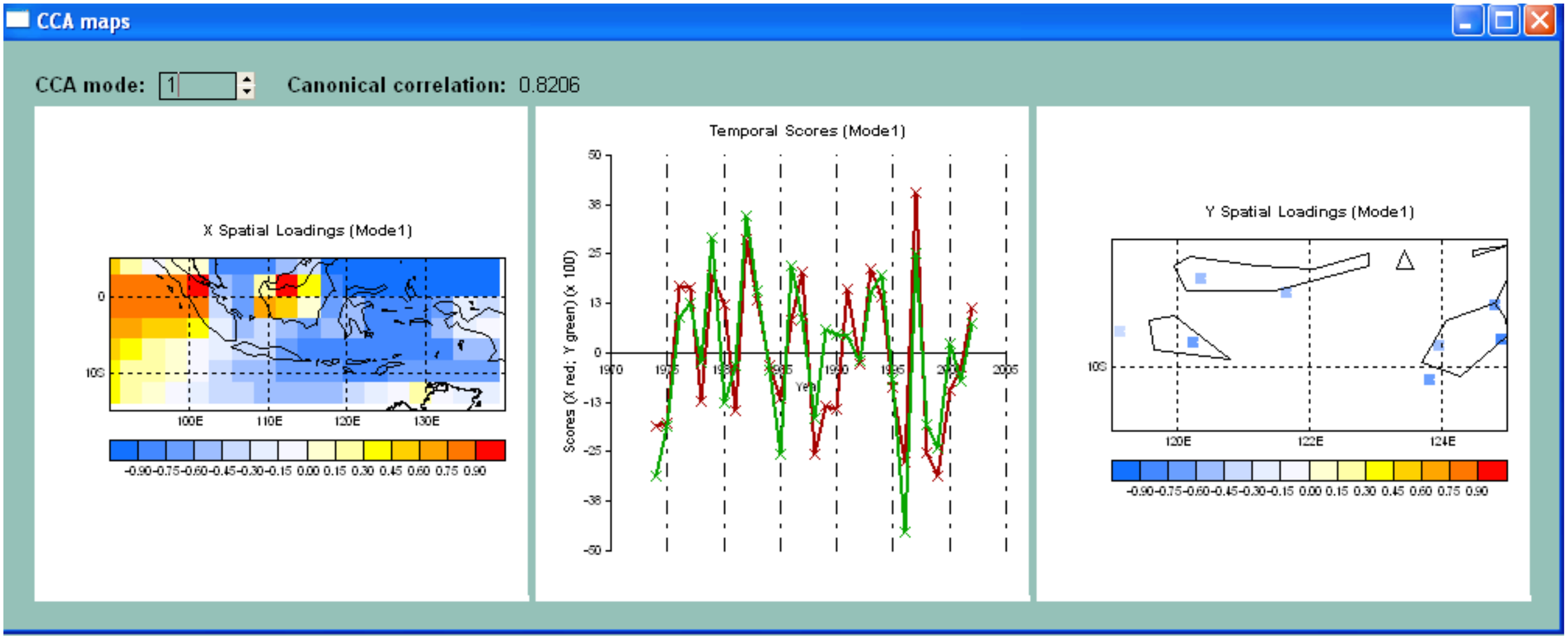
Progress: 100%

Actions:

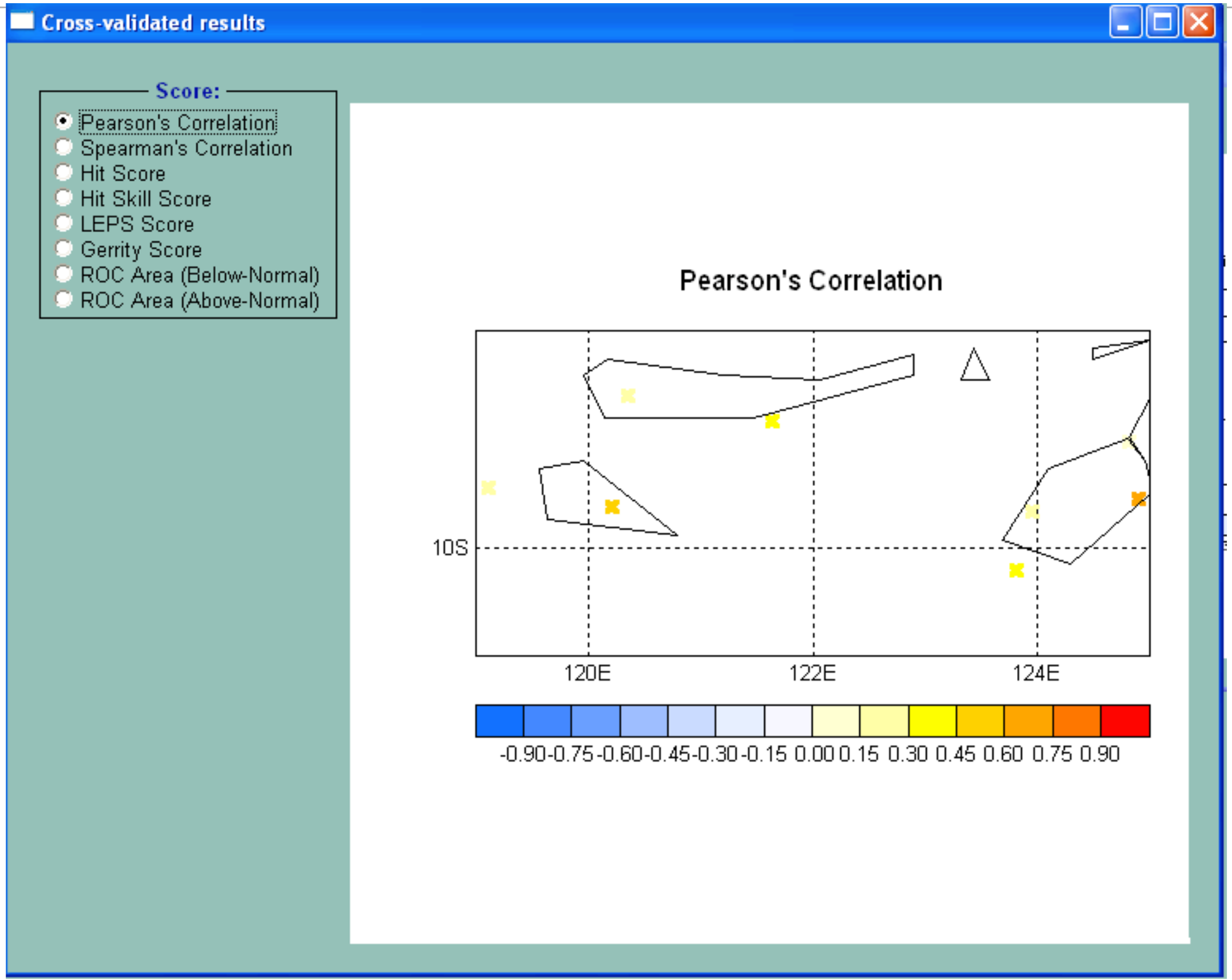
```
Reading C:\Documents and Settings\simon\Application Data\CPT\Data\ECMWF_FMA.tsv ...
Checking for missing values ...
Reading C:\Documents and Settings\simon\My Documents\Slides\Masters\DATA\NE_Brazil.txt ...
Checking for missing values ...
Data read successfully

Calculating climatologies and thresholds ...
Optimizing cross-validated performance ...
Training period: 1971 to 1997

              CURRENT                          OPTIMUM

     Number of Modes    Goodness      Number of Modes    Goodness
        X    Y   CCA       Index         X    Y   CCA       Index

        1    1     1       0.517         1    1     1       0.517
        1    2     1       0.515         1    1     1       0.517
        2    1     1       0.509         1    1     1       0.517
        2    2     1       0.500         1    1     1       0.517
        2    2     2       0.507         1    1     1       0.517
        3    1     1       0.497         1    1     1       0.517
        3    2     1       0.473         1    1     1       0.517
        3    2     2       0.488         1    1     1       0.517
        4    1     1       0.493         1    1     1       0.517
        4    2     1       0.463         1    1     1       0.517
        4    2     2       0.486         1    1     1       0.517
Constructing model using full training period (1971 to 1997) ...
Identifying categories ...
Done!
```

CPT begins the specified analysis in a new “Results Window”. Here you can see the steps of the analysis and of the optimization procedure.

CCA analysis of NTT station data (Oct–Dec) and GCM hindcasts of precip from Sept 1



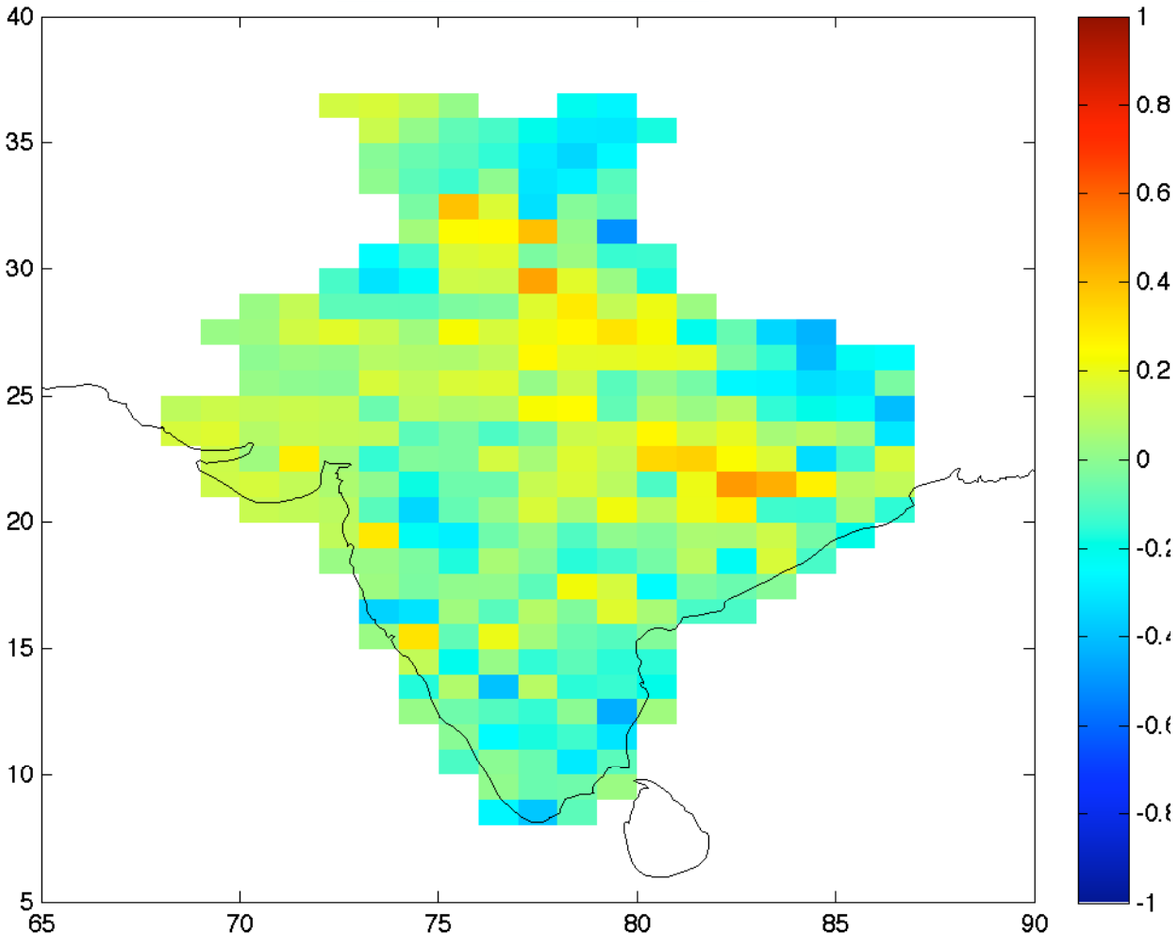
NTT OND hindcast skill (from Sept 1)



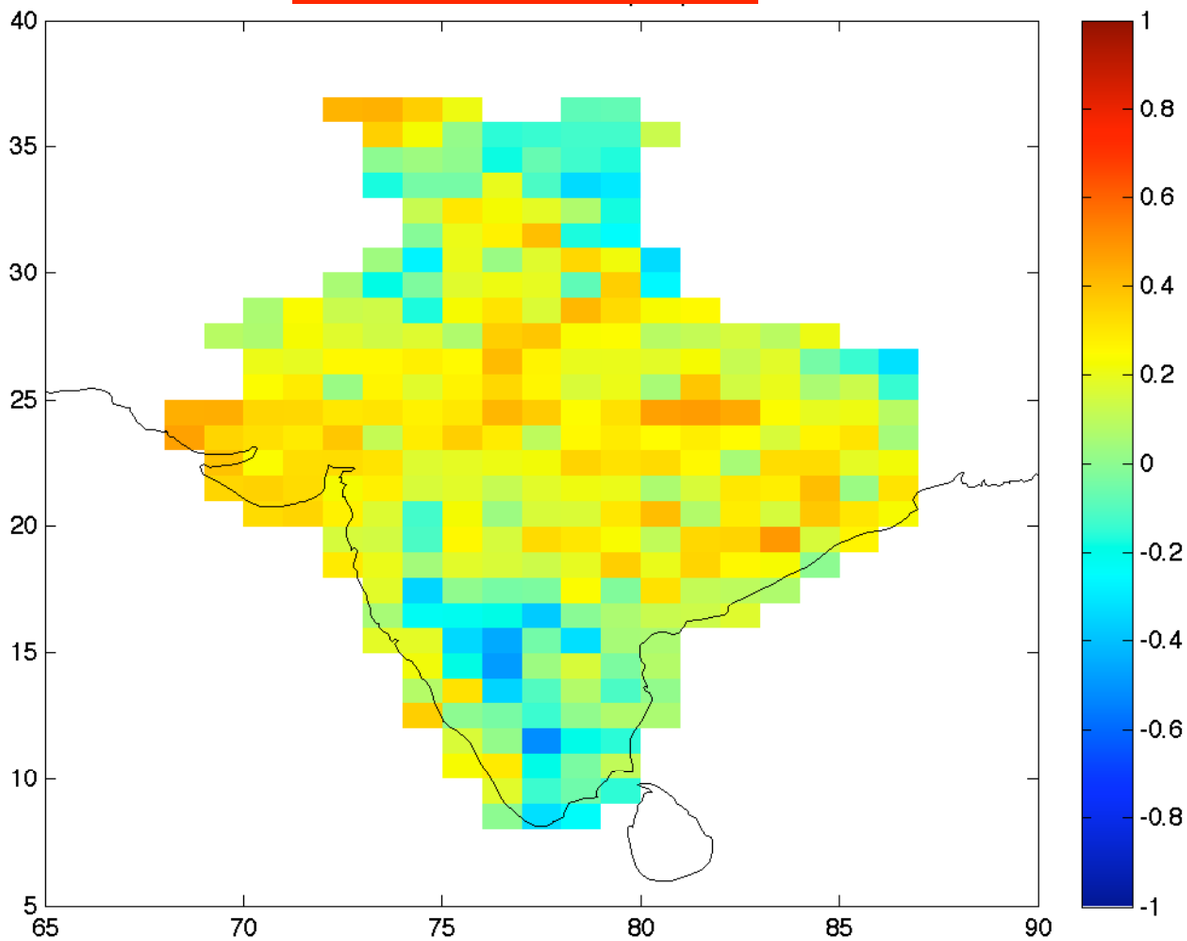
Jun–Sep anomaly correlation skill:

CCA[ECHAM4 precip (60E-90E, 5N-30N) , IMD]

seasonal total



rainfall frequency



model validation/forecast verification

- our goal is to build regression models that not only fit the training timeseries well, but which will also work well for real forecasts

- verification requires testing the model on *independent* historical data

- overfitting (“artificial skill” or “skill inflation”) can result if the regression model is fit to noise

- CPT helps avoid overfitting by (a) using EOF reduction of predictor and predictand fields, and (b) choosing the EOF truncations using rigorous cross-validation and retroactive forecasts

- CPT provides a set of performance measures, both continuous and categorical – following WMO Standard Verification System for Long Range Forecasting (SVS for LRF) (http://www.wmo.ch/web/www/DPS/SVS_for_LRF.html)

Leave-one-out cross-validation

Leave-one-out cross-validation

1951	Predict 1951	Training period							
1952	Training period	Predict 1952	Training Period						
1953	Training period		Predict 1953	Training period					
1954	Training period			Predict 1954	Training Period				
1955	Training period				Predict 1955	Training period			

… then correlate 1951–2000.

Leave-*k*-out cross-validation

Here, k = 5

1951	Predict 1951	Omit 1952	Omit 1953	Training period			
1952	Omit 1951	Predict 1952	Omit 1953	Omit 1954	Training period		
1953	Omit 1951	Omit 1952	Predict 1953	Omit 1954	Omit 1955	Training period	
1954	Training period	Omit 1952	Omit 1953	Predict 1954	Omit 1955	Omit 1956	Training period
1955	Training period		Omit 1953	Omit 1954	Predict 1955	Omit 1956	Omit 1957

Ensure that cross-validation window length is at least twice the decorrelation time

Retroactive forecasting

Another option available in CPT...

1981	Training period (1951-1980)	Predict 1981	Omit 1982+			
1982	Training period (1951-1981)		Predict 1982	Omit 1983+		
1983	Training period (1951-1982)			Predict 1983	Omit 1984+	
1984	Training period (1951-1983)				Predict 1984	Omit 1985+
1985	Training period (1951-1984)					Predict 1985

model validation

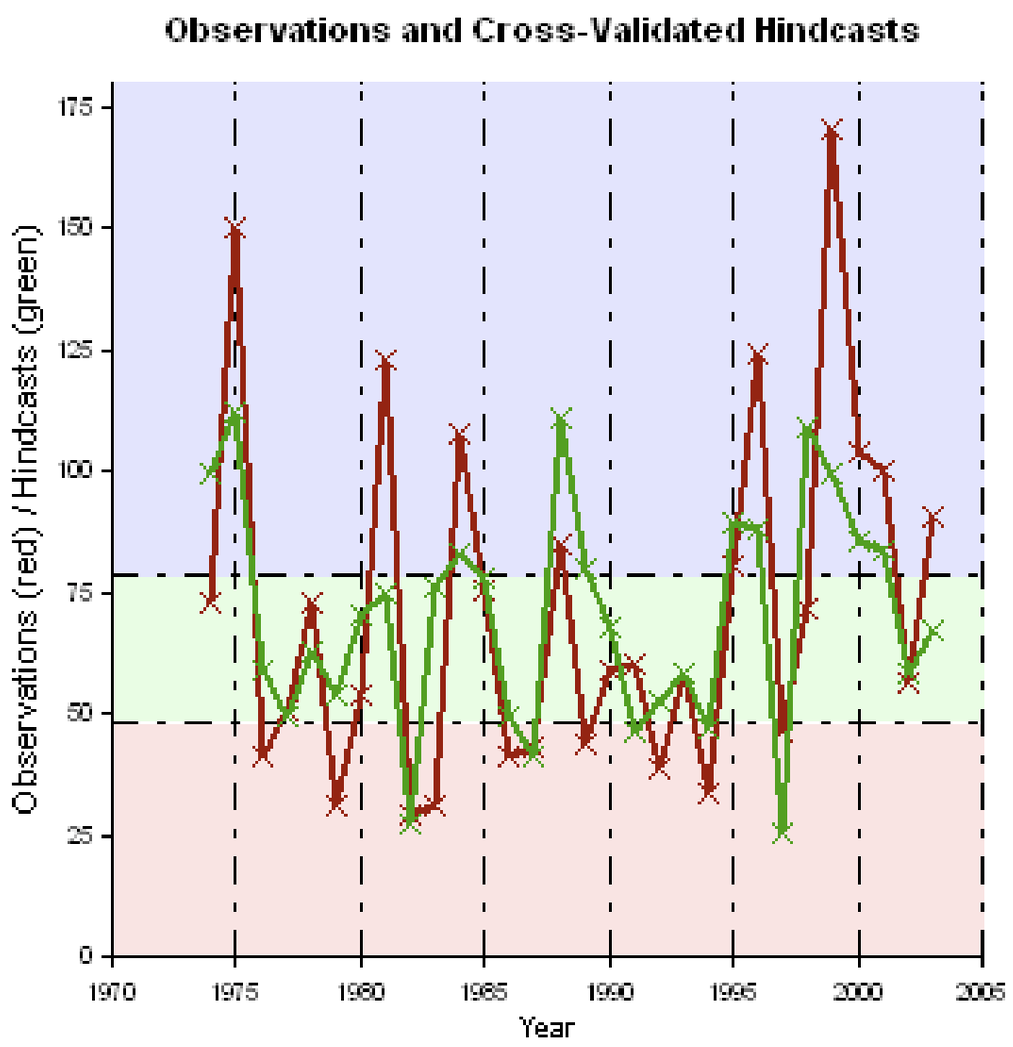
Cross-validated scores

Station 2

WAING 9.62S, 120.22E

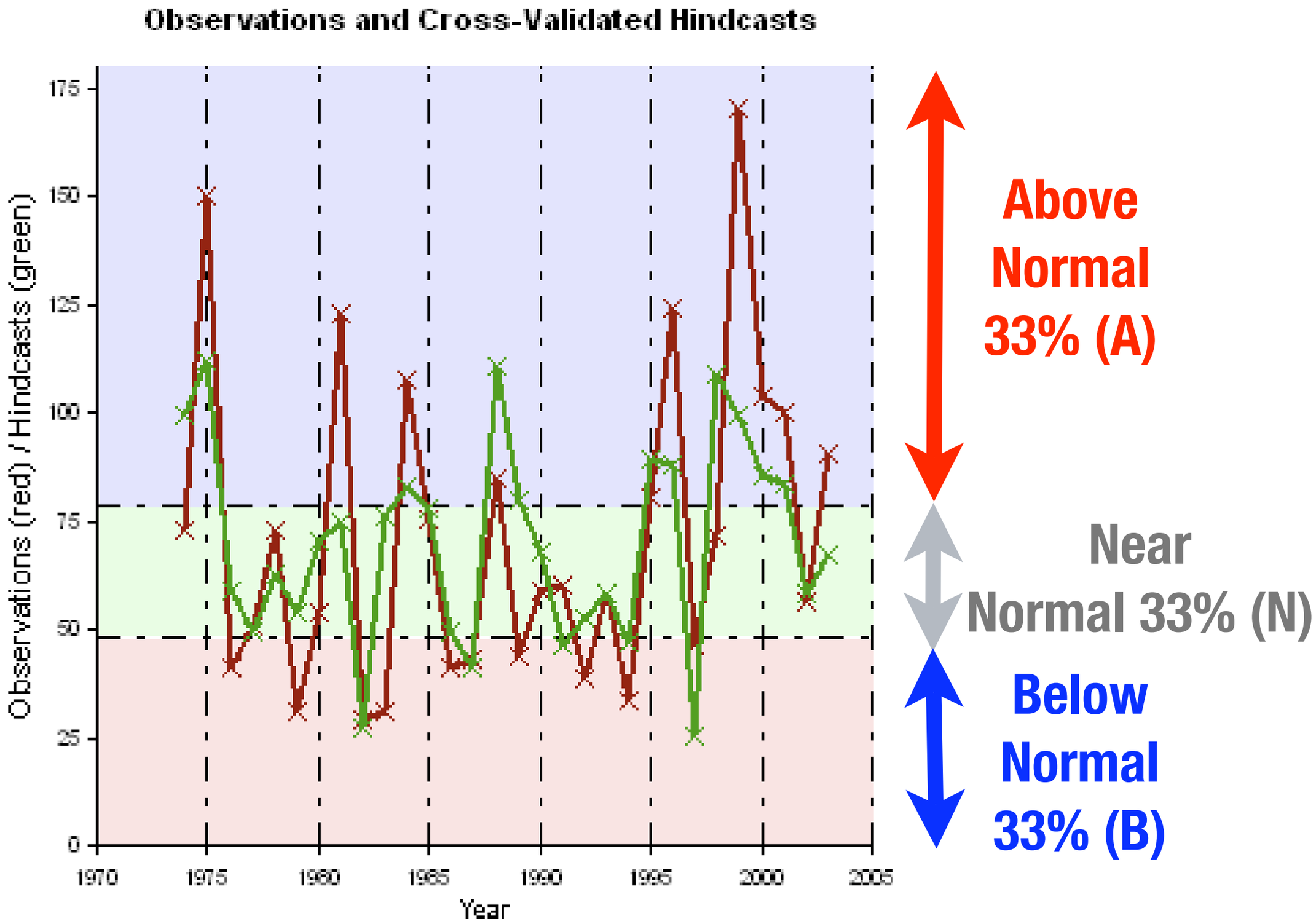
Continuous measures:

Pearson's correlation	0.6840
Spearman's correlation	0.7175
Mean squared error	684.12
Root mean squared error	26.16
Mean absolute error	20.43
Bias	-1.40



- Pearson's product moment correlation coefficient – describes the strength of the linear association between the forecasts and the observations;
- Spearman's rank order correlation coefficient – describes the strength of the monotonic association between the forecasts and the observations;
- Mean squared error – defines the average squared difference between each forecast and observation;
- Root mean squared error – the square root of the mean squared error;
- Mean absolute error – the average amount by which the forecast was incorrect;
- **Bias** – the difference between the mean of the forecasts and the mean of the observations.

probabilistic forecast measures based on categories defined from the observations



categorical measures based on the contingency table: hit scores

---

Cross-validated contingency tables

Station 2

WAING 9.62S, 120.22E

Frequency table:

		Forecast			
		B	N	A	Total
Observed	A	0	2	8	10
	N	1	6	3	10
	B	5	4	1	10
Total		6	12	12	30

Contingency table:

		Forecast			
		B	N	A	All
Observed	A	0%	17%	67%	33%
	N	17%	50%	25%	33%
	B	83%	33%	8%	33%
All		20%	40%	40%	100%

Cross-validated scores

Station 2

- **Hit score** – the percentage of times the forecast category corresponds with the observed category (here 19/30);
- **Hit skill score** – the percentage of times, beyond that expected by chance, the forecast category corresponds with the observed category;
- **LEPS score** – the mean absolute difference between the cumulative frequency of the forecast and the cumulative frequency of the observations;
- **Gerrity score** – variant of LEPS score;

Categorical measures:

Hit score	63.33%
Hit skill score	45.00%
LEPS score	55.00%
Gerrity score	52.50%
ROC area (below-normal)	0.8550
ROC area (above-normal)	0.8800

WAING 9.62S, 120.22E

Frequency table:

		Forecast			
		B	N	A	Total
Observed	A	0	2	8	10
	N	1	6	3	10
	B	5	4	1	10
Total		6	12	12	30

Cross-validated scores

Station 2

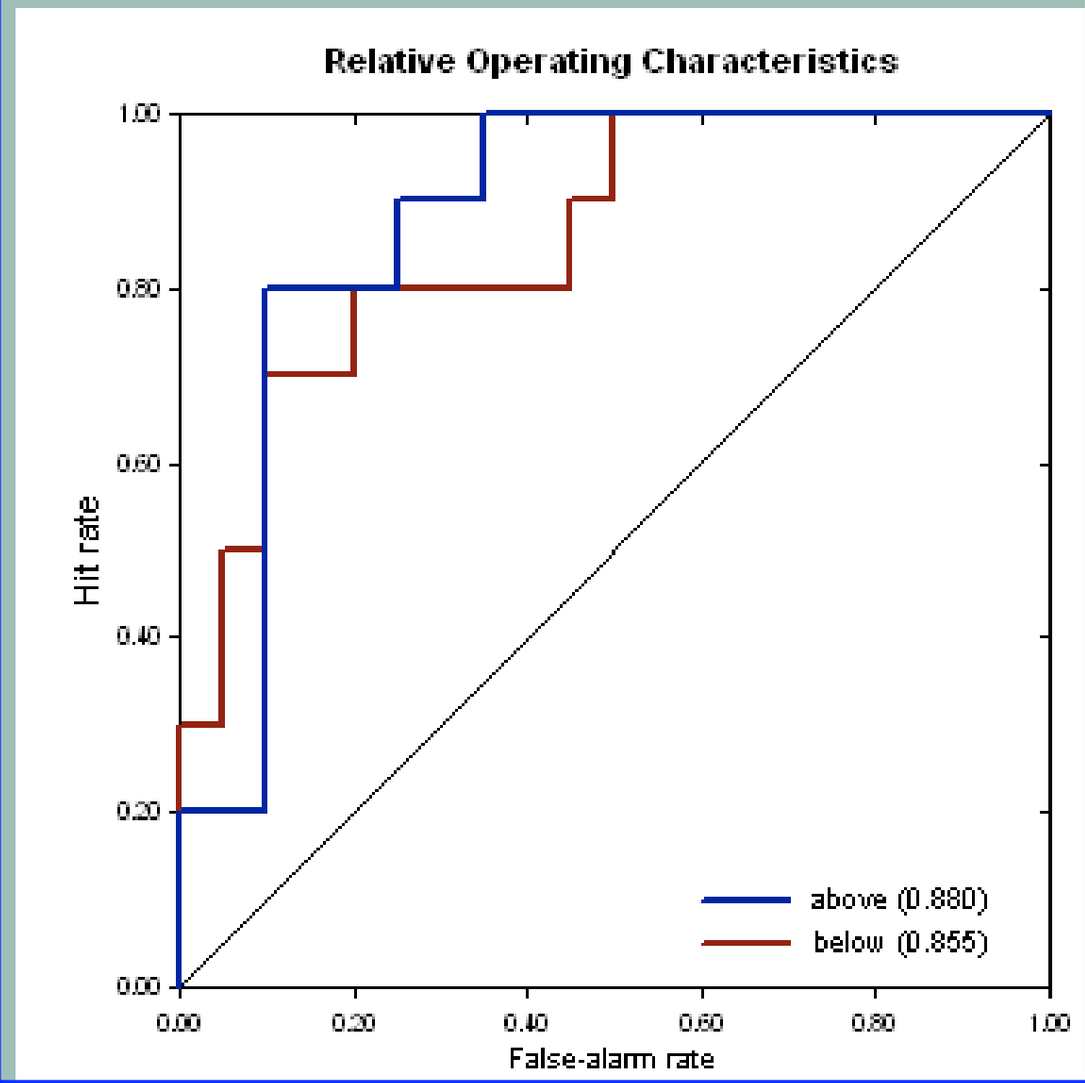
- **ROC area (below-normal)** – the area beneath the ROC curve for forecasts of the below-normal category; gives the proportion of times that below-normal conditions can be distinguished successfully from the other categories;

- **ROC area (above-normal)** – the area beneath the ROC curve for forecasts of the above-normal category; gives the proportion of times that above-normal conditions can be distinguished successfully from the other categories.

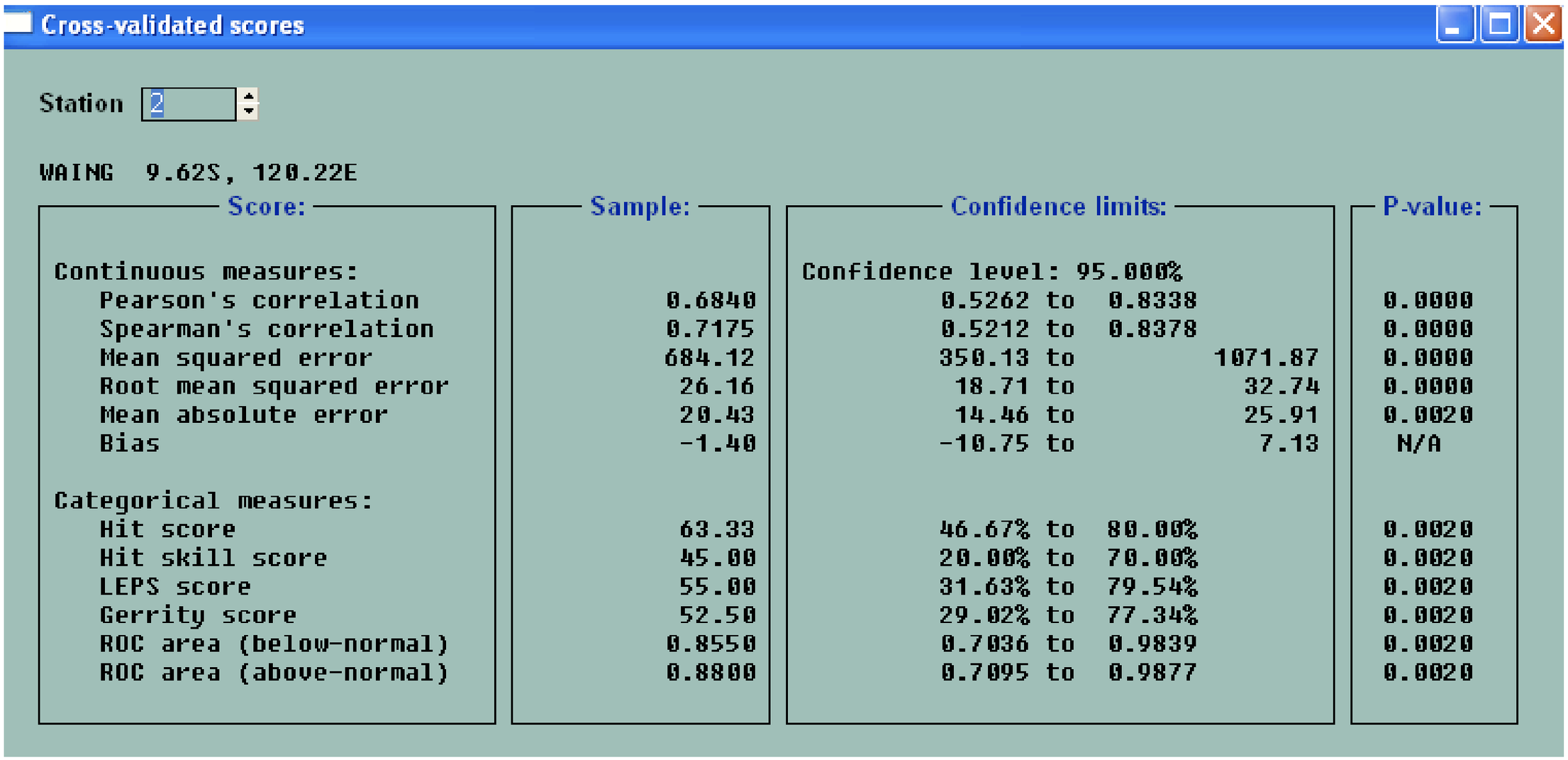
- The forecasts are ranked, and the forecast with the highest value is taken as the most confident forecast for above-normal conditions, and that with the lowest value is taken as the least confident forecast. For forecasts of below-normal conditions, this ranking is inverted so that the forecast with the highest value is taken as the most confident forecast for below-normal conditions, and that with the highest value is taken as the least confident forecast.

Categorical measures:

Hit score	63.33%
Hit skill score	45.00%
LEPS score	55.00%
[illegible]	[illegible]
ROC area (below-normal)	0.8550
ROC area (above-normal)	0.8800



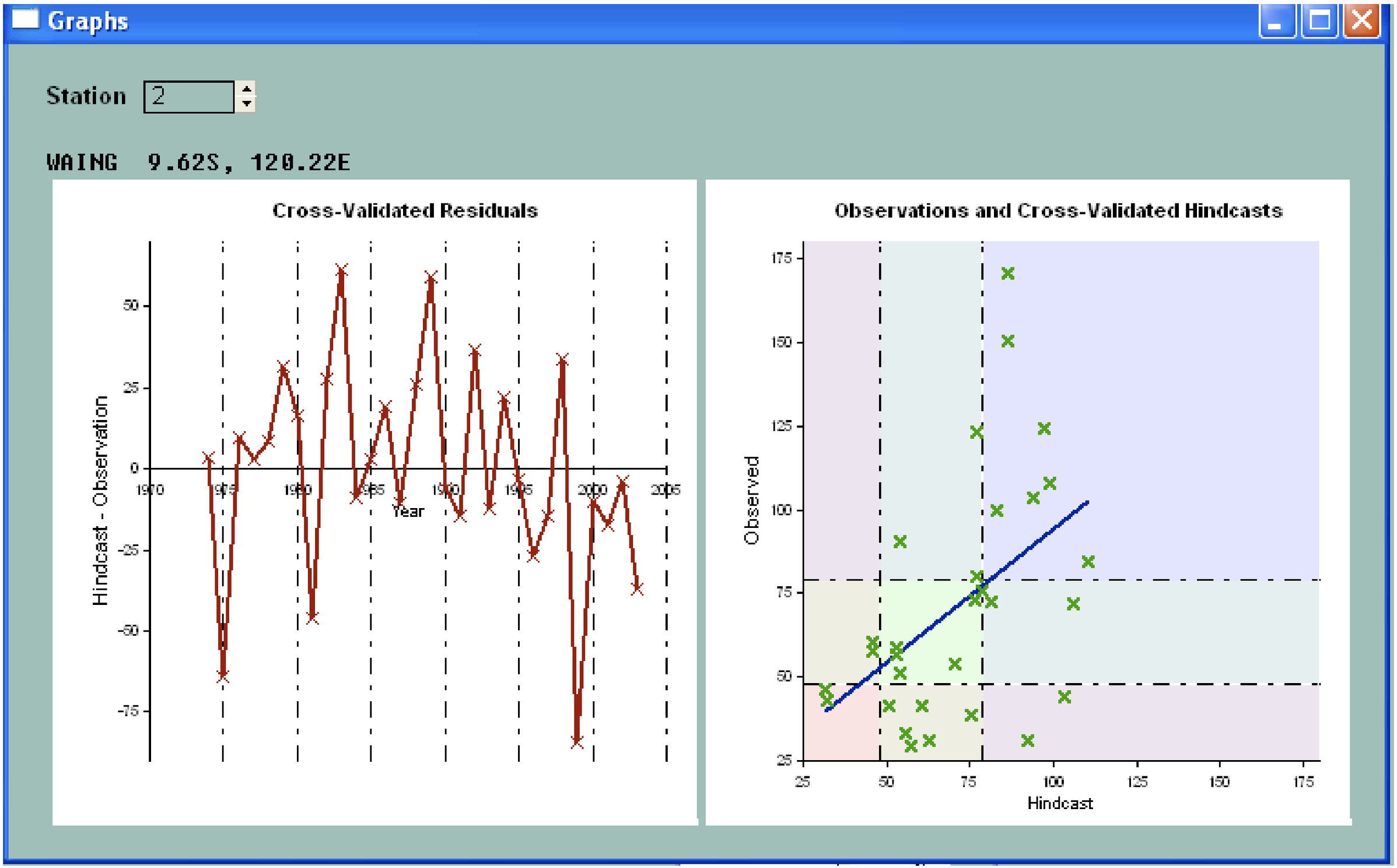
The Bootstrap window – Confidence limits & Significance Tests



- Confidence limits are calculated using bootstrap resampling, and provide an indication of the sampling errors in each performance measure. The bootstrap confidence level used can be adjusted using the Customise~Resampling Settings menu item.

- The p-value indicates the probability that the sample score would be bettered by chance. Permutation procedures are used to calculate the p-values. The accuracy of the p-values depends upon the number of permutations, which can be set using the Customise~Resampling Settings menu item. It is recommended that at least 200 permutations be used, and more if computation time permits.

scatter plots and cross-validated residuals



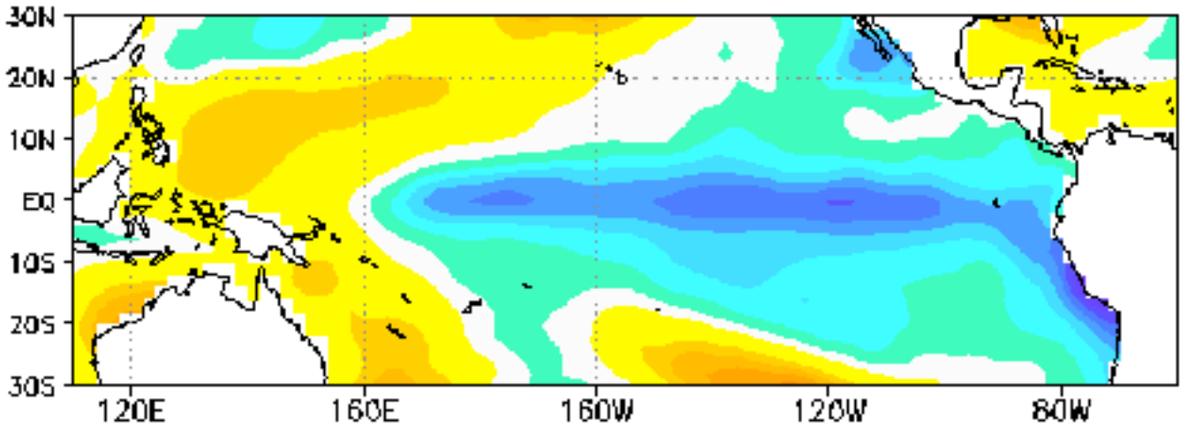
summary

- GCM seasonal forecasts usually need to be transformed in order to make useful regional forecasts for applications
    - this is often referred to as “MOS” or “statistical downscaling”, or sometimes “calibration” or “forecast assimilation”
- regression analysis provides a mathematical framework to combine together GCM predictions with local data
- testing the quality of the tailored forecasts is vital – note the need for cross-validation and probabilistic skill metrics
- the CPT toolbox provides convenient access the CCA and PCR tools, as well as many skill metrics

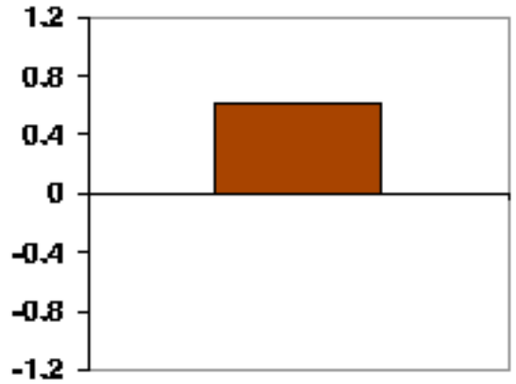
postscript

- CPT may be (hopefully!) easy to learn, but constructing successful tailored forecasts takes years of practice!

SST anomaly DJF 1971



PC Coefficient DJF 1971



PC 1 of SST anomaly 1971-2006 DJF