Tailored Forecast Information

Andrew W Robertson
awr@iri.columbia.edu

International Research Institute
for Climate and Society
EARTH INSTITUTE | COLUMBIA UNIVERSITY

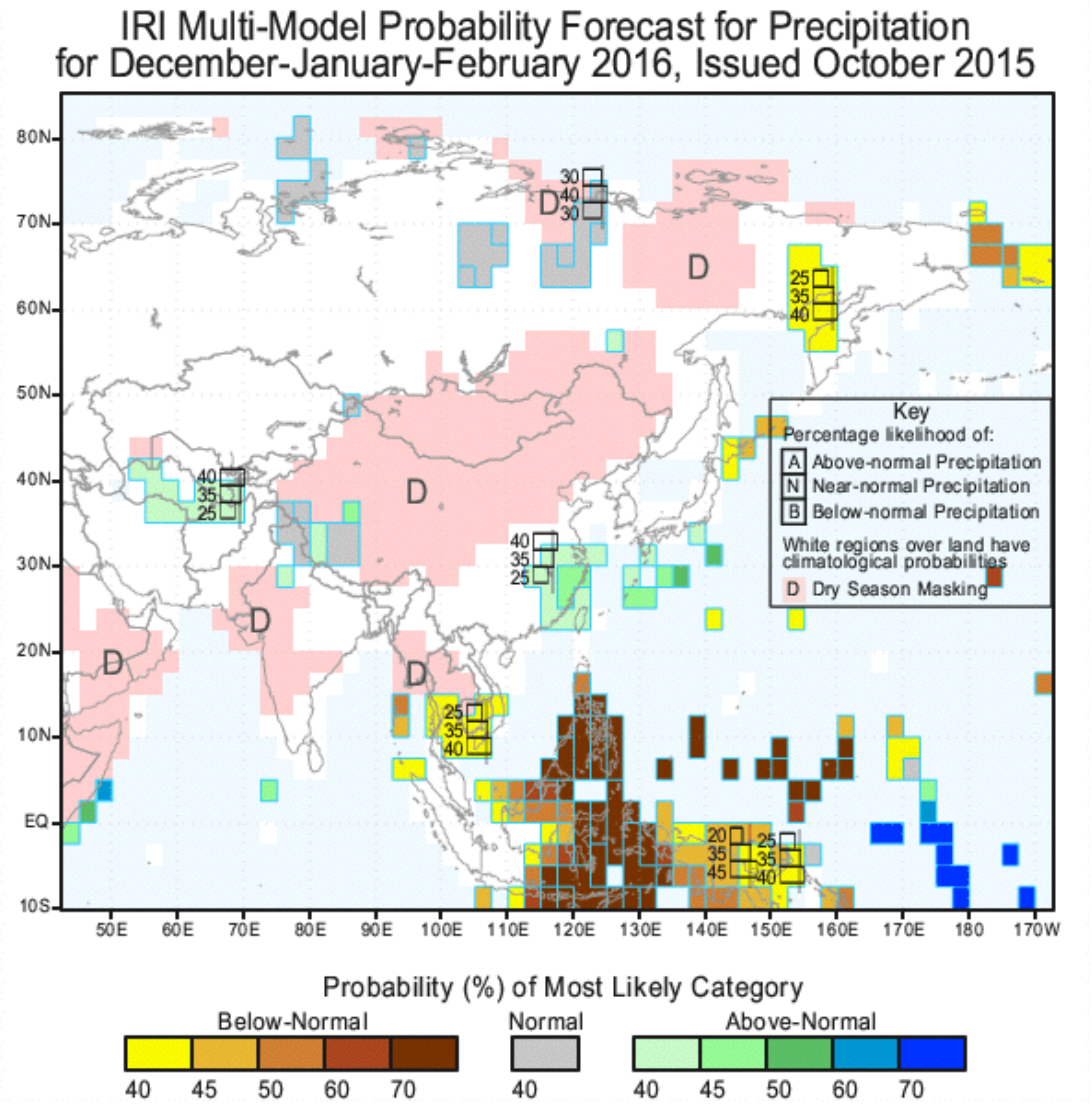
Advanced School and Workshop on Subseasonal to Seasonal (S2S) Prediction and Application to Drought Prediction, ICTP, Trieste, Nov 23 – Dec 4, 2015

Outline

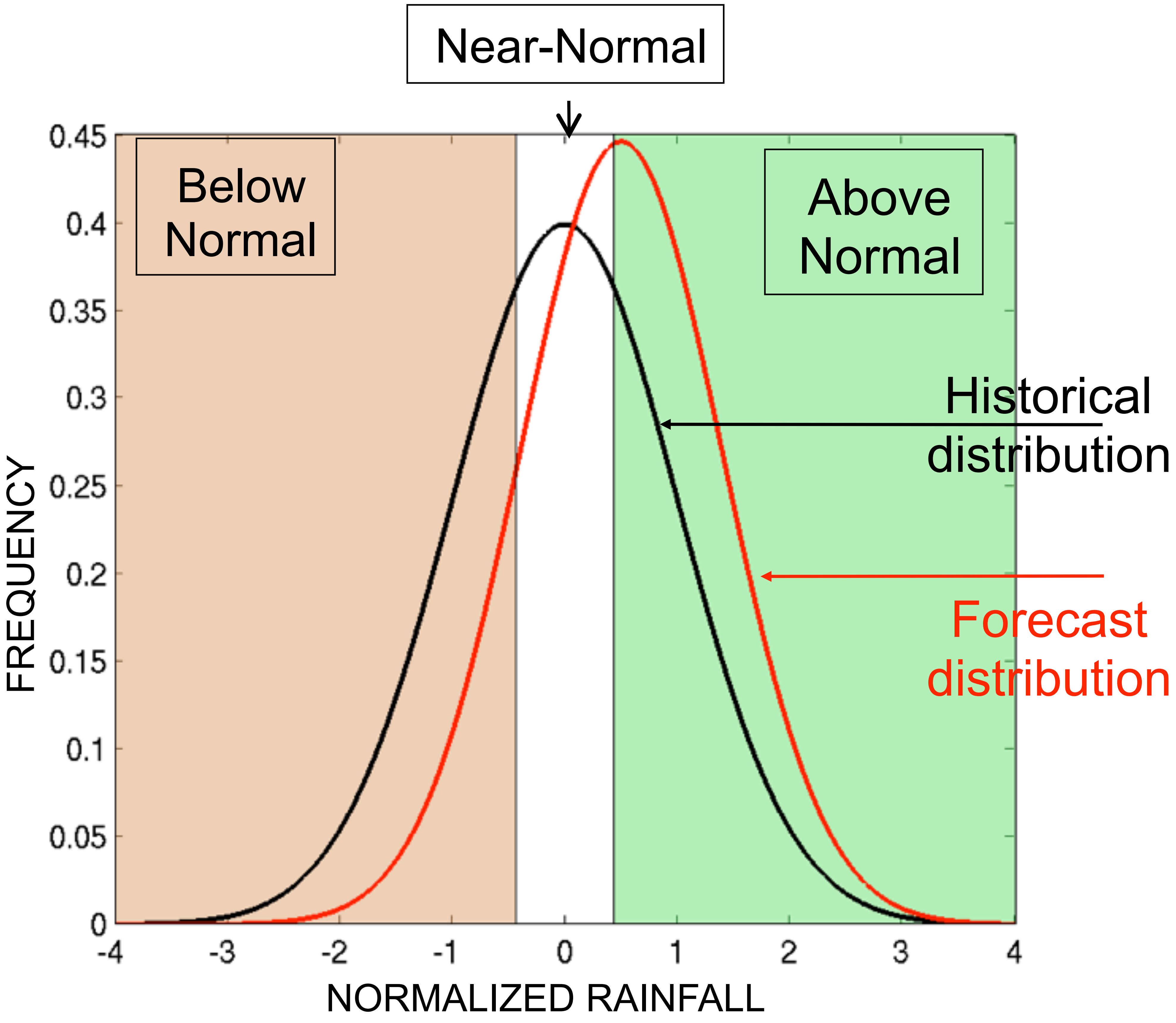
1. Regression models for tailoring and calibrating seasonal forecasts

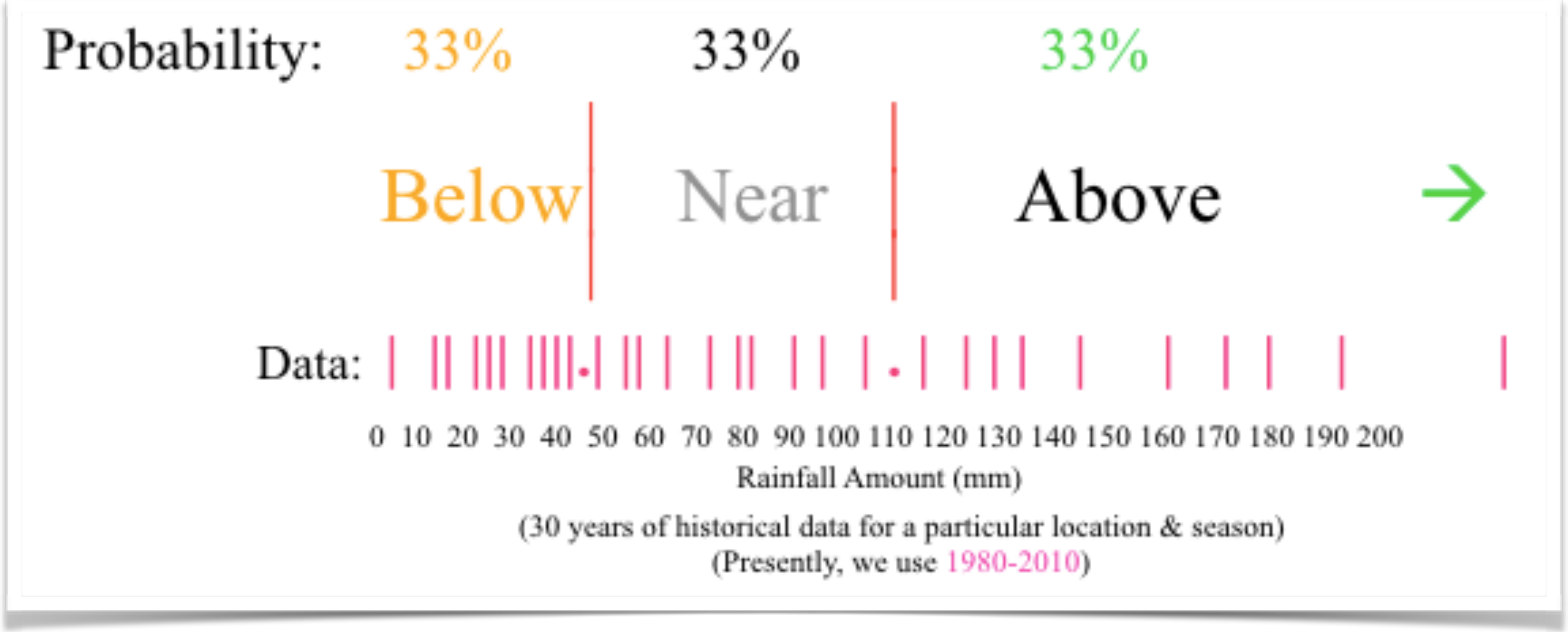
2. Examples of tailored forecasts

3. Quantile regression


IRI Multi-Model Probability Forecast for Precipitation
for December-January-February 2016, Issued October 2015
80N
70N
60N
50N
40N
30N
20N
10N
EQ
10S
50E
60E
70E
80E
90E
100E
110E
120E
130E
140E
150E
160E
170E
180
170W
Key
Percentage likelihood of:
A Above-normal Precipitation
N Near-normal Precipitation
B Below-normal Precipitation
White regions over land have climatological probabilities
D Dry Season Masking
Probability (%) of Most Likely Category
Below-Normal
40 45 50 60 70
Normal
40
Above-Normal
40 45 50 60 70

Displaying forecast probabilities





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.

Linear Regression Models

- Given a set of GCM hindcasts or other predictors x(t) and a set of observations y(t), we can build a regression model to relate them.

  $$y(t) = ax(t) + b + \text{error residual}$$

- In this equation, x(t) is the "predictor" and y(t) is the "predictand"
- b is the mean bias
- The coefficients a and b are estimated by minimizing the sum of squares of the residual error term
- Regression models trained on GCM hindcasts vs historical data are called "MOS Correction" (for Model Output Statistics)
- Generalized linear models can be used for nonlinear relationships

Choice of predictor(s)

$y(t) = ax(t) + b +$ error residual

- Note that in this equation, x(t) does not have to be the same physical quantity as y(t).
- Multiple linear regression is often used to get smaller error residual.
- This leads to the main pitfall. Fact: the error residual can be reduced to *zero* by including enough *random* predictors. How many?
  - ➡ If too many predictors are included, this is called *overfitting.*
  - ➡ Rule of thumb: need 5–10 samples per predictor
- Raises the question of how to choose x(t)'s?
  - ➡ Golden rule: (1) predictors need to be chosen from *physical considerations*, and (2) the model error (or skill) needs to be estimated using *independent data*

Choice of Predictand

$$y_t = ax_t + b + error$$

- The predictand could be station-scale precip, yielding a statistical downscaling

- It could even be a more-relevant variable like reservoir inflow, or crop production data

➡ we can thus “tailor” the forecasts to specific users using regression models

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)

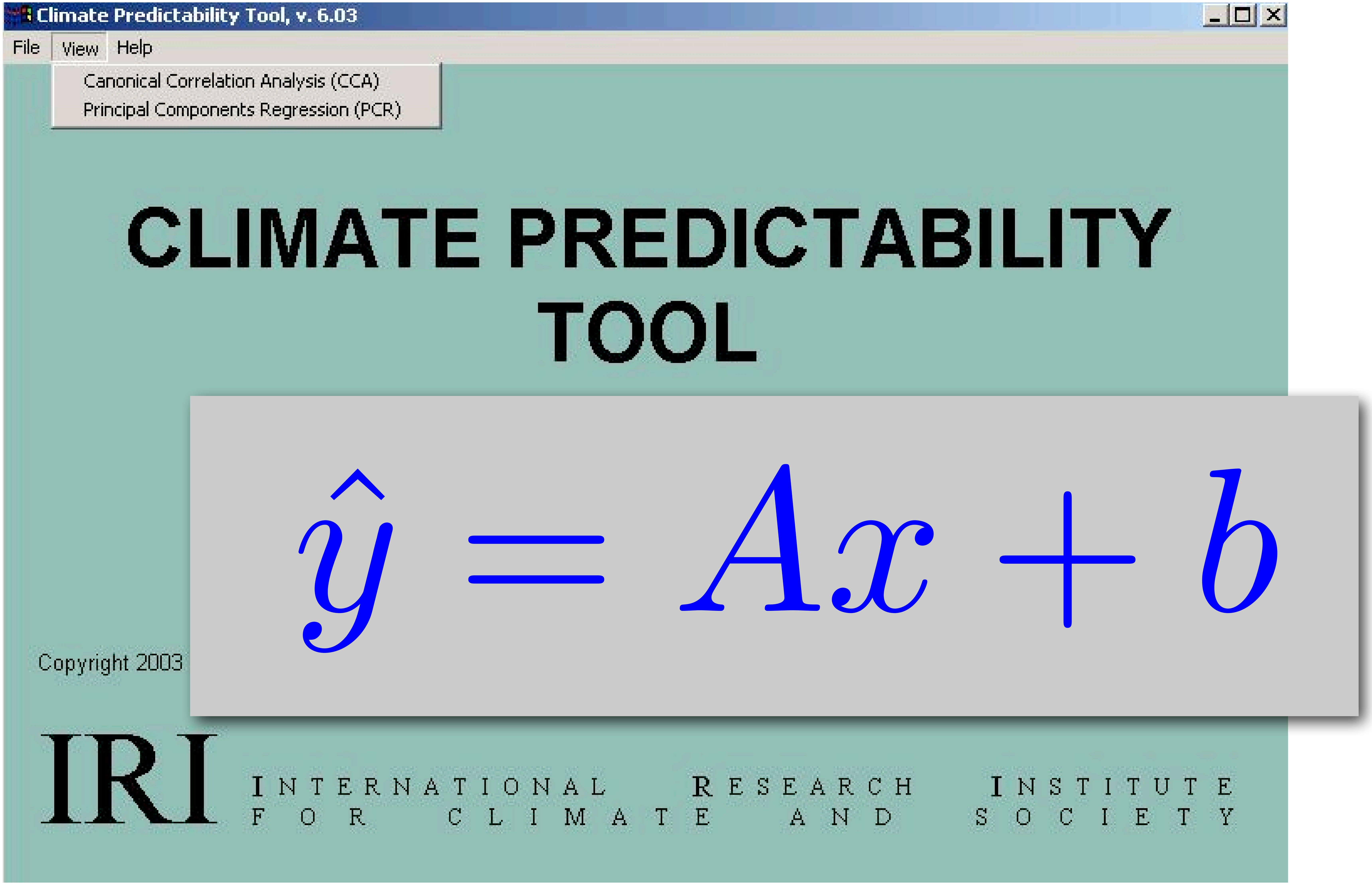
  -- e.g., **Principal Components Regression (PCR)**

- multivariate (pattern) regression: **two or more** predictors, **two or more** predictands
  $\mathbf{y} = \mathbf{Ax+b}$ (matrix **A**, vectors **y,x,b**)

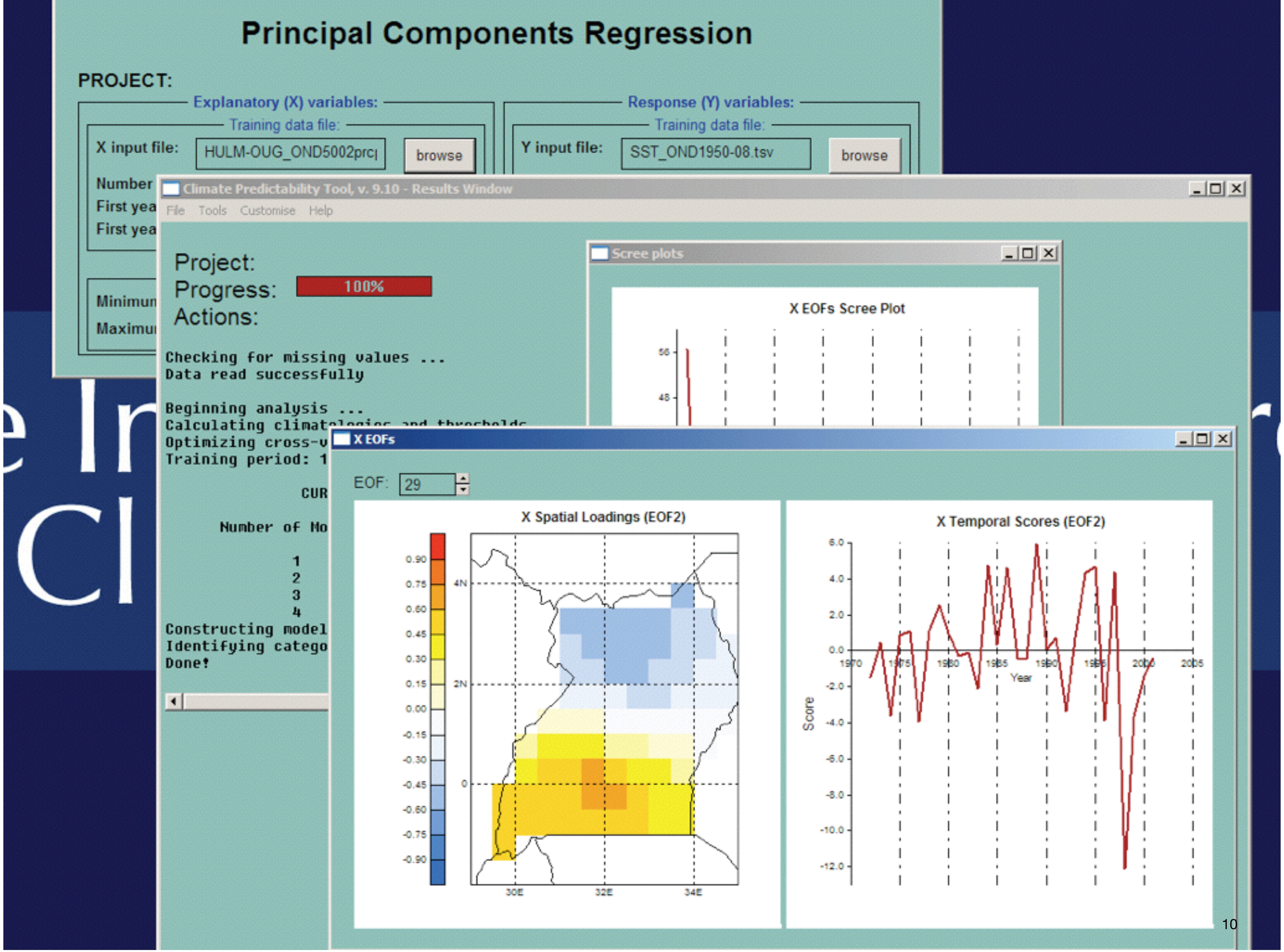
  -- e.g., **Canonical Correlation Analysis (CCA)**

IRI Tool for MOS correction & downscaling seasonal forecasts

motivated by experience at Climate Outlook Fora (COFs) in Africa



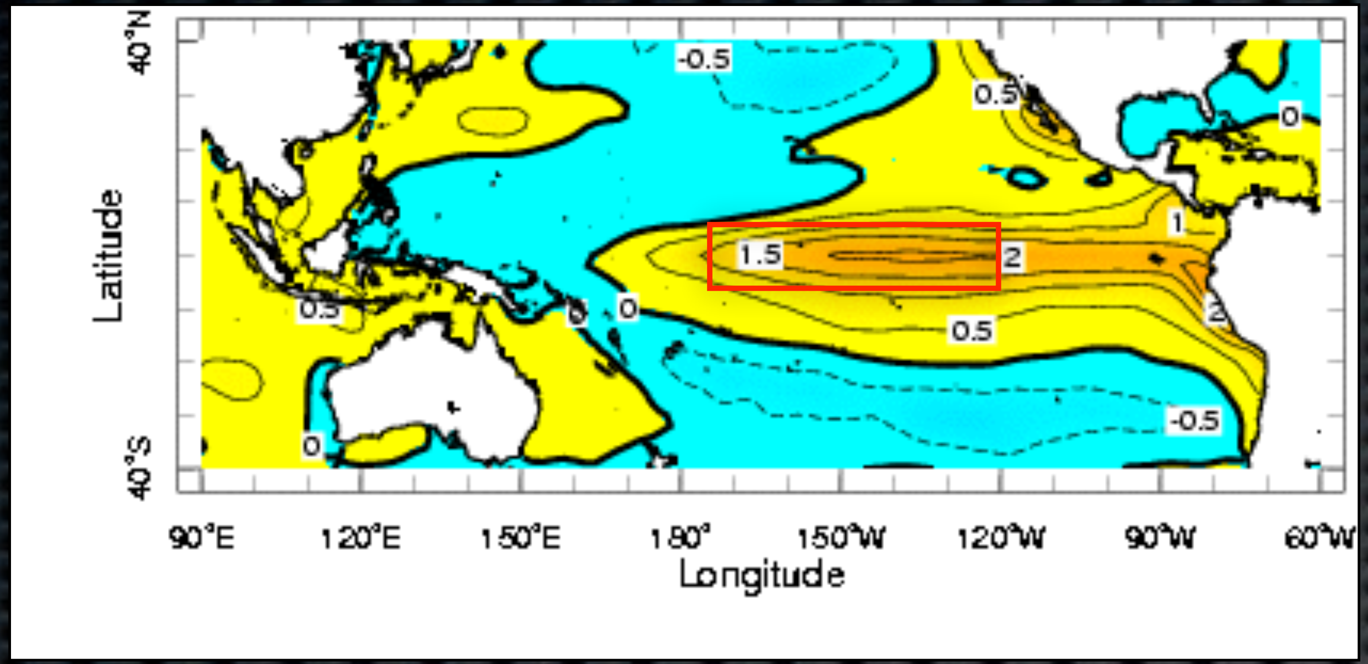
$$\hat{y} = Ax + b$$


Principal Components Regression
PROJECT:
Explanatory (X) variables:
Training data file:
X input file:
HULM-OUG_OND5002prc
browse
Response (Y) variables:
Training data file:
Y input file:
SST_OND1950-08.tsv
browse
Climate Predictability Tool, v. 9.10 - Results Window
File Tools Customise Help
Project:
Progress:
100%
Actions:
Checking for missing values ...
Data read successfully
Beginning analysis ...
Scree plots
X EOFs Scree Plot
X EOFs
EOF: 29
X Spatial Loadings (EOF2)
X Temporal Scores (EOF2)
Year
Score
Constructing model
Done!

Tailoring seasonal forecasts to reservoir inflow

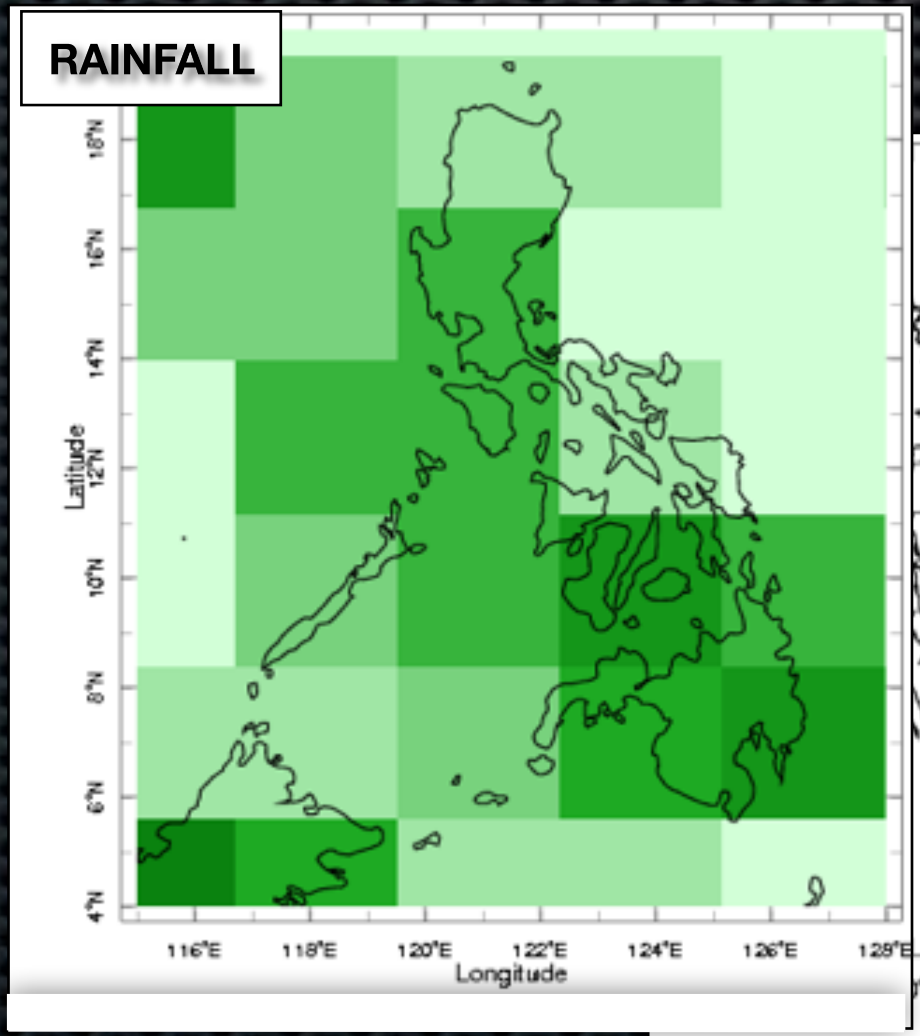
B. Lyon (IRI)
A. Lucero (PAGASA)

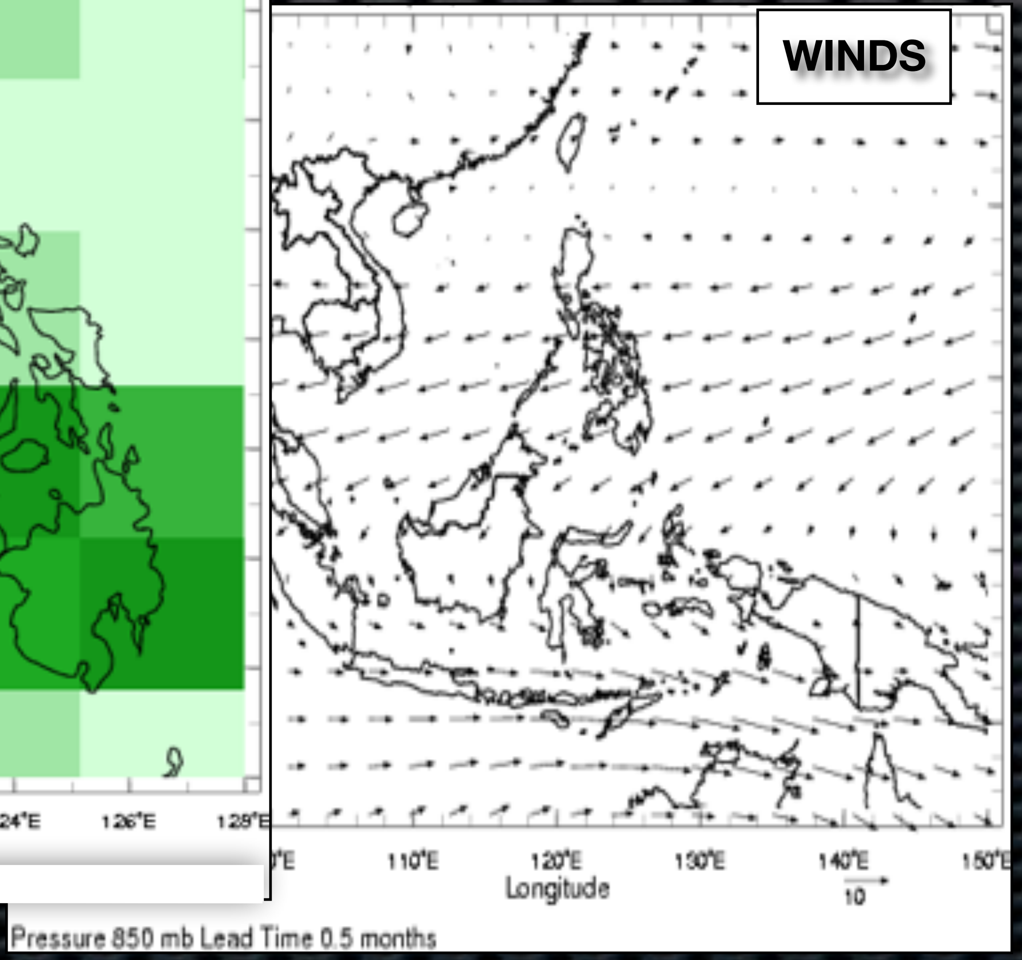
Sea Surface Temperatures



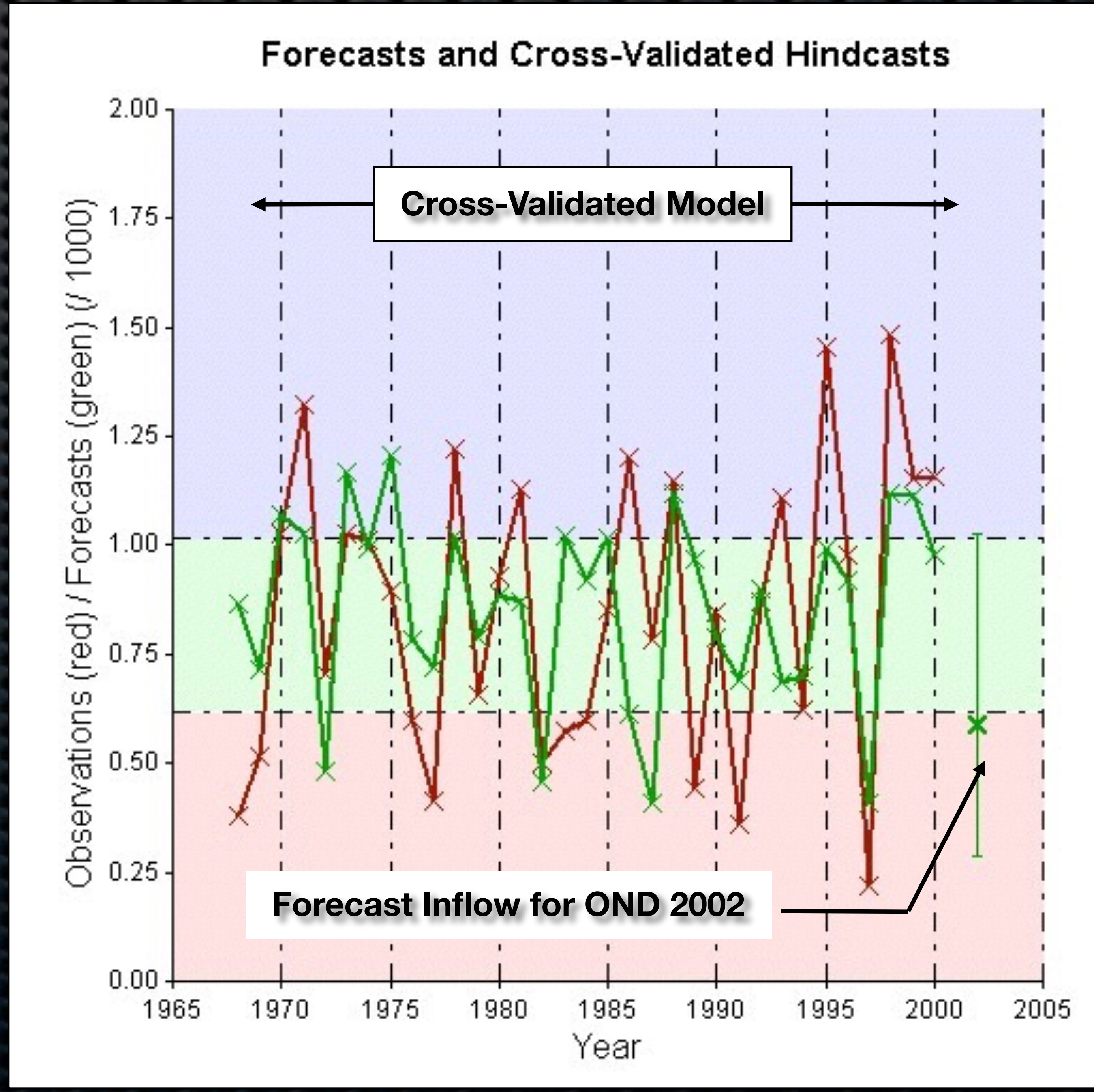
Global Climate Model

Historical Angat Inflow Observations





Statistical Model



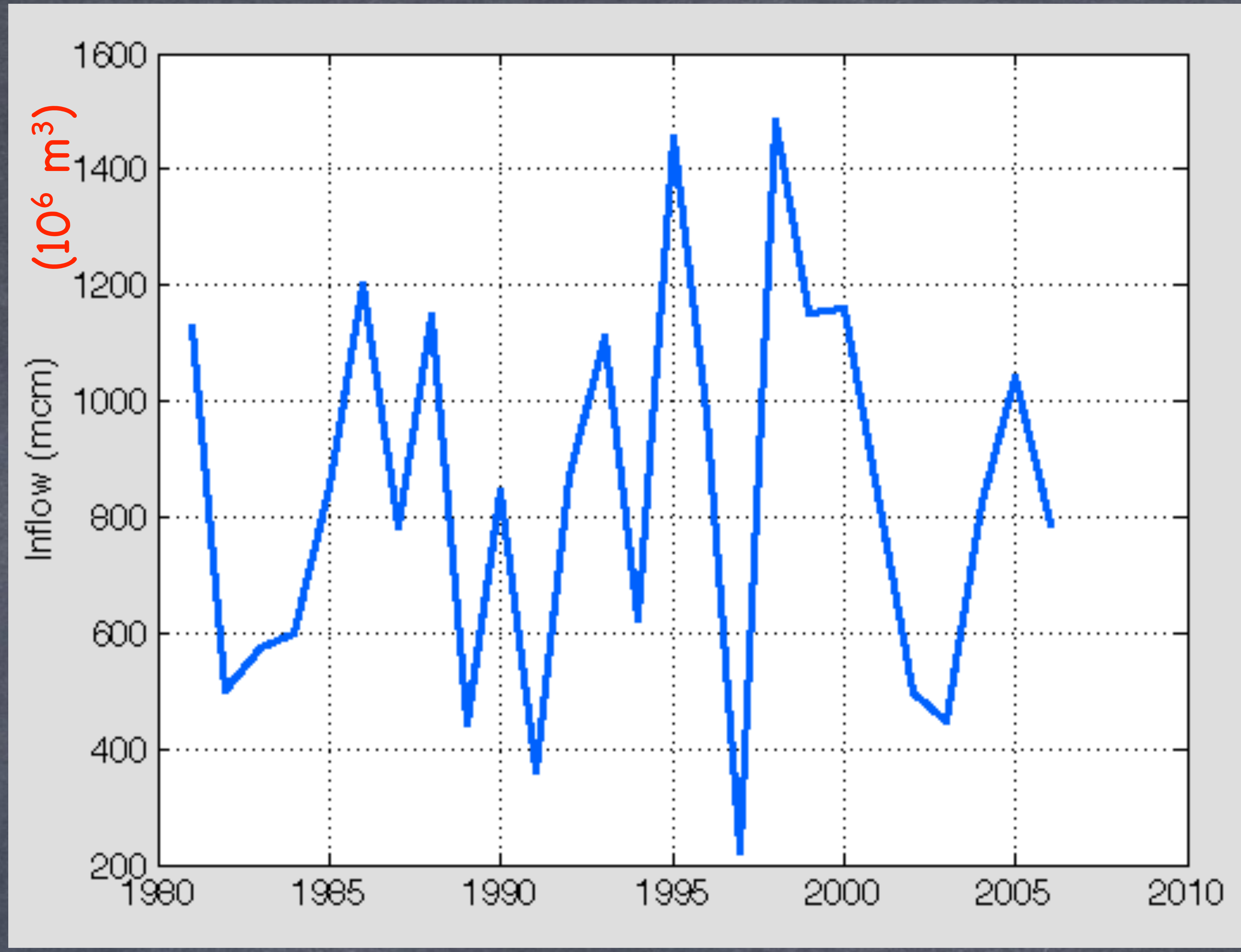
Basic requirements for regression model fitting

$$y_t = ax_t + b + \text{error}$$

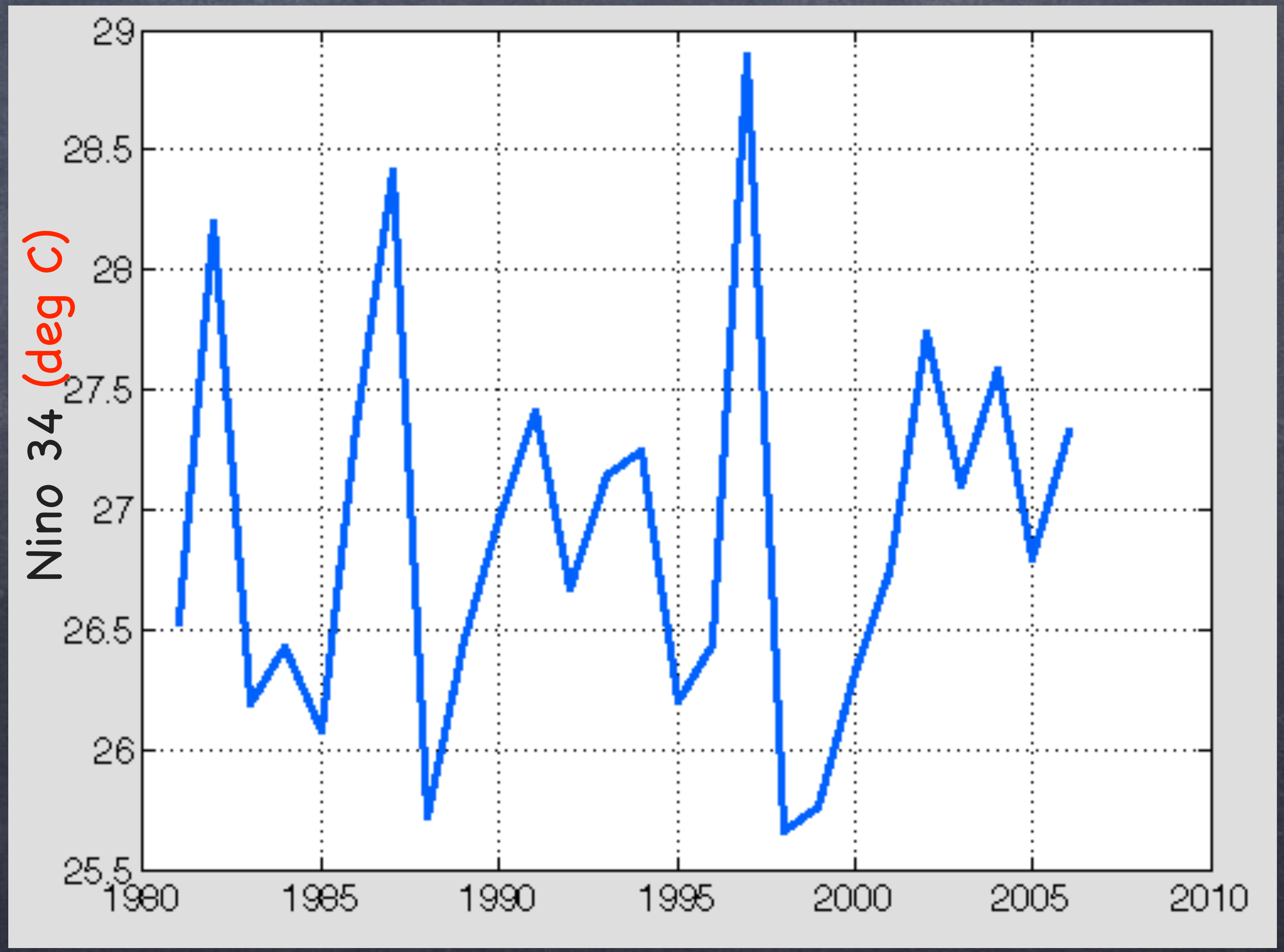
- A <u>long</u> historical time series of y (OND streamflow)
- A <u>matching</u> historical time series of x (e.g. September Nino3.4 SST)

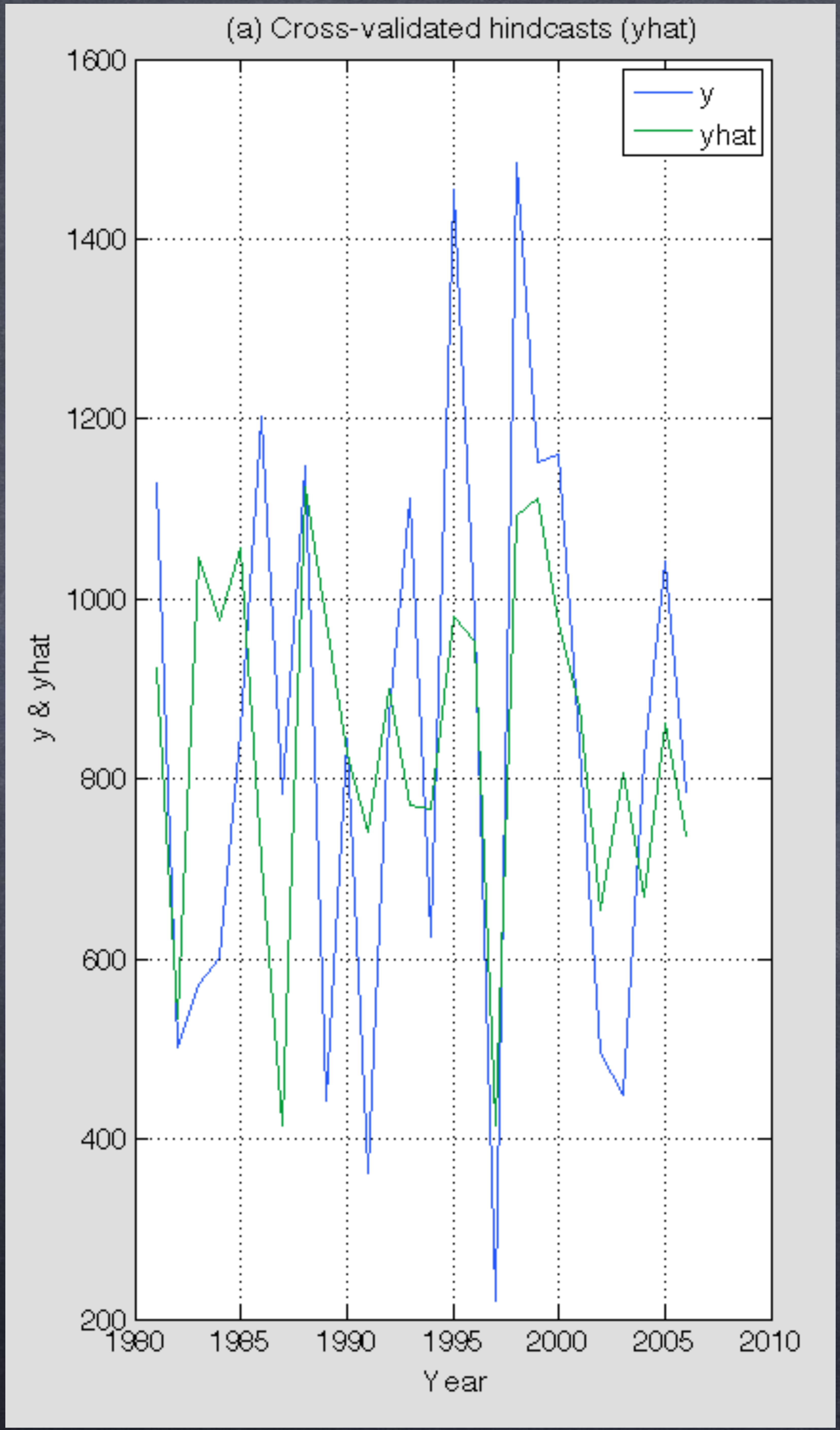
	Sept Nino3.4 x (C)	OND Inflow y (10^8 m^3)
1981	26.5285	11.2799
1982	28.2017	5.009
1983	26.1886	5.7266
1984	26.4288	6.0093
1985	26.0805	8.5389
1986	27.352	12.021
1987	28.4074	7.8353
1988	25.7203	11.4695
1989	26.4601	4.4186
1990	26.9618	8.4525
1991	27.4065	3.6189
1992	26.6667	8.6925
1993	27.1416	11.1192
1994	27.2457	6.2394
1995	26.1964	14.5434
1996	26.4368	9.7648
1997	28.8881	2.2057
1998	25.6589	14.8412
1999	25.7636	11.5271
2000	26.3237	11.5968
2001	26.7461	8.394
2002	27.7331	4.9591
2003	27.1061	4.4899
2004	27.5801	8.2306
2005	26.7958	10.4253
2006	27.3255	7.8595

y:



x:





Cross-validated Hindcasts of OND Inflow

Hindcast =
forecast made for previous years

Cross-validation =
the year to be forecast is excluded from the data used to train the model used for that year, to mimic the eal-time forecast situation, and prevent statistical "overfitting"

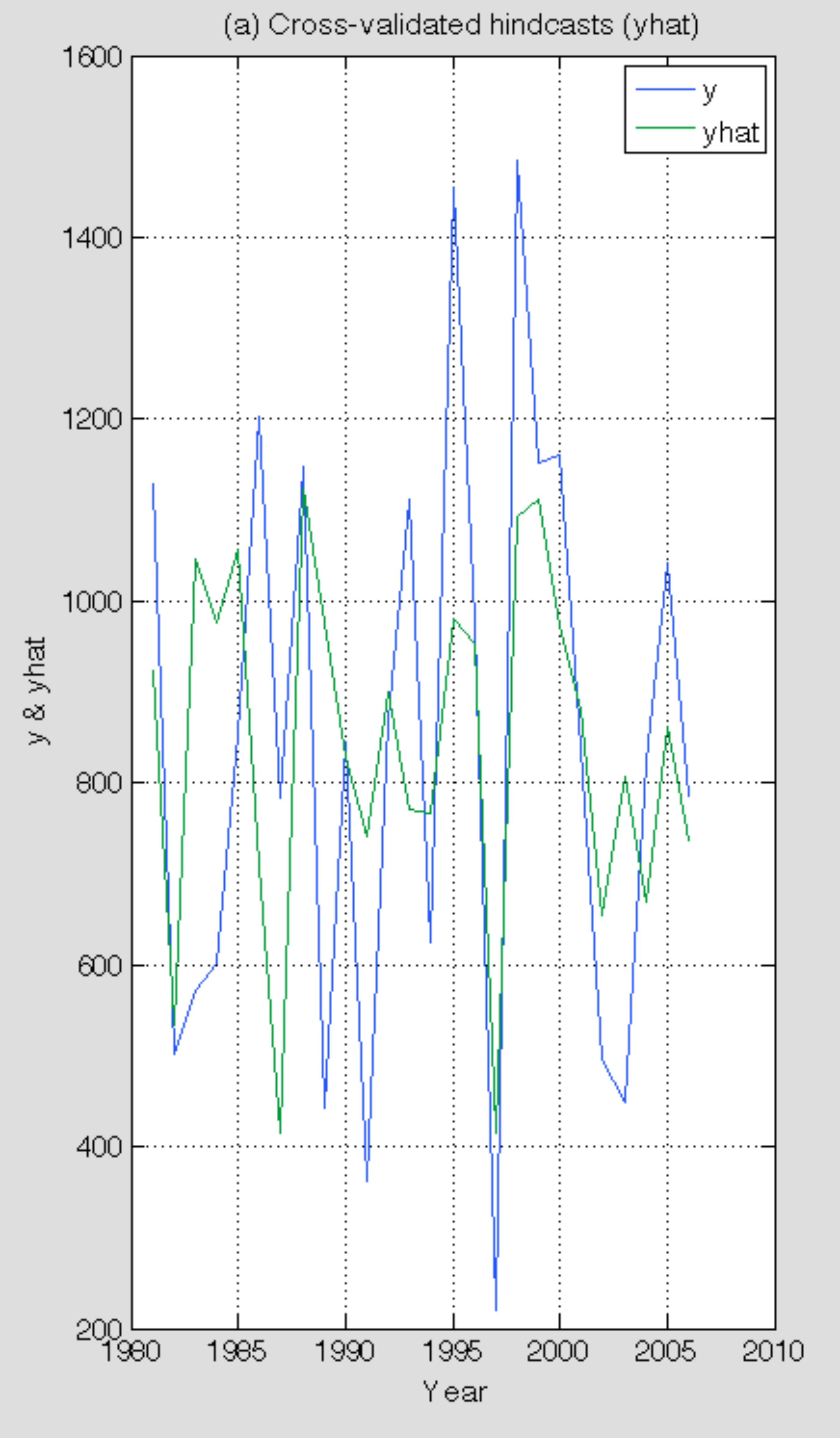
Leave-one-out cross-validation

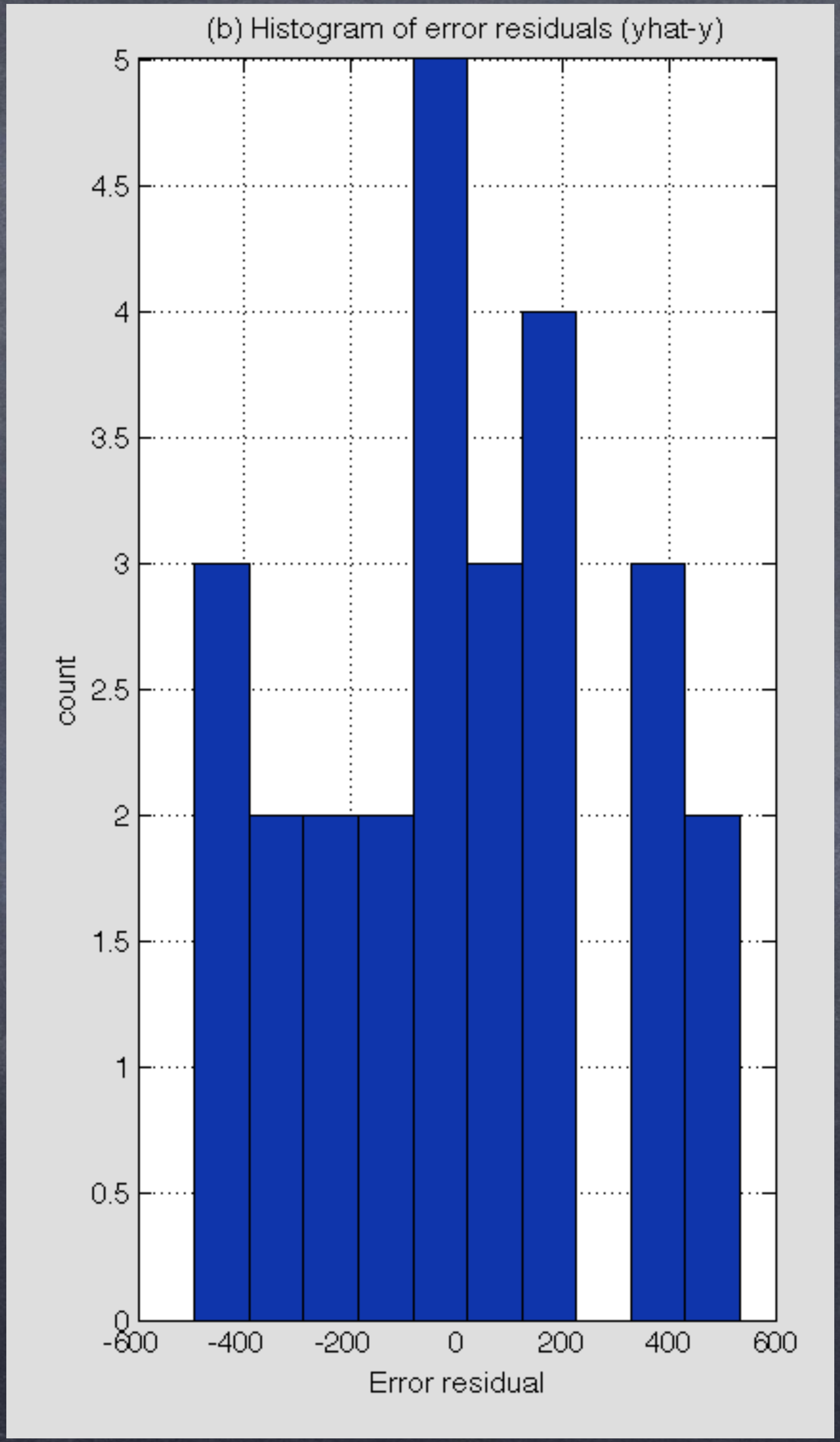
Leave-one-out cross-validation

<table>
<tr><td>1951</td><td>Predict 1951</td><td colspan="5">Training period</td></tr>
<tr><td>1952</td><td>Training period</td><td>Predict 1952</td><td colspan="4">Training Period</td></tr>
<tr><td>1953</td><td colspan="2">Training period</td><td>Predict 1953</td><td colspan="3">Training period</td></tr>
<tr><td>1954</td><td colspan="3">Training period</td><td>Predict 1954</td><td colspan="2">Training Period</td></tr>
<tr><td>1955</td><td colspan="4">Training period</td><td>Predict 1955</td><td>Training period</td></tr>
</table>

… then correlate 1951–2000.

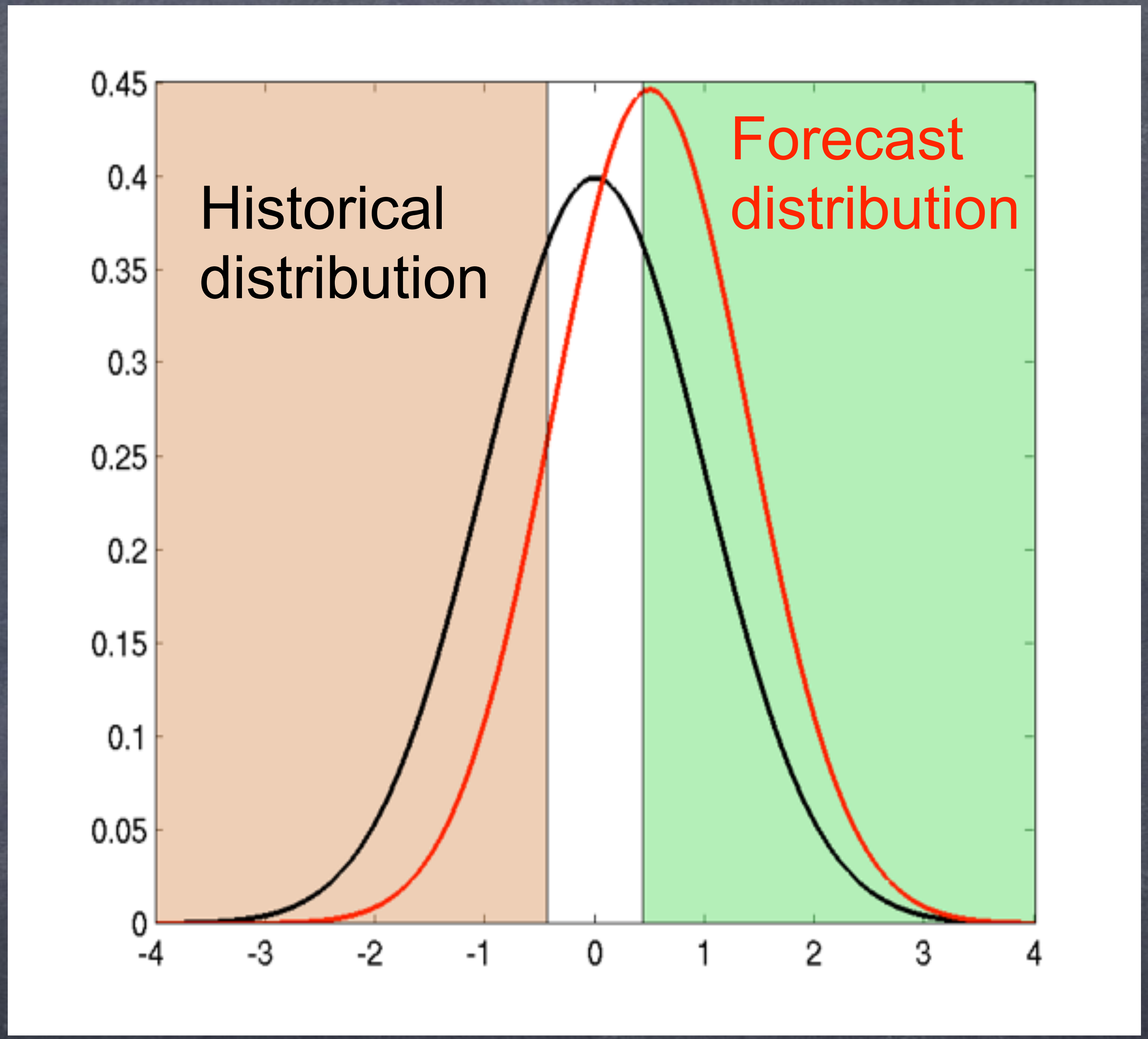
Error residuals of OND-inflow hindcasts



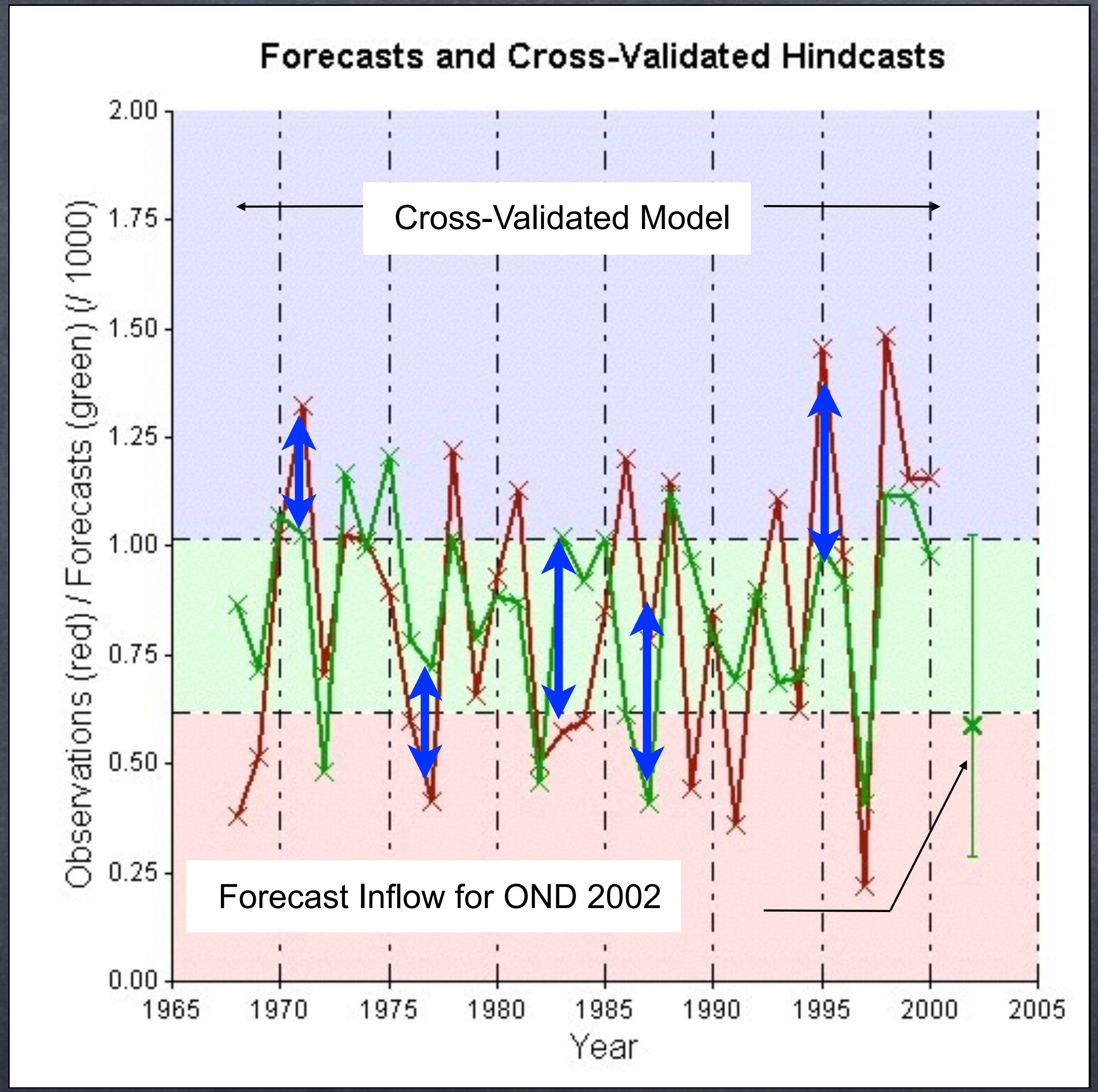


These residuals should be approx. normally distributed for the regression assumptions to be valid.

- How do we get the forecast PDF using regression models?
  - Assume a normal distribution (transformation can be applied), with the mean given by the regression model
  - Estimate the spread from the errors of past forecasts



How do we make probabilistic forecasts from this? Assume a normal distribution with mean given by regression prediction y(t)

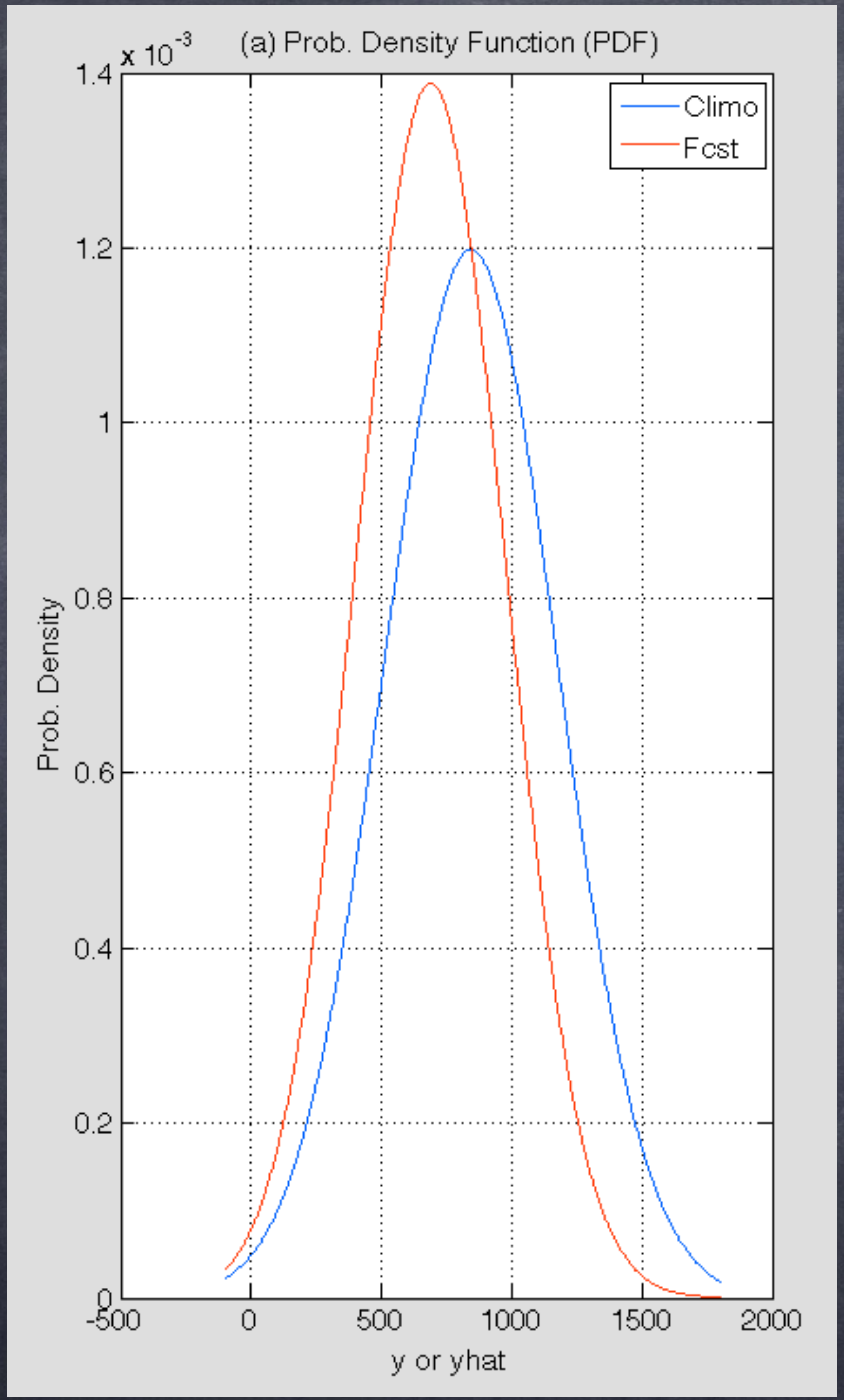


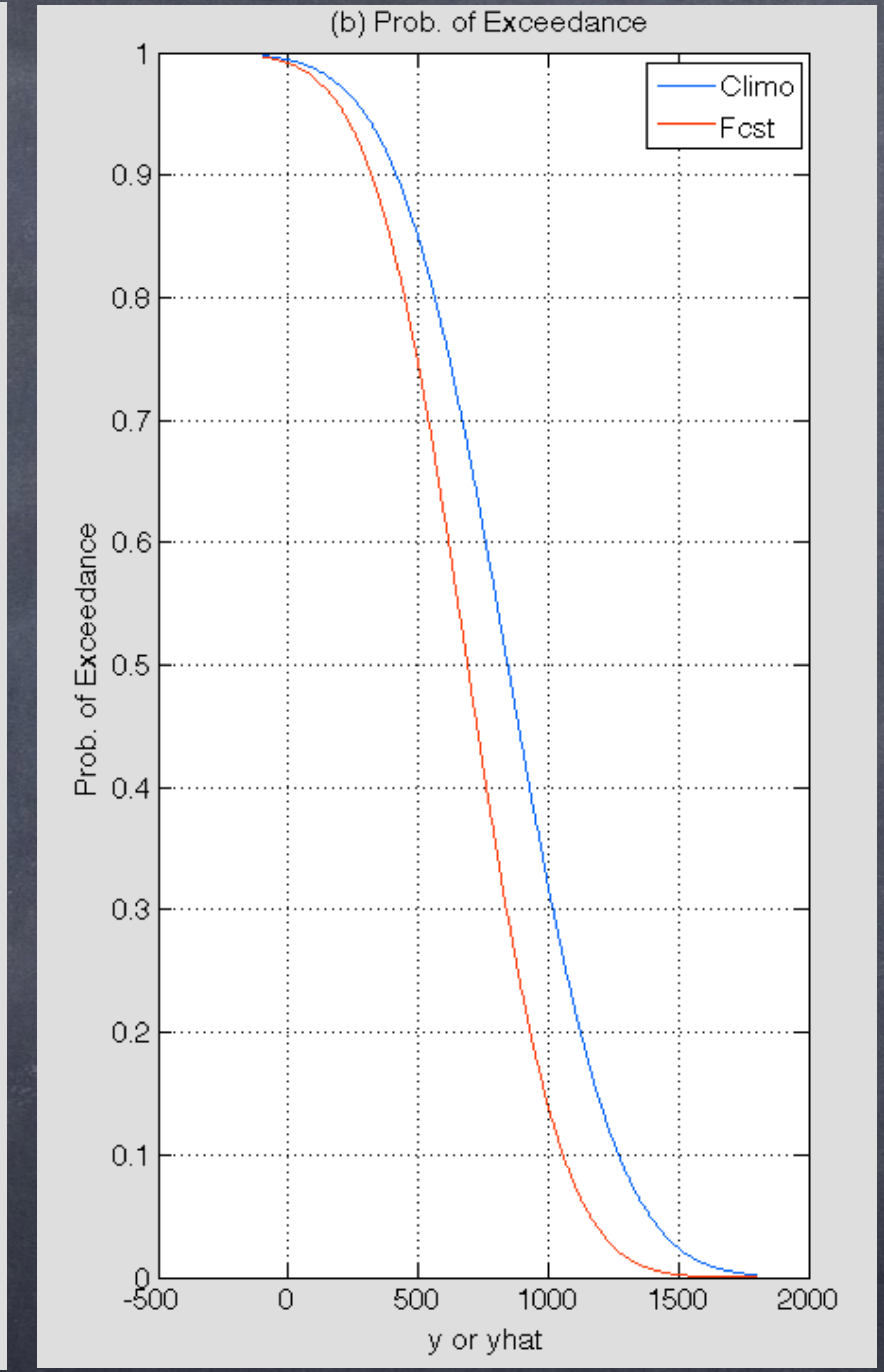
$$y_f \sim N(\hat{y}, \sigma^2_{errors})$$

$$\sigma^2_{errors} = \frac{1}{N-1}\sum_{i=1}^{N}(y-\hat{y})^2$$

yhat - result of regression model; y - obs data. The forecast distribution y_f is assumed to be normal with mean yhat and variance from the squared errors of the cross-validated hindcasts.

Probabilistic forecast of 2009 OND-inflow





Regression results:

===================

Cross-validated anomaly correlation skill = 0.512

2009 Forecast distribution mean = 689

2009 Forecast distribution standard deviation = 287

Climatological distribution mean & st devn: 843, 333

Predictability of Philippines Rice Production from GCM hindcasts and published rice production data

Jan–Jun (Dry Season)
from prev. Jun 1

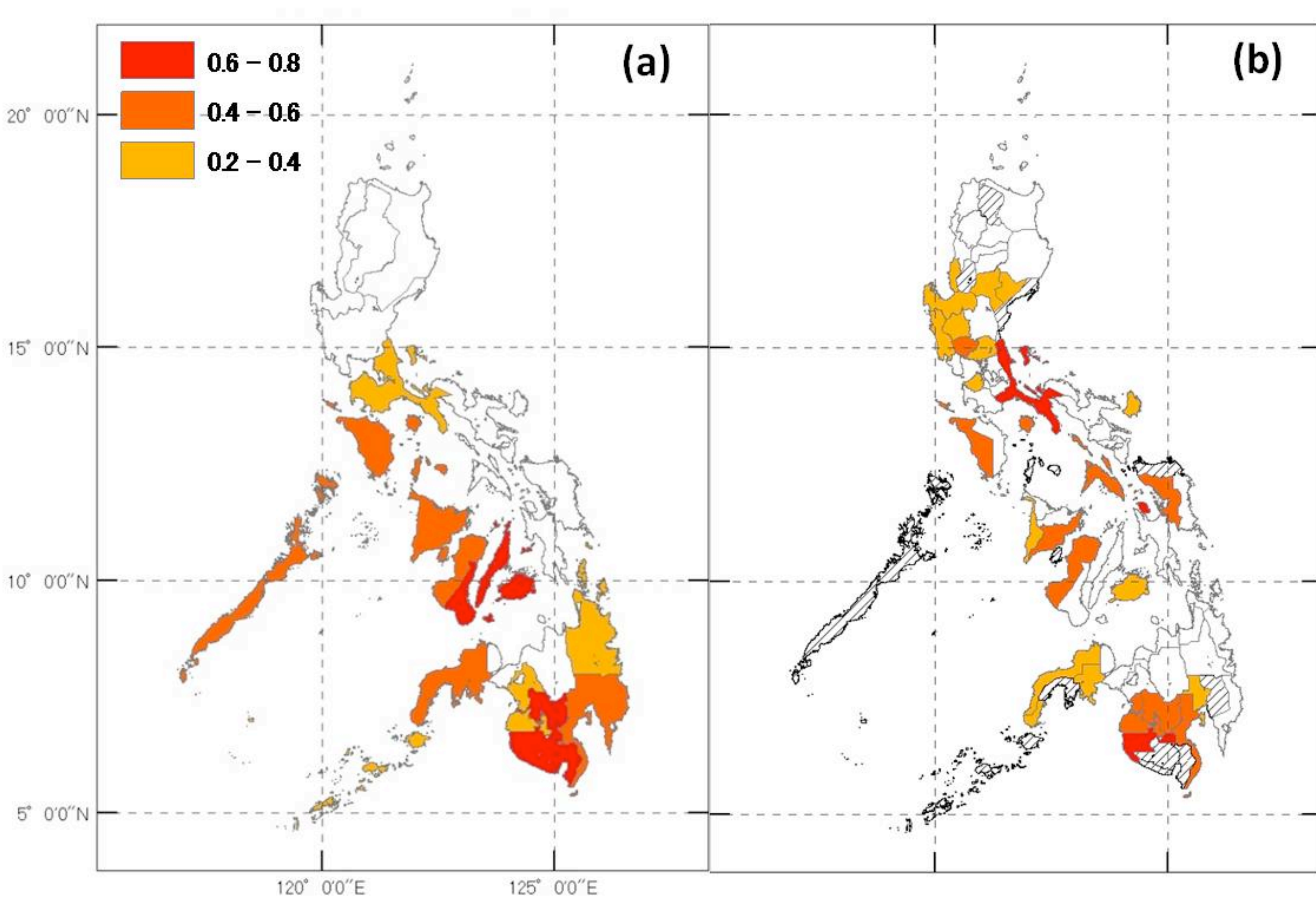
Canonical correlation analysis of
x=GCM predicted Oct-Dec precip,
y=Jan-Jun rice production

y=**Ax**+**b**

1980–2007

Koide et al (2013, JAMC)

ACC Skill of (a) Regional & (b) Provincial Production



Statistical Hindcasts of Monsoon Onset Date

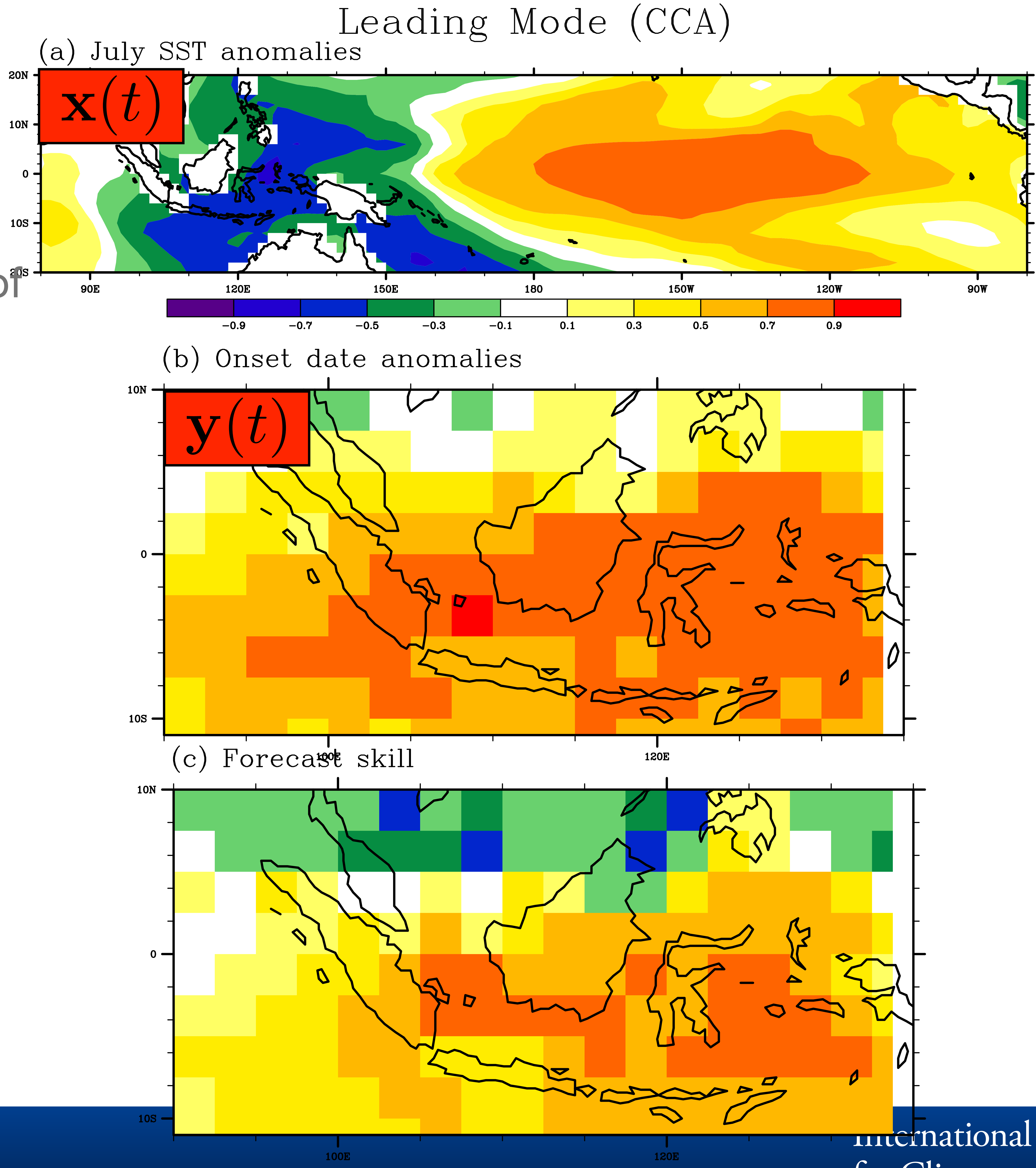
- Canonical correlation analysis of CMAP onset dates vs. July monthly SST field

$$\mathbf{y}(t) = \mathbf{A}.\mathbf{x}(t) + \mathbf{C}$$

$$t : 1979, 1980, \ldots, 2009.$$

- Cross-validated anomaly correlation skill

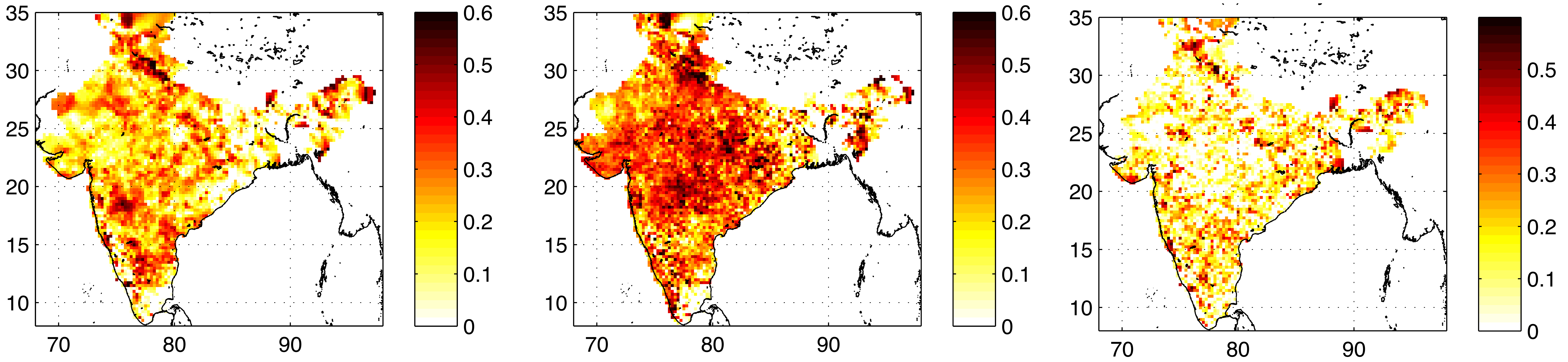
$$r(\hat{y}(t), y(t))$$



after Moron, Robertson & Boer (2009)

Seasonal predictability of daily rainfall statistics

Seasonal Total = Rain Day Frequency x Mean Intensity



IMD 0.25-degree daily rainfall data

Anomaly Correlation "Skill"

Cross-validated regression with observed tropical Indo-Pacific SST

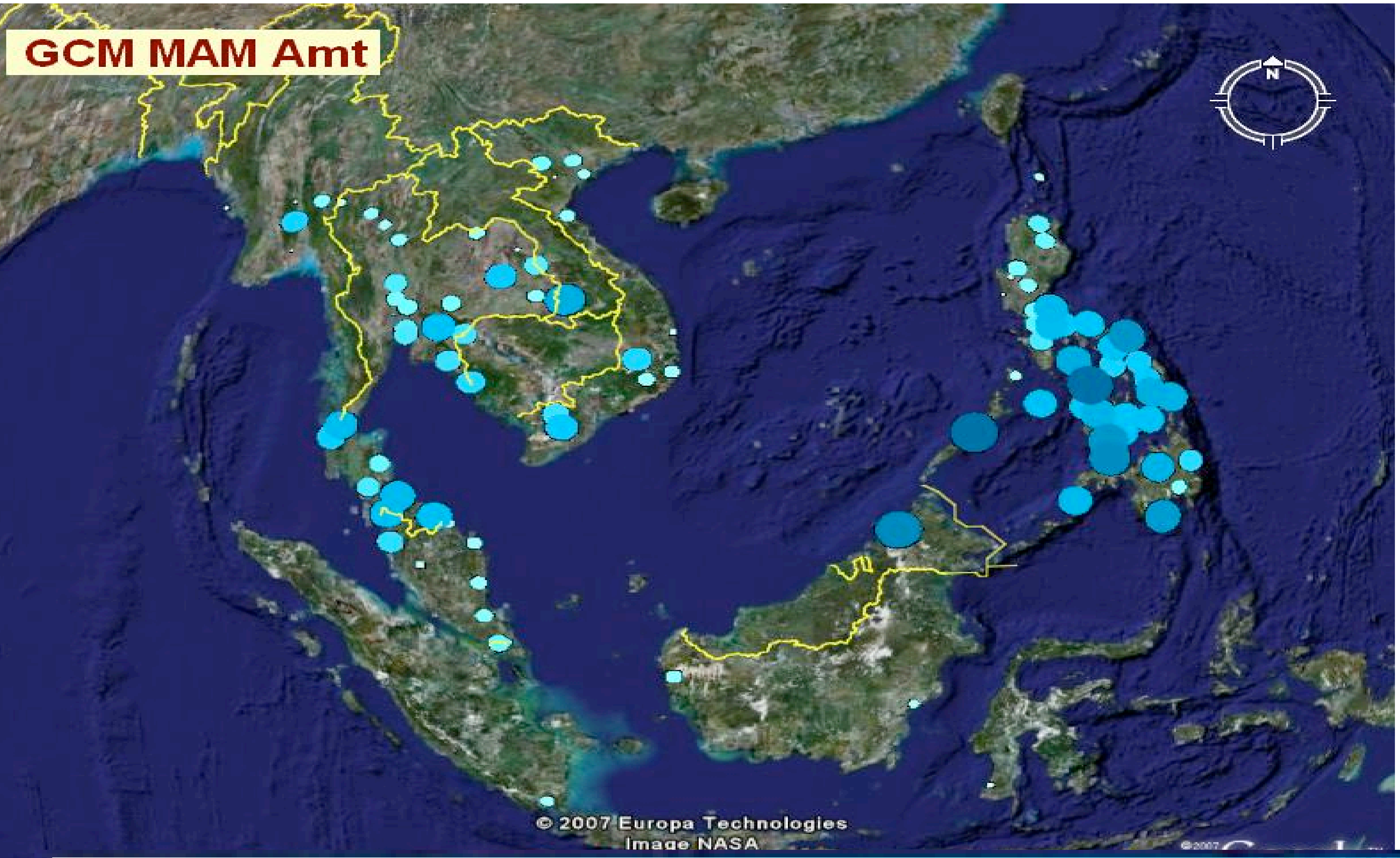
Jun–Sep 1901–2004

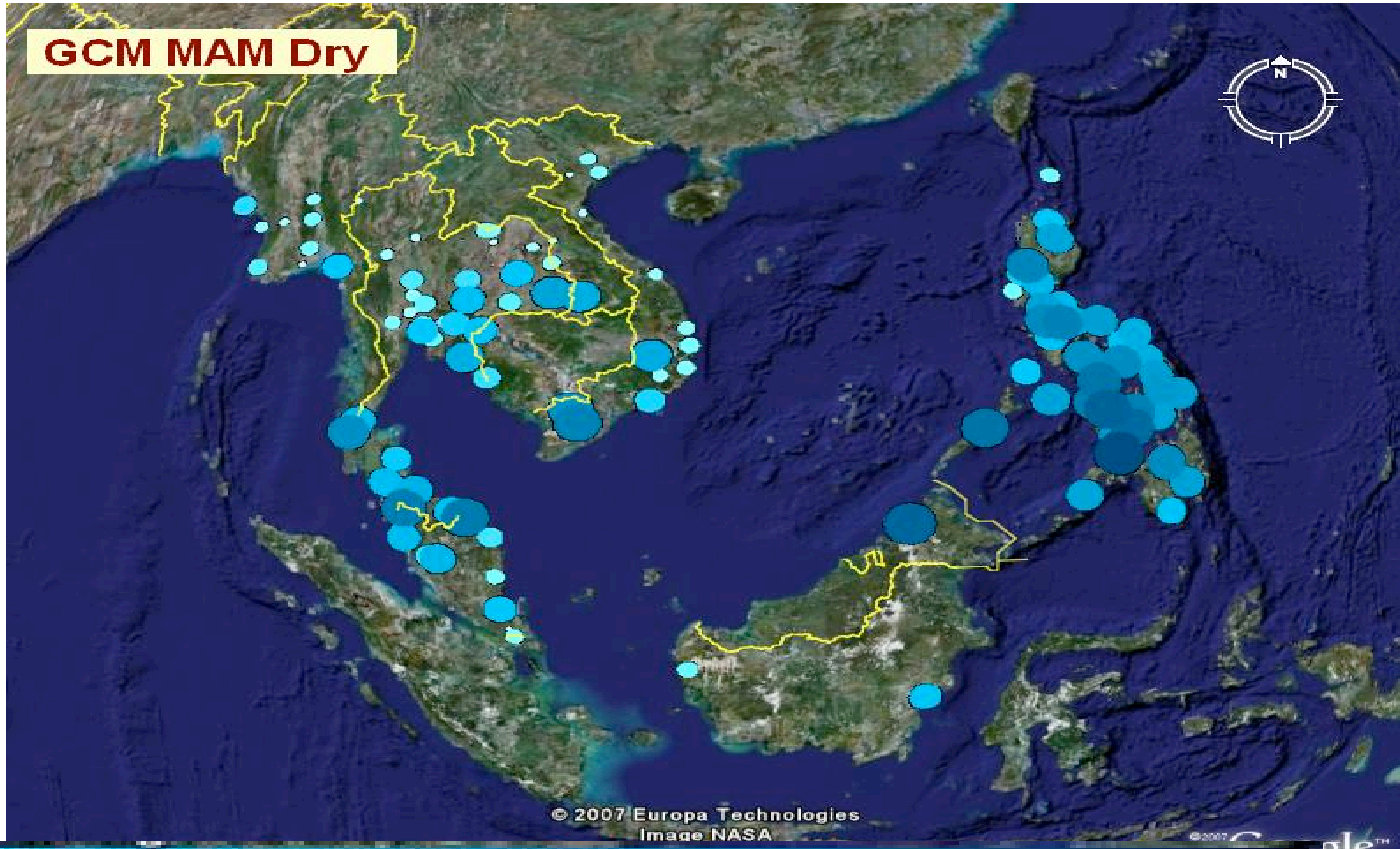
ASMC/IRI Seasonal-Intraseasonal Climate Prediction and its Applications Workshop
21st May—30th May 2007, Singapore

GCM Downscaled Precip. Anomaly Correlation Skill (from 2007 Singapore Workshop ASEAN participants)

Season Rainfall Total

Number of Dry Days





Quantile regression

- The ultimate goal of regression analysis is to model the **conditional distribution** of the response variable given a set of explanatory variables - this is called **Distributional Regression**

- **Quantile regression** is a reduced form in which the predictand is a quantile of the forecast PDF. **Logistic regression** is well suited to predicting a probability rather than a measurable physical quantity

$$\ln\left[\frac{p}{1-p}\right] = f(\mathbf{x}) \qquad p = \Pr\{V \leq q\}$$

p is the probability of not exceeding quantile q
This equation is linear on the logistic, or log-odds scale

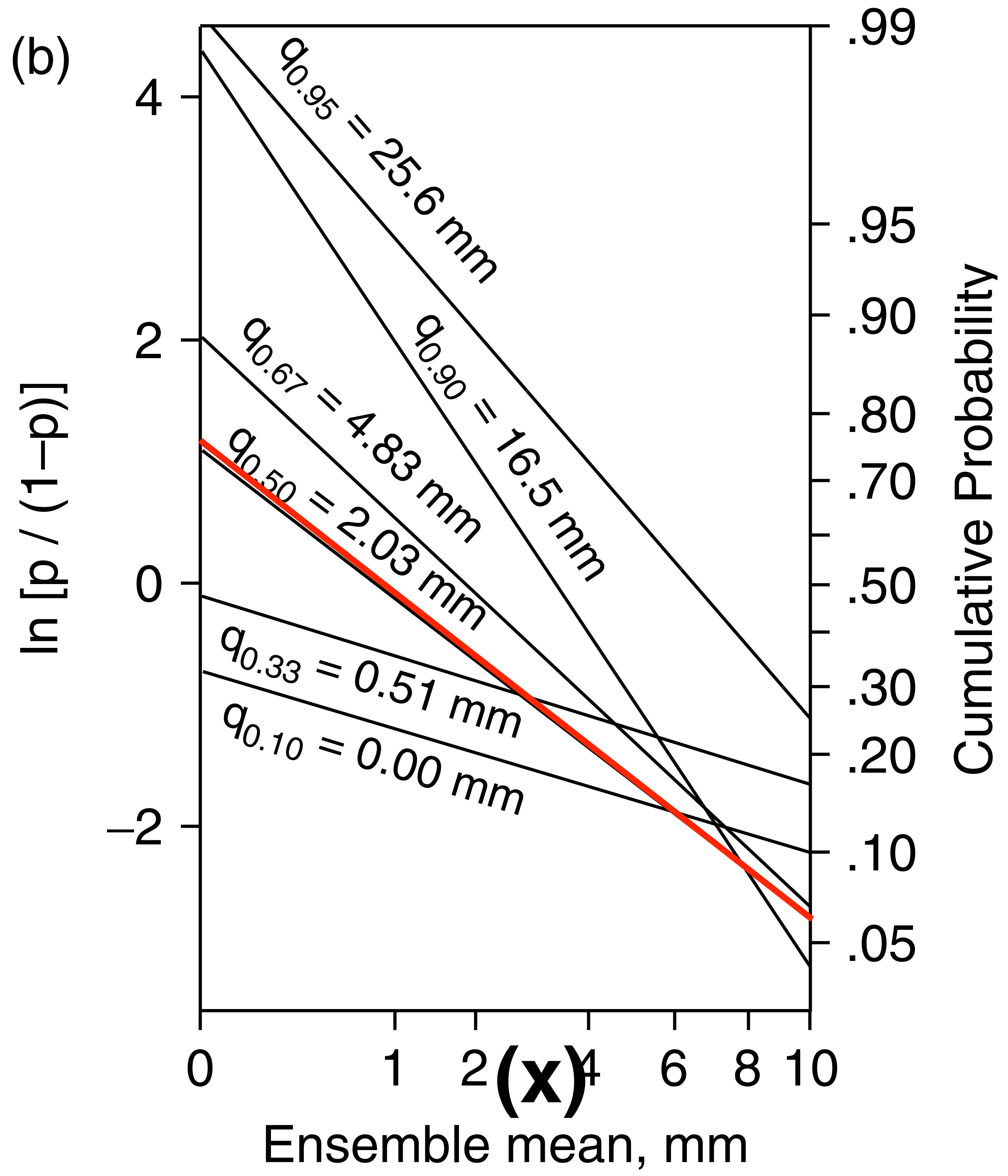
GFS Day 6–10 Accumulated Precip Forecast for Minneapolis

28 Nov – 2 Dec 2001

x is GFS ensemble mean precip at nearest grid point, square rooted

p is probability of not exceeding various quantiles (cumulative probability)

training-data window of ±45 days around the forecast date



Individual regressions

$$\ln\left[\frac{p}{1-p}\right] = f(\mathbf{x})$$

Wilks (2009)

Extended logistic regression

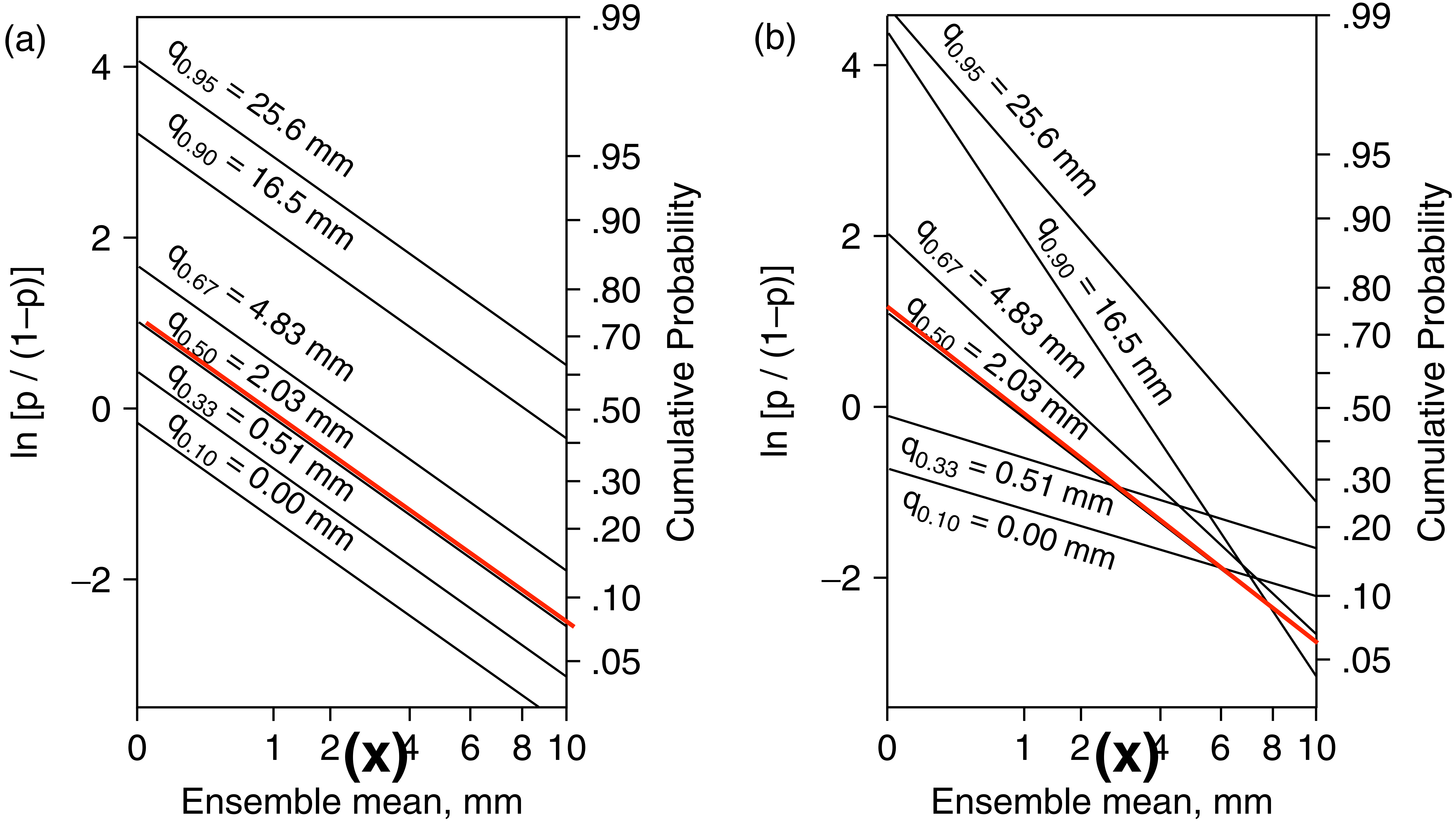
- Logistic regression lines obtained separately for each quantile can cross
- Extended logistic regression alleviates this:

$$\ln\left[\frac{p(q)}{1-p(q)}\right] = f(\mathbf{x}) + g(q)$$

this specifies parallel functions of the predictors x, whose intercepts b0 $*$ (q) increase monotonically with the threshold quantile, q

GFS Day 6–10
Precip Forecast for
Minneapolis
28 Nov – 2 Dec 2001

$$\ln\left[\frac{p(q)}{1-p(q)}\right] = f(\mathbf{x}) + g(q) \qquad \ln\left[\frac{p}{1-p}\right] = f(\mathbf{x})$$



Wilks (2009)

An S2S example

CFSv2 re-forecasts calibrated with extended logistic regression (Wilks 2009)

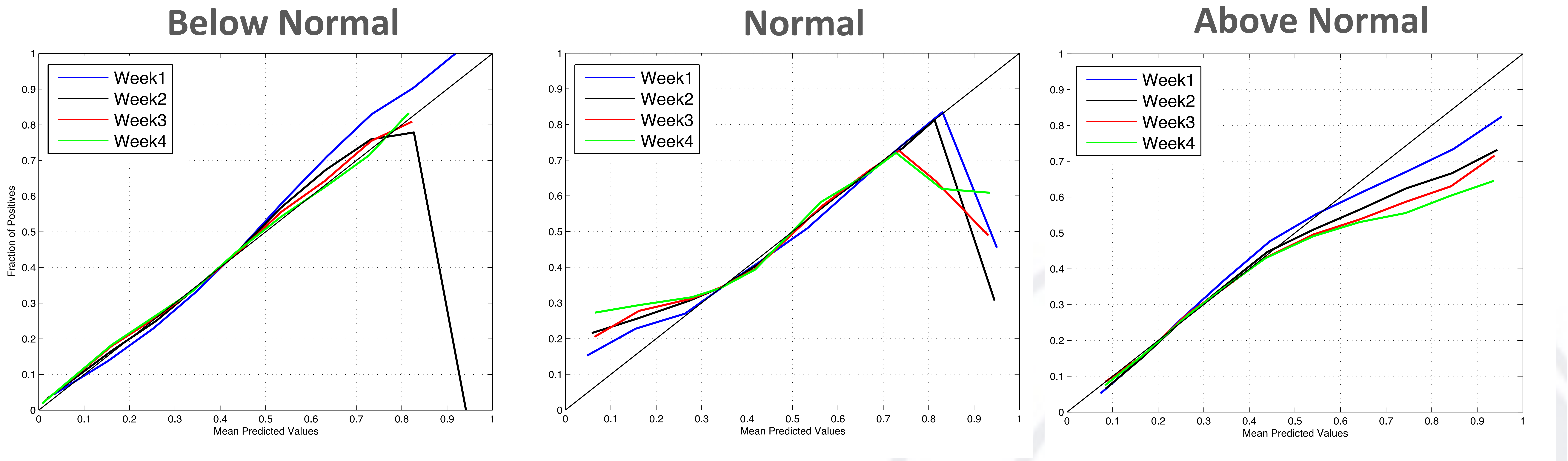


Figure 5: Reliability diagram for forecasts issued in JAS over the 1999-2010 period from one to four weeks lead over a US only window (land and ocean points)

courtesy of Nicolas Vigaud, IRI

done separately for each gridpoint
4-member ensemble averages, every day

Main points

- Seasonal forecasts are sometimes **tailored**, expressing the forecast in terms of a **predictand of interest** (e.g. rainfall frequency, monsoon onset date, drought probability, river flow, crop yield..).
- This can also be a form of **forecast calibration** or **statistical downscaling**, according to the choice of predictand.
- **Regression models** are the workhorse of forecast tailoring and calibration, with predictor (explanatory) variables taken from GCM ensemble-mean forecasts or antecedent climate conditions.
- Usually a **Gaussian** or transformed Gaussian distribution is assumed.
- Most regression approaches are limited to the **conditional mean** as a function of the predictor variables. The spread needs to be estimated separately.
- **Quantile regression** using **extended logistic regression** has been used in weather forecasting and seems well suited to calibrating sub-seasonal forecasts.