Multivariate Analysis of Regression Patterns

Timothy DelSole and Xiaosong Yang

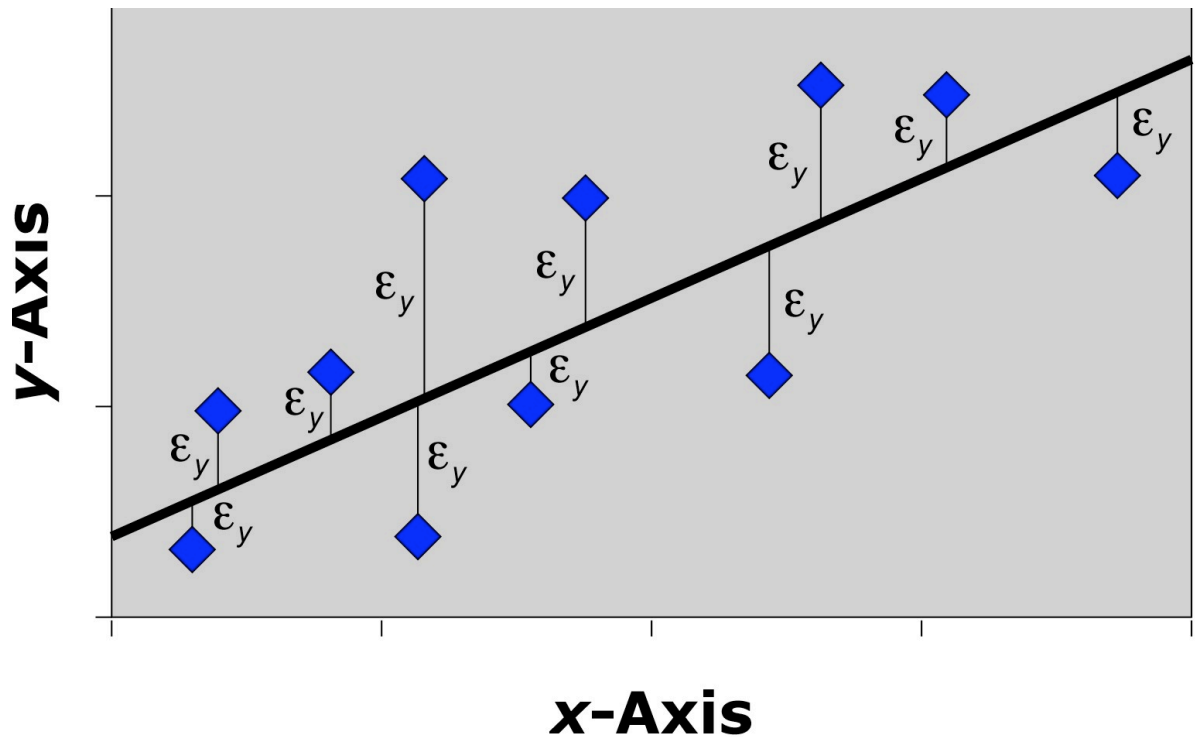
George Mason University, Fairfax, Va and
Center for Ocean-Land-Atmosphere Studies, Calverton, MD

July 25, 2010

What is a Regression Pattern?

A regression pattern is the set of regression coefficients between a pre-specified time series and each variable in the data set.

Fit $y = ax + b + \epsilon$. The "regression coefficient" is the slope.



[1]

[1] figure http://www.palass.org/modules.php?name=palaeo_math&page=7

Example of Regression Pattern: Trend Maps

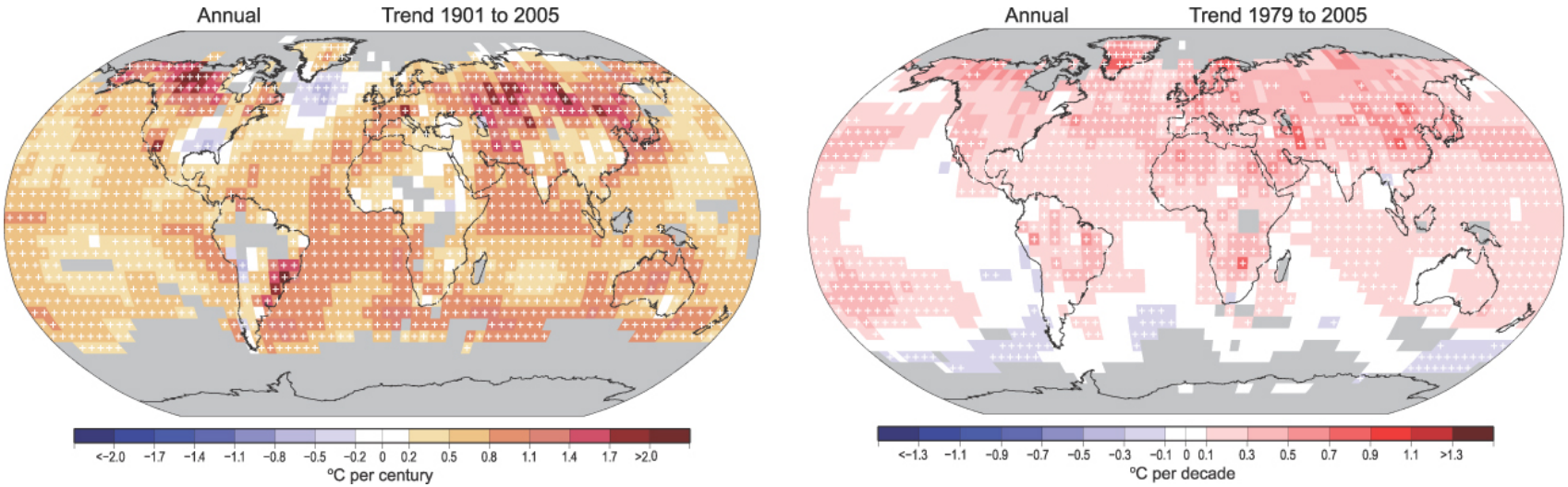


Figure: Linear trend of annual temperatures. Trends significant at 5% level indicated by white + marks. Grey areas have insufficient data to estimate reliable trends. IPCC-AR4, fig. 3.9

Example of Regression Pattern: ENSO and SLP

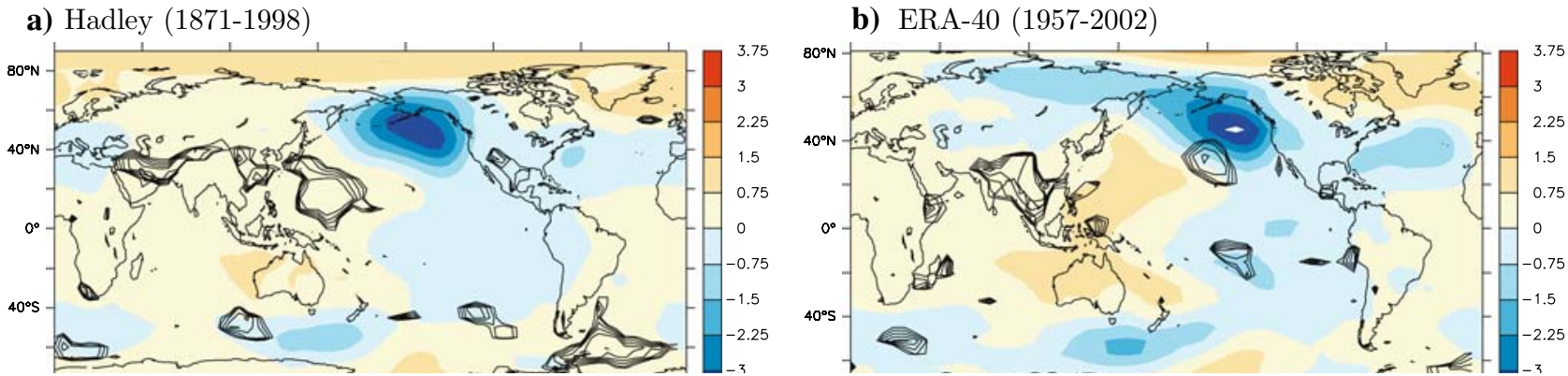
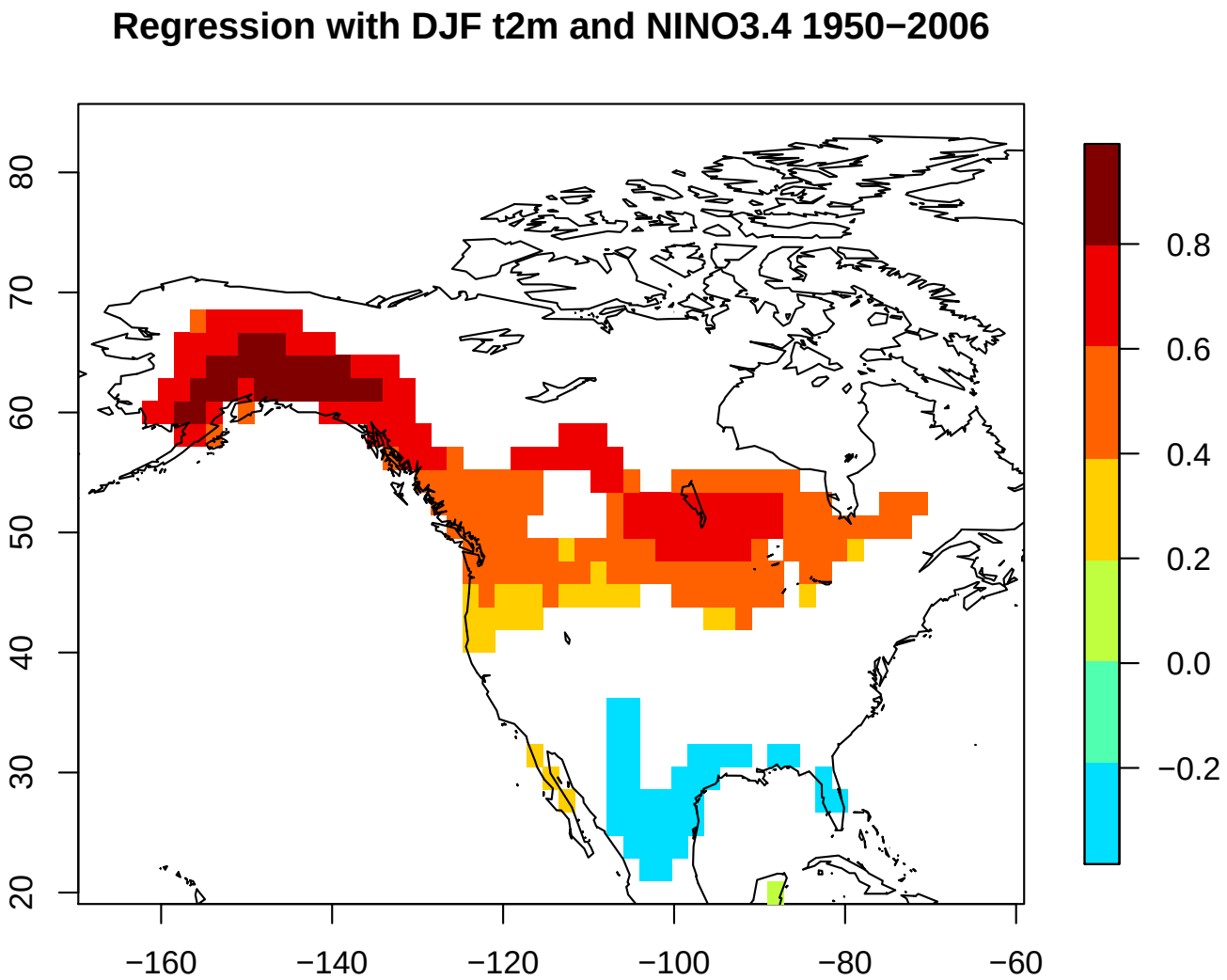


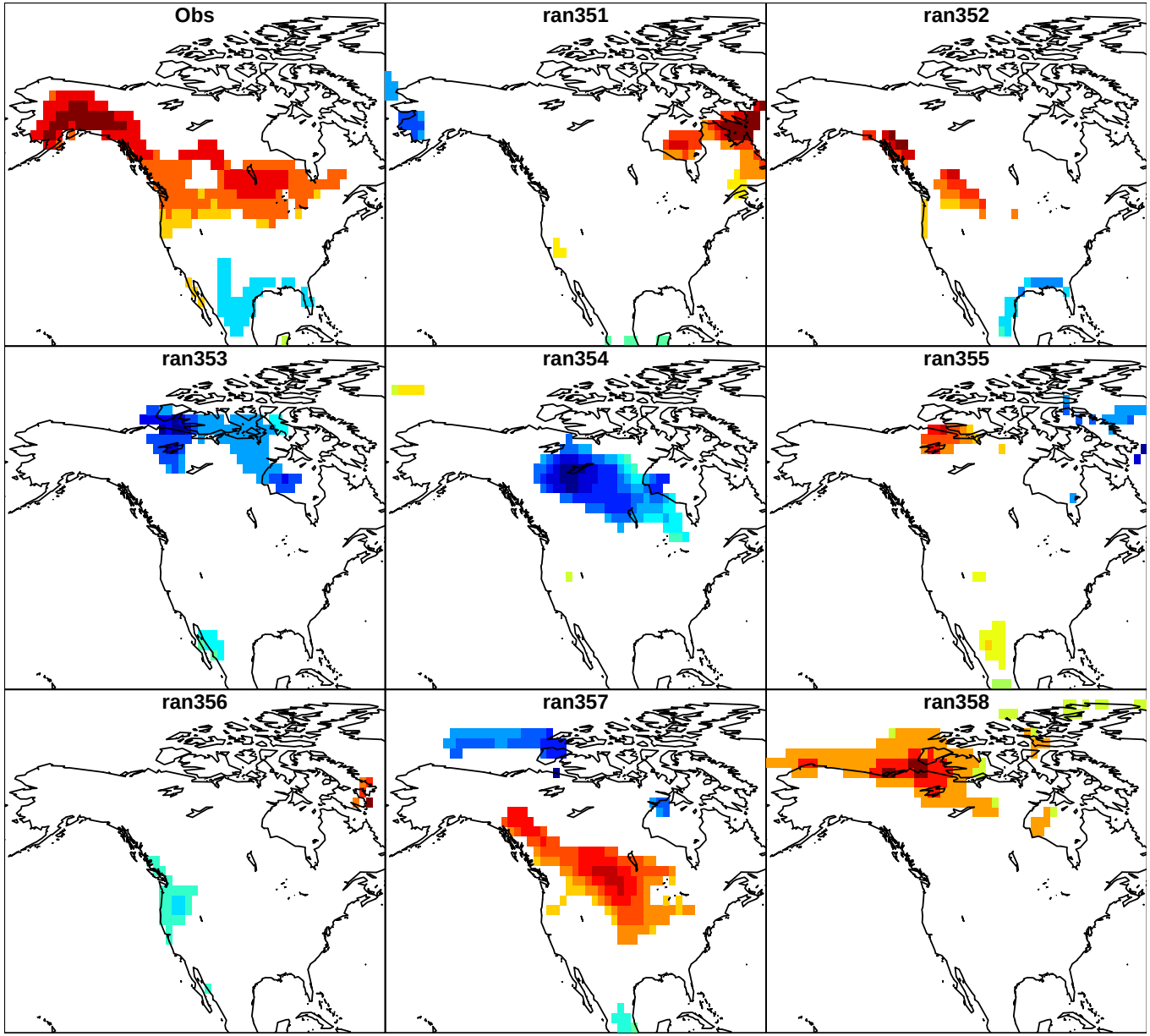
Figure: Regression coefficients between NINO3.4 index and Sea Level Pressure in January for (a) observations (HadSLP1), (b) ERA-40. Sterl et al., 2007, *Clim. Dyn.*

Can We Trust Regression Patterns at Face Value?



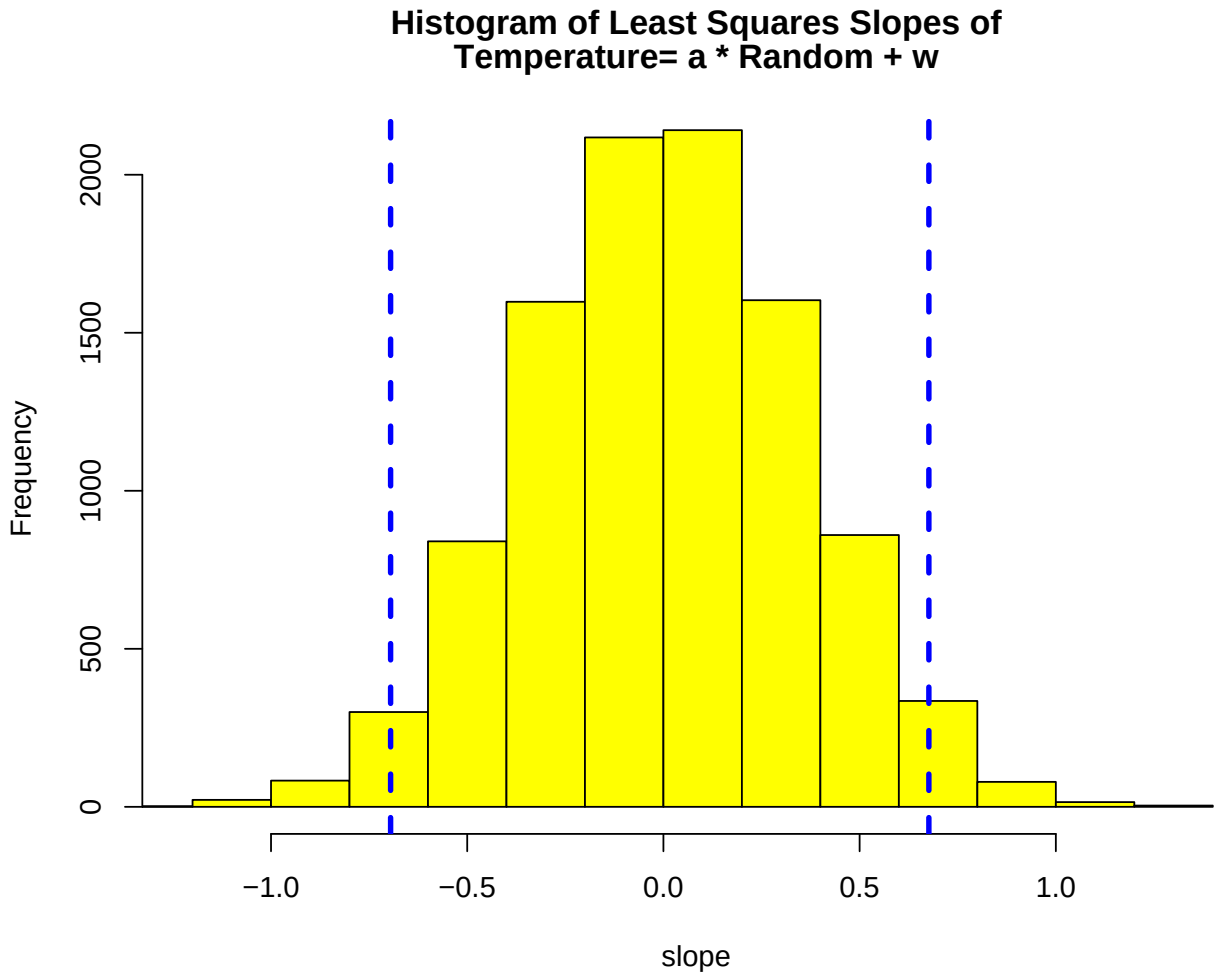
Coefficients that are insignificant at the 5% level are masked out.

Regression Between T2m and Random Noise



How Can Regression With Random Noise be Significant?

Answer: because we are testing it a 1000 times



How Can Regression With Random Noise be Significant?

- Even if true coefficient is zero, sample coefficient can be large.
- The distribution of sample coefficient is known exactly.
- The probability that sample coefficient exceeds a critical value when true coefficient vanishes is the significance level.
- Histogram shows that 5% significance level is 0.68, so we expect sample correlation to exceed 0.68 5% of the time.

Gridded Data Tends to be Spatially Correlated Too

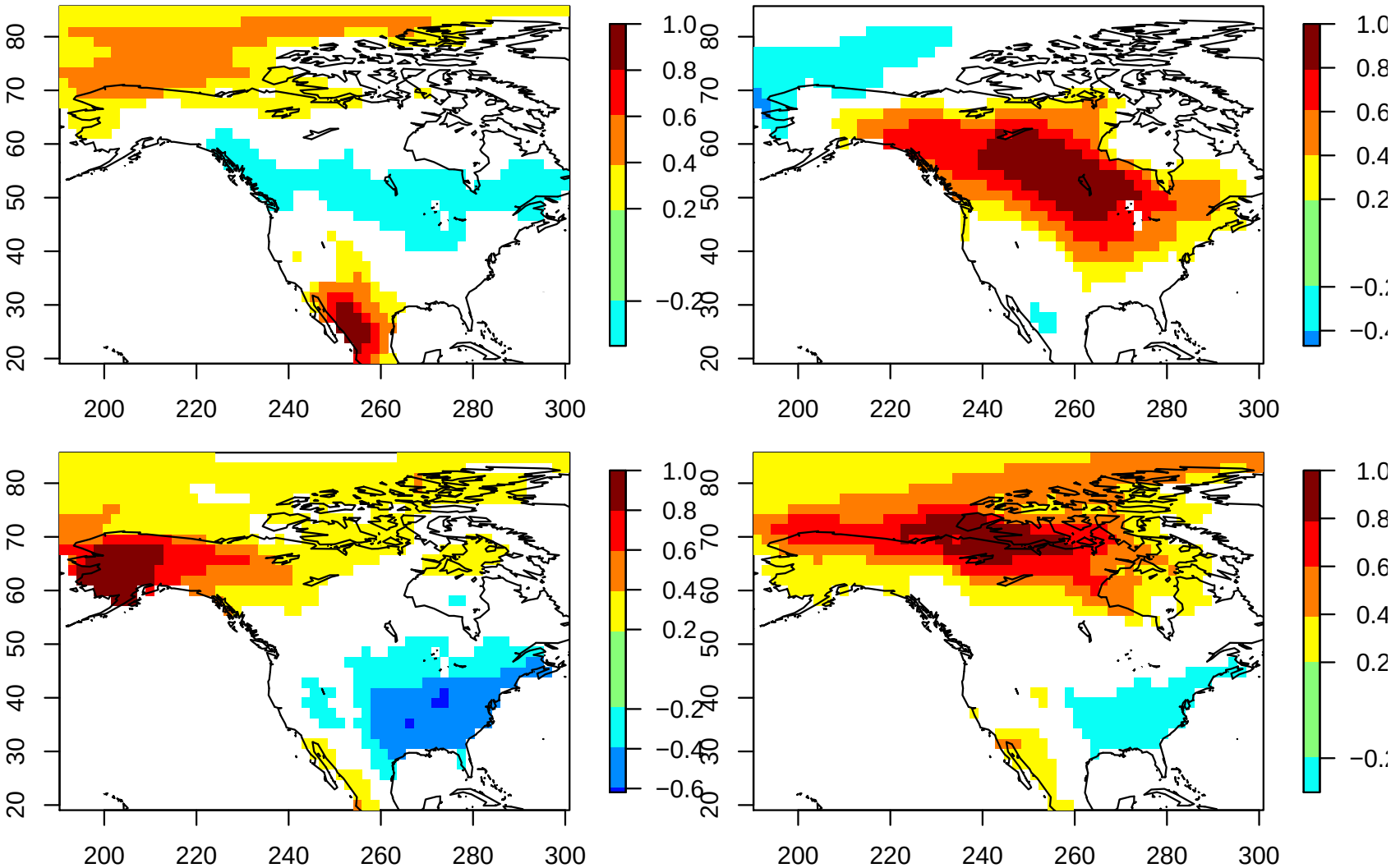


Figure: Correlation between base point and neighboring grid points for T2m reanalysis 1950-2009.

Effect of Spatial Correlations

Correlations in space imply that if one grid point has a high correlation with a time series, then the neighboring points also will have a strong correlation, *no matter the true correlation*.

This means that the significance tests are not independent.

We do not expect just 5% of the grid points to exceed the significance threshold randomly, but rather 5% of the spatially coherent structures to exceed the threshold.

This means that more than 5% of the area tends to exceed the significance threshold for spatially correlated data.

Field Significance

Field significance is the statistical significance of the hypothesis that all regression coefficients vanish simultaneously.

To perform a field significance test, need to account for:

- multiple hypothesis are being tested simultaneously.
- variables are interdependent.

Gilbert Walker

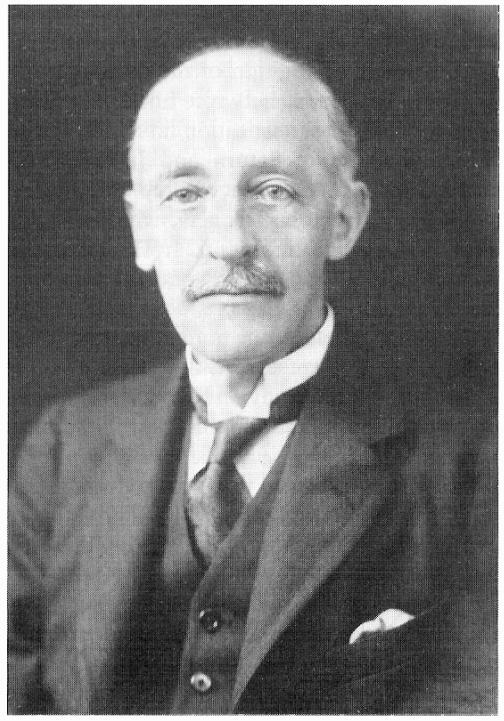


Figure: Sir Gilbert Walker

"[Let c be the probability that the correlation between independent quantities is less than p.] Then the chance of all coefficients [between m pairs of independent quantities] being less than p will be c^m." -Walker 1914

Experimentwise Error Rate

The 5% significance level is the absolute correlation below which sample correlations will fall 95% of the time, for independent data.

However, if the sample correlation is calculated for M different indices, then the probability that at least one correlation out of M exceeds the α-significance level is

$$prob = 1 - (1 - \alpha)^M$$

M	1	2	3	4	5	10	20
prob	5%	10%	14%	19%	23%	40%	64%

Table: Probability that event occurs at least once in M trials when probability of the event occurring in one trial is 5%

The probability of at least one false rejection of the null hypothesis over multiple comparisons is called the experimentwise error rate.

Multiple Comparisons

The comparisonwise $\alpha_c = 5\%$ significance level should NOT be used for multiple comparisons.

For multiple comparisons, one should use the experimentwise significance level:

$$\alpha_e = 1 - (1 - \alpha_c)^{1/M}$$

Livezey and Chen 1983

Even if 5% significance tests are independent, the number of "passed tests" that occur by accident is often more than 5%.

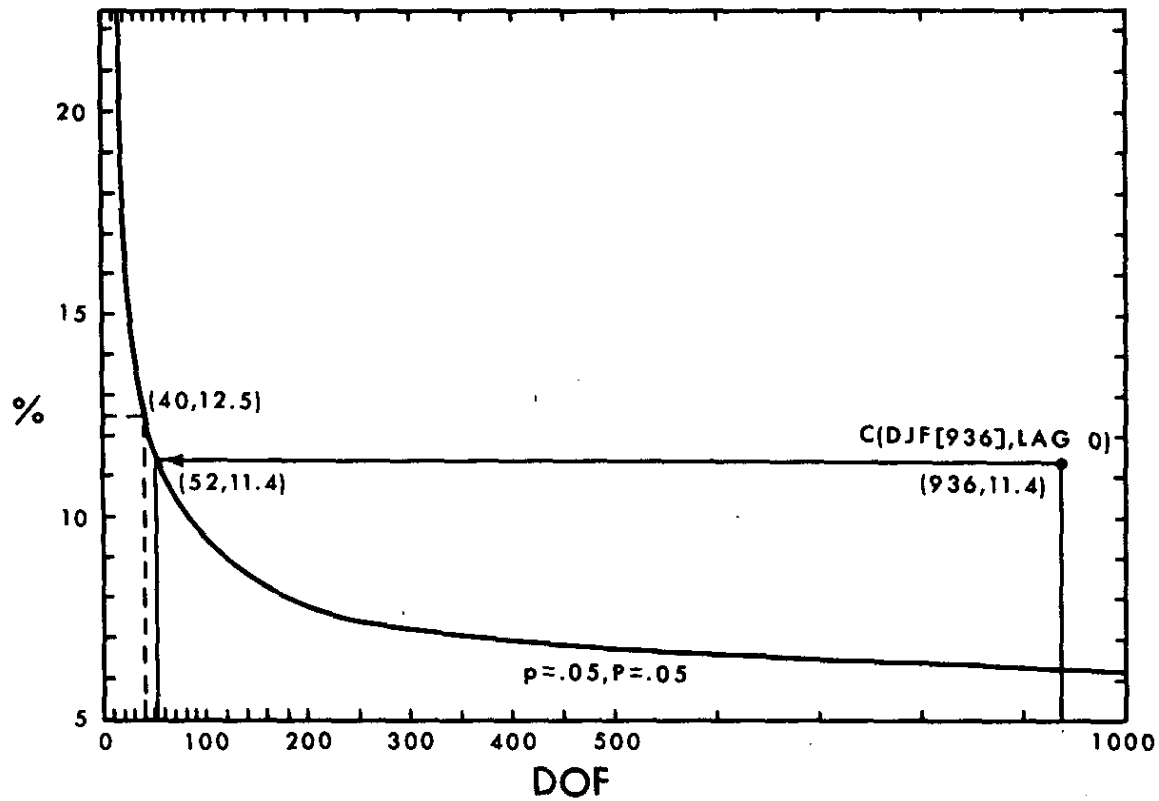


FIG. 3. Estimated percent of independent 95% ($p = 0.05$) significance tests passed that will be equalled or exceeded by accident 5% ($P = 0.05$) of the time versus the number of independent tests N (labeled "DOF" for "degrees of freedom"). The curve is based on the binomial distribution. The plotted point and coordinate lines and points refer to the significance test of Chen's experiment described in the text.

Effect of Spatial Correlations

Large cross-correlations in the field reduce the "degrees of freedom" in the field. -Livezey and Chen 1983

The "effective" number of degrees of freedom is not known, but can be estimated by a variety of techniques.

Monte Carlo Estimate of Field Significance

Livezey and Chen (1983) procedure:

- Replace pre-specified time series with random numbers drawn from same distribution as pre-specified time series.
- Calculate correlation maps between field and random numbers.
- Count the number of "passed tests" in the field.
- Repeat many times and record the counts.
- Compare observed count with counts from random numbers.
- If observed count falls in the upper 5th percentile, reject hypothesis that all correlation vanish.

Monte Carlo Estimate of Field Significance (Livezey and Chen 1983)

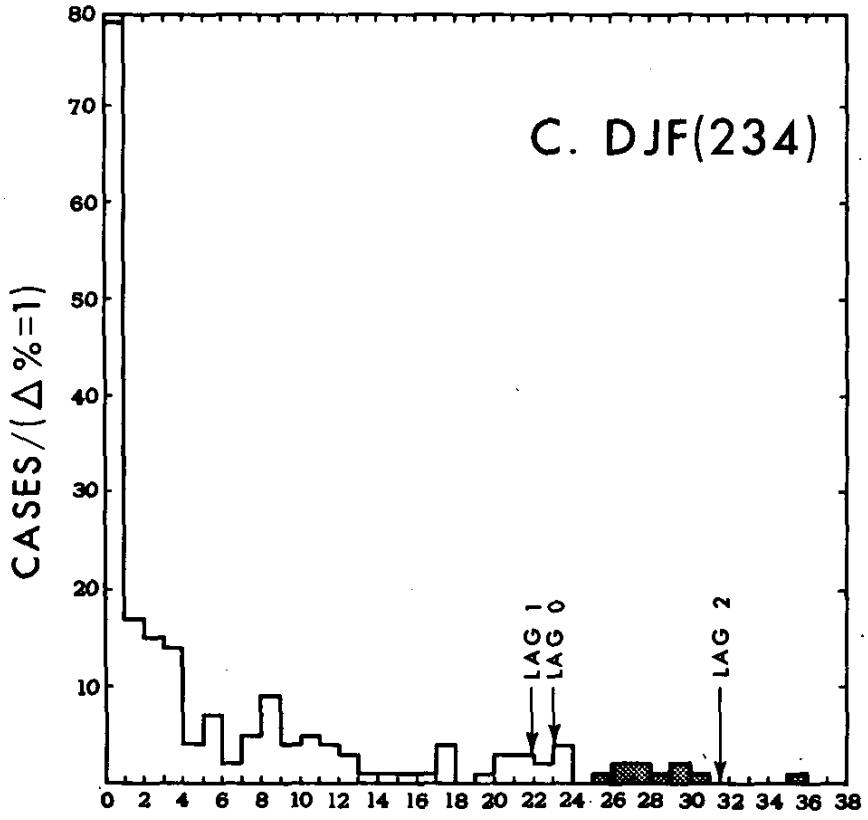


FIG. 5. Histograms of percent of area with correlations of 700 mb heights and Gaussian noise statistically significant at the 95% level ($p = 0.05$) in 200 Monte Carlo trials for: (A) the winter hemisphere; (B) the summer hemisphere; and (C) the winter north Pacific basin (outlined in heavy dashed line in Fig. 1). The abscissa is percent of area while the ordinate is number of cases for each one percent interval. The 5% tail ($P = 0.05$) is schematically shown by shading the 10 of 200 largest percents. The results for correlations with seasonally averaged SOI's, with heights lagging by the indicated number of seasons, are shown by vertical arrows.

Limitations of "Counting" Methods

Counting methods count the number of passed tests regardless of spatial location or degree of significance.

False Discovery Rate (Wilks 2006)

The False Discovery Rate is the expected proportion of rejected local null hypotheses that are actually true.

Wilks (2006) proposed testing field significance based on FDR:

- Perform M independent hypothesis tests $H_1, H_2, \ldots, H_M$.
- Calculate the corresponding p-values $p_1, p_2, \ldots, p_M$.
- Order p-values: $p_{(1)}, p_{(2)}, \ldots, p_{(M)}$
- For given α, find the largest k such that $p_{(k)} \leq \alpha k/M$.
- Reject the corresponding hypotheses $H_{(i)}$ for $i = 1, 2, \ldots, k$.

Wilks claims that Walker's test and the FDR approach are "relatively" insensitive to correlations among local tests.

Multivariate Regression Approach

Fit the model

$$\underset{\text{data}}{Y_{nm}} = \underset{\substack{\text{regression}\\ \text{coefficient}}}{z_m} \underset{\substack{\text{pre-specified}\\ \text{time series}}}{t_n} + \underset{\text{constant}}{\mu_m} + \underset{\text{noise}}{\epsilon_{nm}}$$

Least squares estimate of $\mathbf{z}$ is the regression pattern

$$\hat{\mathbf{z}} = \mathbf{Y}^T\mathbf{t}/(\mathbf{t}^T\mathbf{t}),$$

assuming the time series $\mathbf{t}$ has zero mean.

Hypothesis Test in Multivariate Regression

Test hypothesis that regression coefficients vanish simultaneously:

$$z_1 = z_2 = \cdots = z_M = 0.$$

The likelihood ratio test is a standard procedure for testing hypothesis in multivariate regression. This test leads to the statistic

$$\lambda = \left(\frac{\mathbf{t}^T \mathbf{t}}{N} \right) \hat{\mathbf{z}}^T \hat{\mathbf{\Sigma}}_N^{-1} \hat{\mathbf{z}}$$

where $\hat{\mathbf{\Sigma}}_N$ is the "noise" covariance matrix

$$\hat{\mathbf{\Sigma}}_N = \frac{1}{N} \left(\mathbf{Y} - \mathbf{1}\hat{\mu}^T - \mathbf{t}\hat{\mathbf{z}}^T \right)^T \left(\mathbf{Y} - \mathbf{1}\hat{\mu}^T - \mathbf{t}\hat{\mathbf{z}}^T \right).$$

If the null hypothesis $\mathbf{z} = \mathbf{0}$ is true, then

$$\lambda \frac{N - M - 1}{M} \sim F_{M, N-M-1}.$$

Problem With Multivariate Hypothesis Test

In typical climates studies, noise covariance matrix $\hat{\mathbf{\Sigma}}_N$ is singular because the number of coefficients exceeds the sample size.

This means the regression problem is underdetermined.

Discriminant Analysis Approach

Find the linear combination of variables that maximizes the fraction of variance explained by the pre-specified time series.

Equivalently, find the linear combination of variables that maximizes the correlation with the pre-specified time series.

Discriminant Analysis Approach

Let weights be q_m. Then linear combination gives the time series

$$r_n = \sum_m Y_{nm} q_m.$$

Fit the time series to the pre-specified time series:

$$r_n = \alpha t_n + \epsilon_n.$$

The fraction of variance explained by the pre-specified time series is

$$\text{“Signal-to-noise ratio”} = STR = \frac{\text{var}[\hat{\alpha} t_n]}{\text{var}[r_n]}$$

We seek weights $\mathbf{q}$ that maximizes the signal-to-total ratio STR.

Solution to Discriminant Analysis

The weights that maximize STR can be found analytically as

$$\mathbf{q} = \hat{\mathbf{\Sigma}}_N^{-1}\hat{\mathbf{z}}.$$

The signal-to-noise ratio (SNR) turns out to be

$$SNR = \lambda = \left(\frac{\mathbf{t}^T\mathbf{t}}{N}\right)\hat{\mathbf{z}}^T\hat{\mathbf{\Sigma}}_N^{-1}\hat{\mathbf{z}}.$$

Discriminant analysis and multivariate analysis lead to the same λ.

Discriminant analysis shows that λ is the optimal signal-to-noise ratio of a linear combination of variables.

Significance test for SNR is exactly the same as the significance test of $\mathbf{z} = 0$ in multivariate regression.

Limitation of Discriminant Analysis

Exactly the same as multivariate regression: $\hat{\mathbf{\Sigma}}_N$ is singular.

Practical Approach

- Transform data into principal components.
- Select a small number of PCs to represent the data.
- Solve regression and discriminant problems in PC-space.
- Transform solutions back into data space.

Example: Trend Analysis

- Annual mean sea surface temperature (SST) 1948-2009.
- Data from ERSSTv3b, Smith and Reynolds (2004).
- $2^\circ \times 2^\circ$ grid.
- $\mathbf{t}$ is a linear function of year, incrementing by 1/10.

Trend Pattern for SST

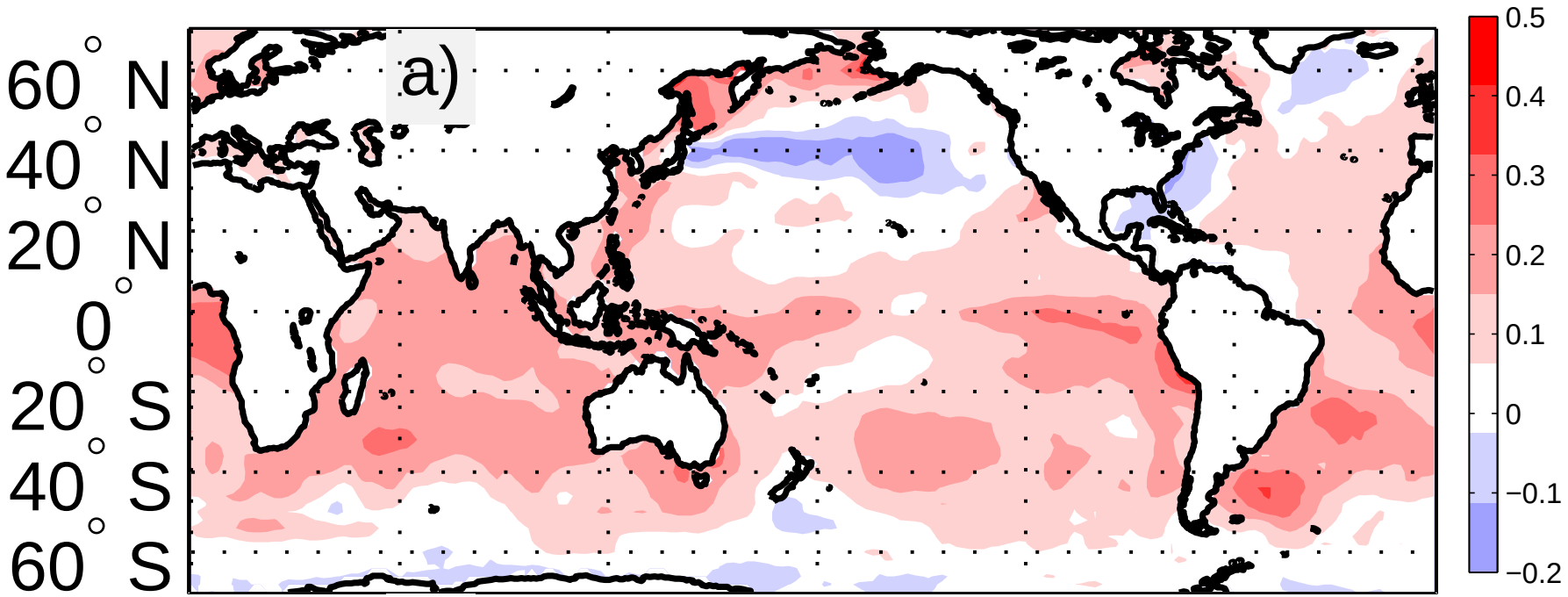
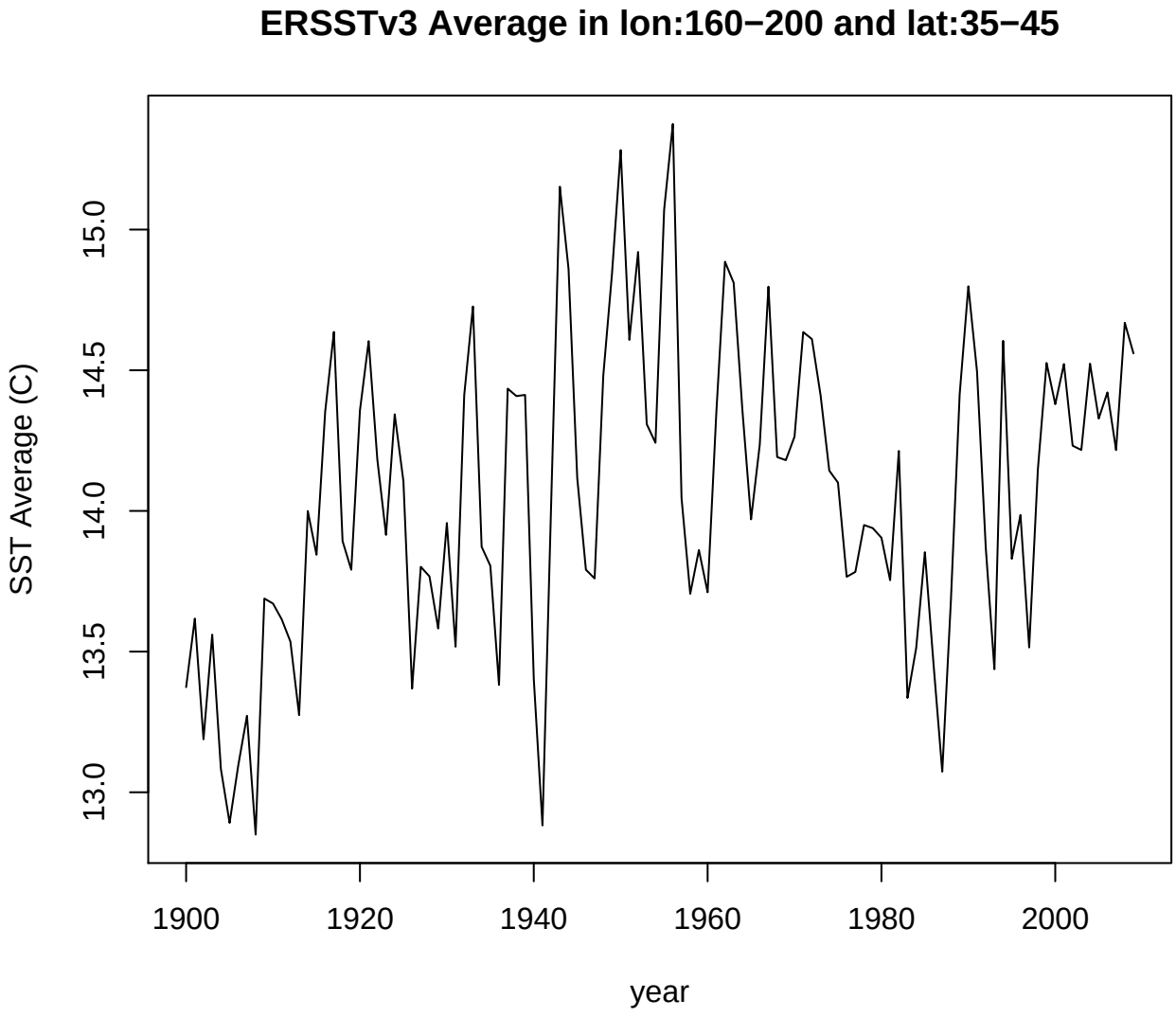


Figure: Point-by-point regression coefficients for the trend of annual mean SST during 1948-2009 (in degrees K per decade)

Cooling in the North Pacific?



Discriminant Analysis

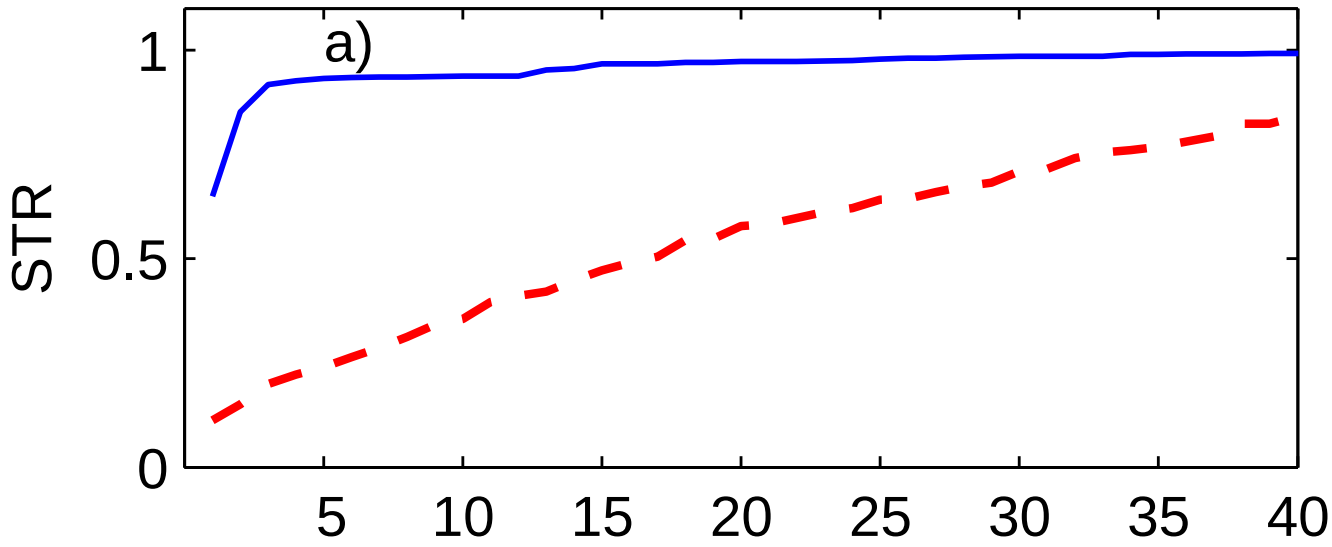


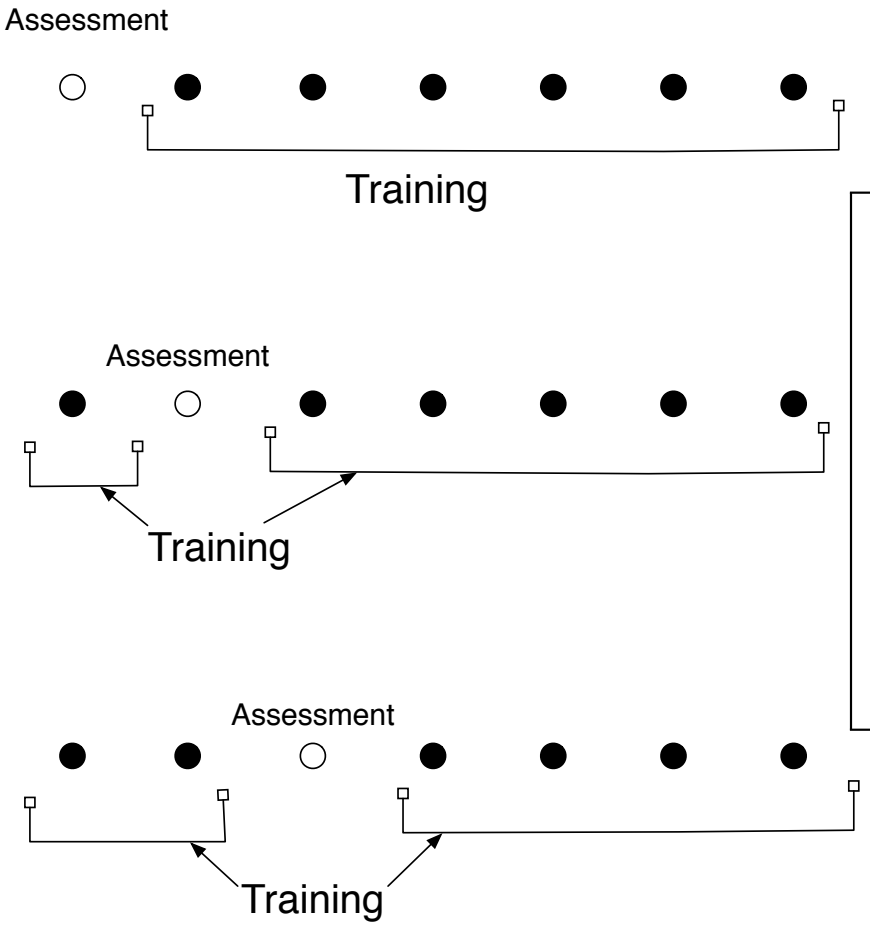
Figure: Optimal signal-to-noise ratio (solid blue) and 5% significance level (red dashed) of a linear trend for annual mean SST during 1948-2009, as a function of the number of PCs used to represent the data.

Comments About Discriminant Analysis Results

- The STR is statistically significant at all PC truncations.
- Most of the STR arises from the first three PCs.
- Little gain in STR using more than three PCs.

How Many PCs Should Be Chosen?

Leave-One-Out Cross Validation



Training Sample: Sample used to estimate model parameters.

Assessment Sample: Sample used to assess/test model predictions.

Cross Validation of Discriminant Analysis

For fixed number of PCs and time series $\mathbf{t}$:

- Withhold year n.
- Calculate discriminant component from remaining years.
- Use resulting projection vector to predict amplitude of discriminant component in year n: r_n
- Fit discriminant time series to $\mathbf{t}$ using training sample, use resulting equation to predict time series in year n: $\hat{r}_n$.
- Calculate squared error $\epsilon_n^2 = (r_n - \hat{r}_n)^2$.
- Compute mean square error $E[\epsilon_n^2]$.

Plot cross-validated mean square error vs. number of PCs.

Extreme Example Using 50 PCs (For Illustration)

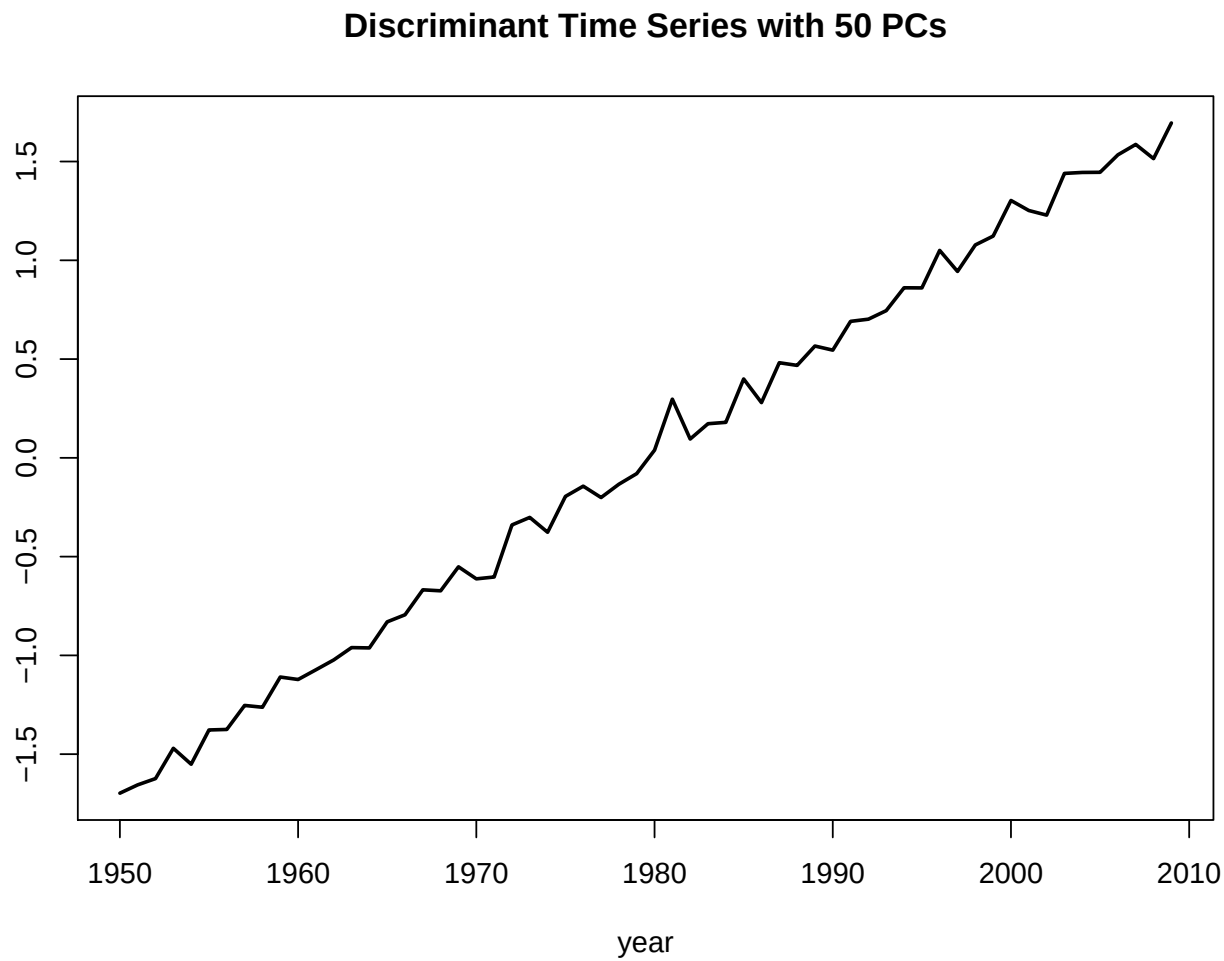


Figure: Discriminant time series for linear trend in annual mean SST using all data in 1948-2009.

Extreme Example Using 50 PCs (For Illustration)

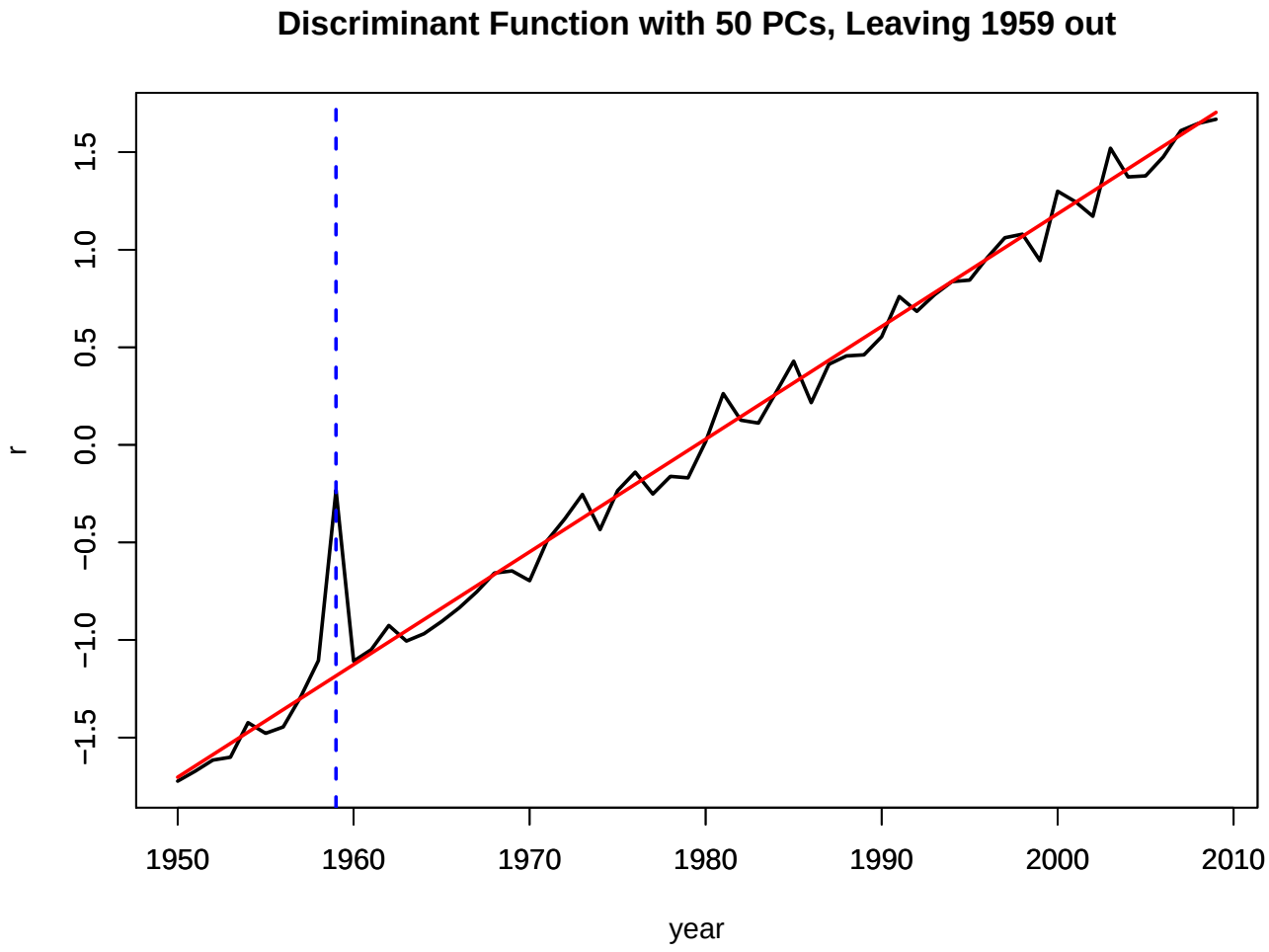


Figure: Discriminant time series for linear trend in annual mean SST using 50 PCs, leaving out 1959 (black). Trend fit based on data leaving 1959 out (red line).

Cross Validated Mean Square Error

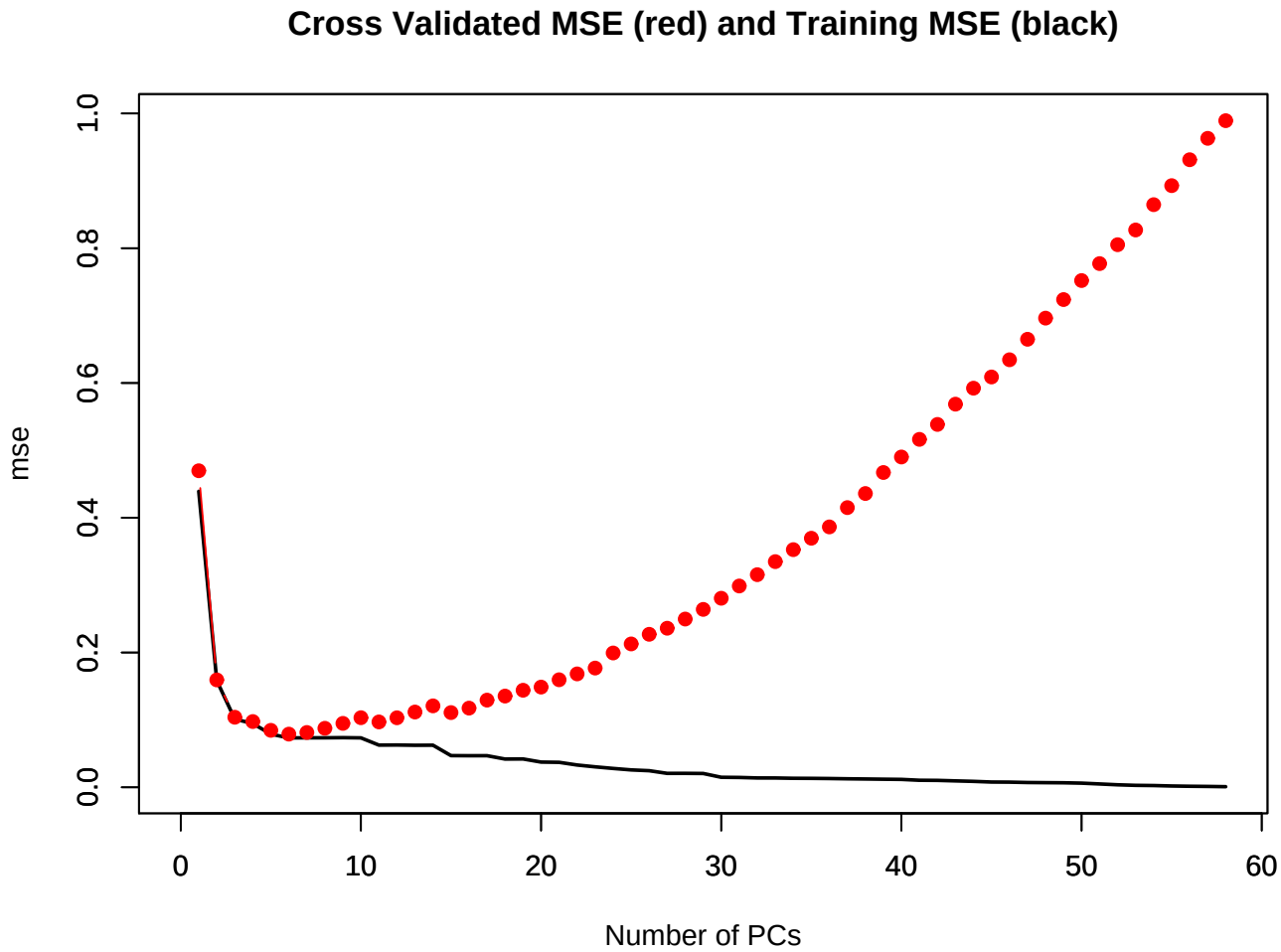


Figure: Cross validated mean square error (red) and in-sample mean square error (black) of the trend discriminant for annual mean SST.

Cross Validated Mean Square Error

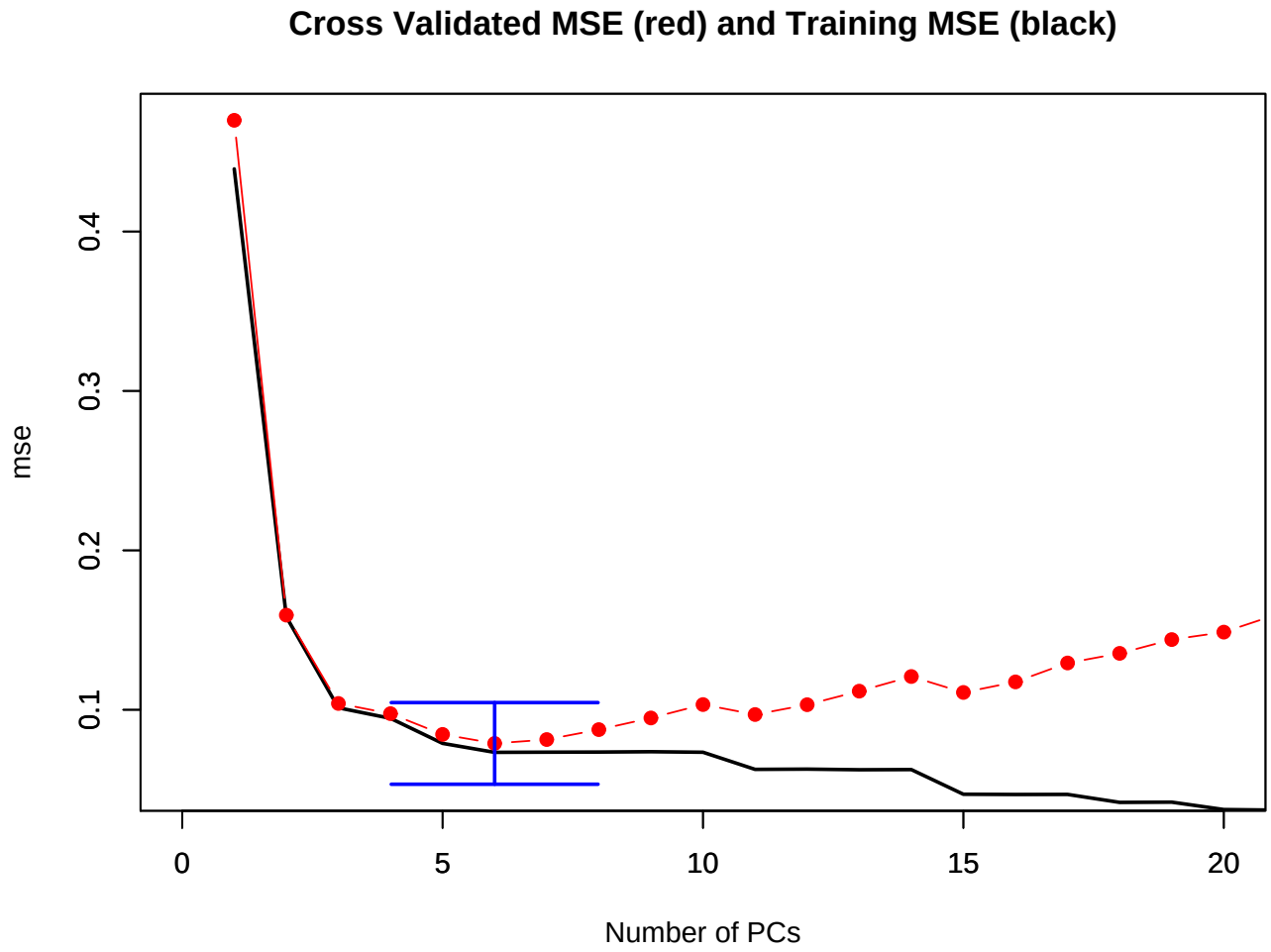


Figure: Cross validated mean square error (red) and in-sample mean square error (black) of the trend discriminant for annual mean SST.

Trend Discriminant for SST

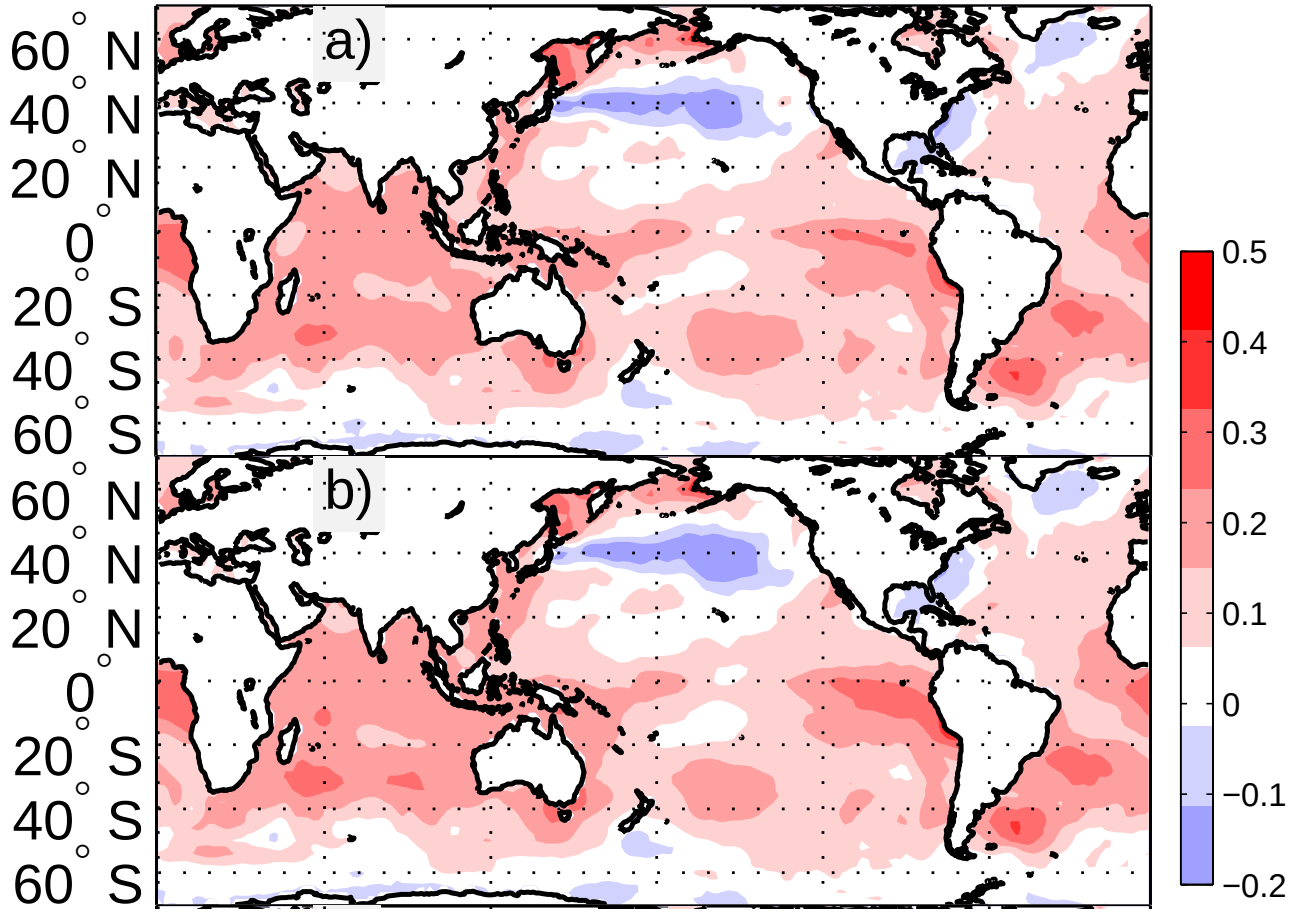


Figure: Point-by-point regression coefficients (a) and discriminant pattern based on 3 PCs (b) for the trend of annual mean SST during 1948-2009 (in degrees K per decade)

Comparison With EOF

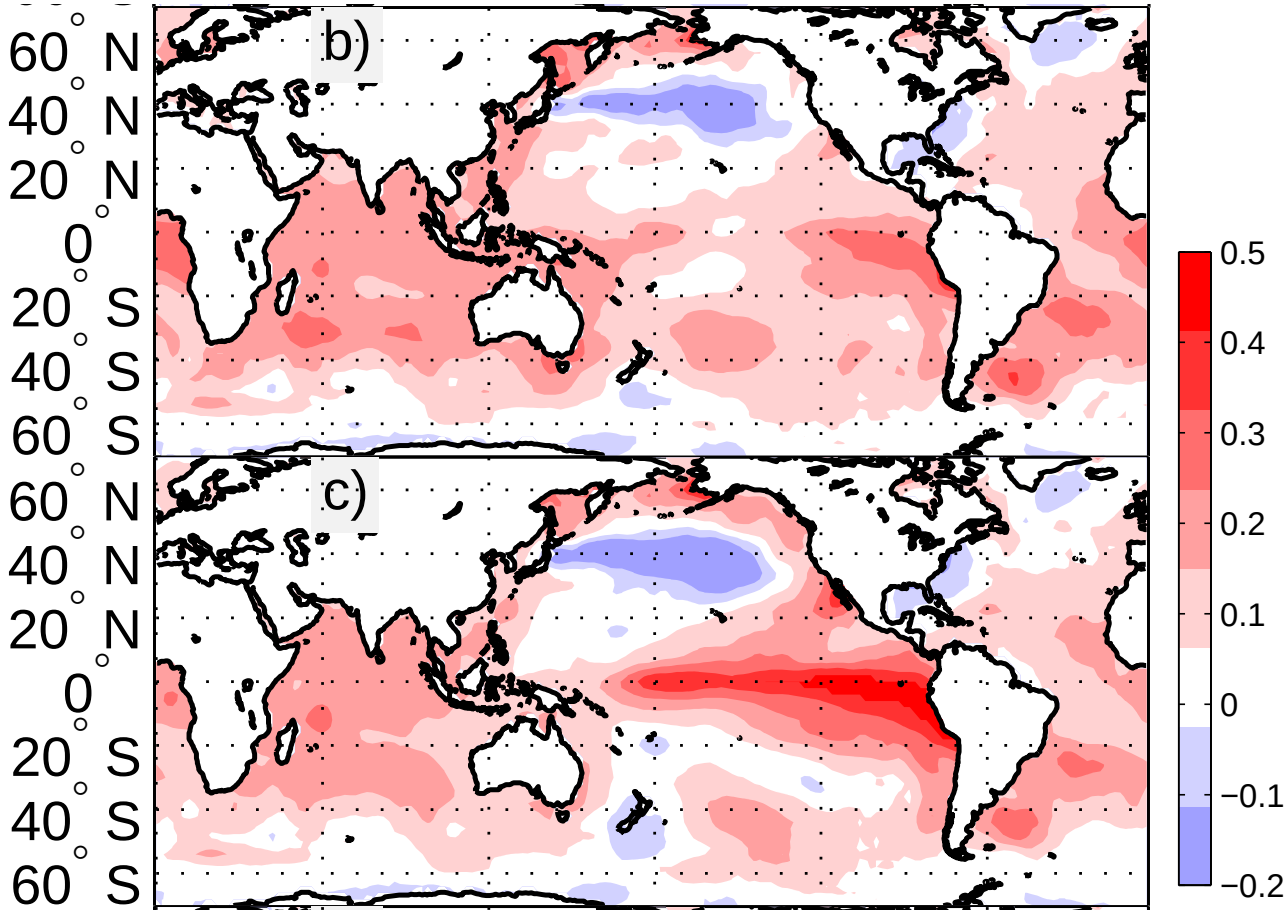


Figure: Discriminant pattern based on 3 PCs (b) and leading EOF (c) for the trend of annual mean SST during 1948-2009.

Discriminant Projection Pattern for SST

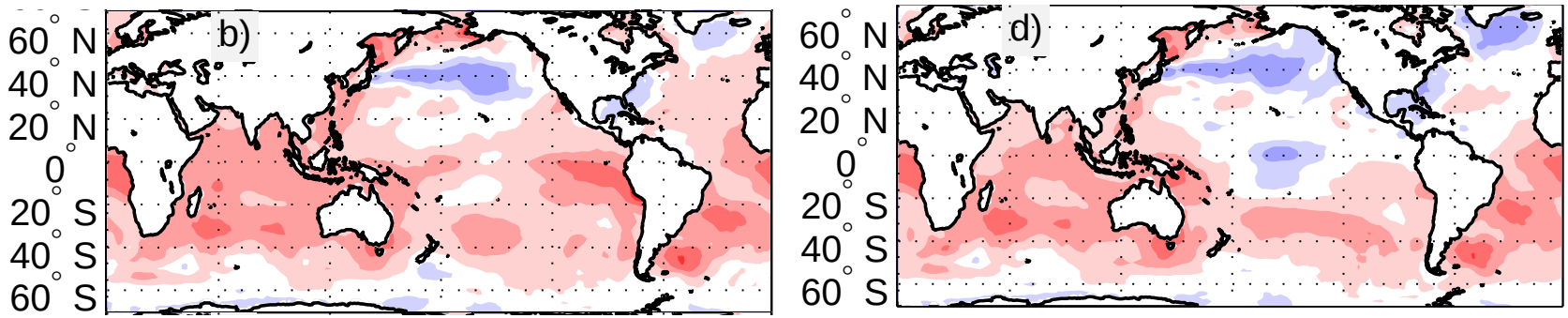


Figure: Discriminant pattern based on 3 PCs (b) and corresponding projection pattern (d) for the trend of annual mean SST 1948-2009.

Time Series for Discriminant and Principal Component

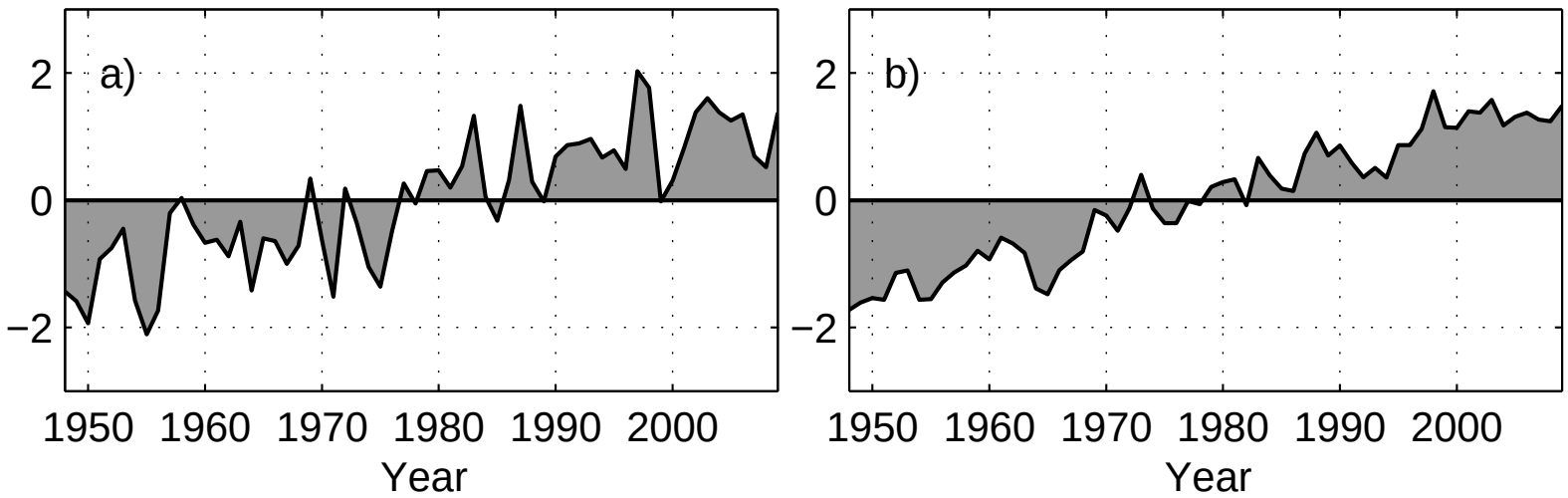
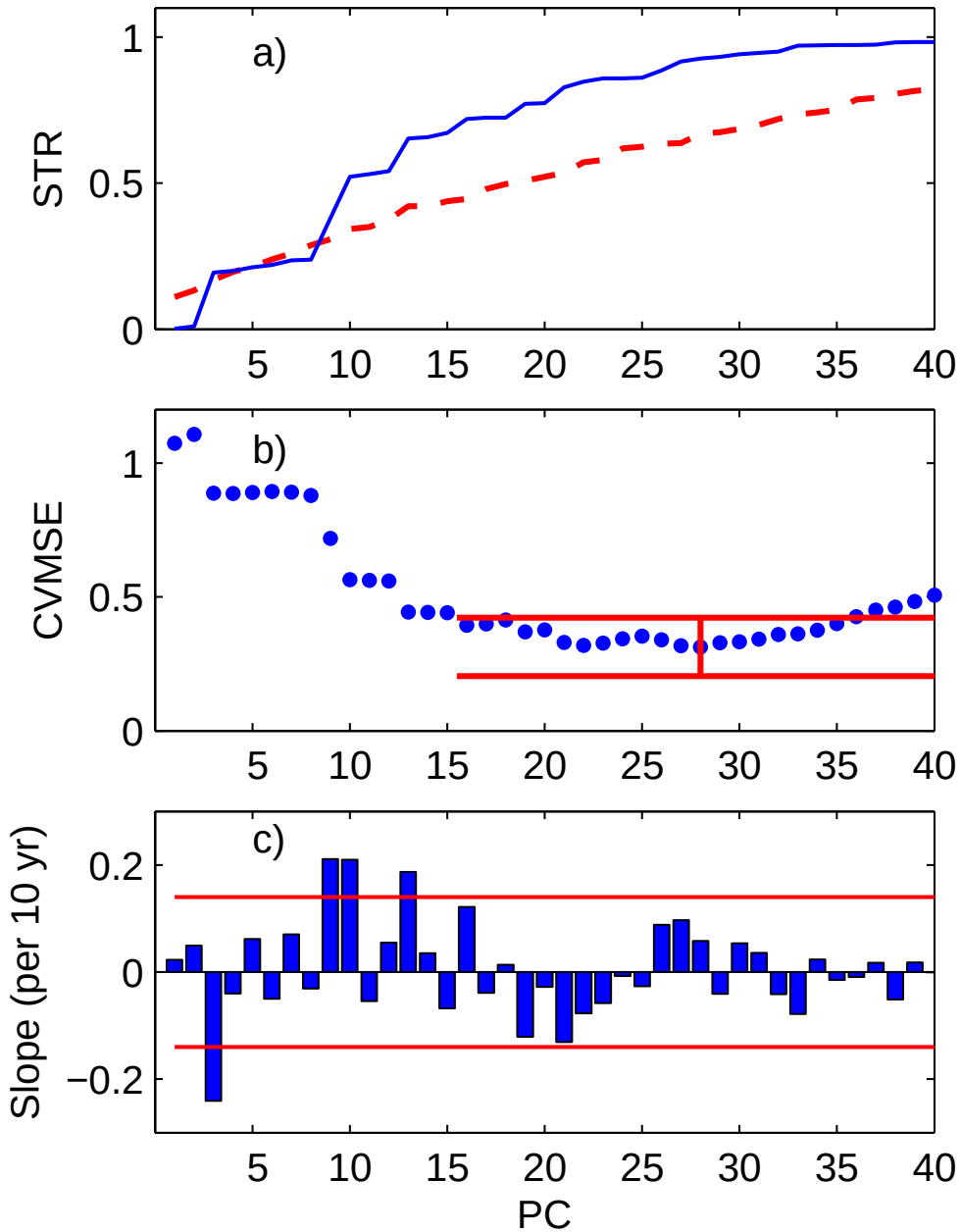


Figure: Time series for the leading principal component (a) and trend discriminant (b) for annual average SST during the period 1948-2009.

Results for DJF 300-hPa Zonal Wind



Results for DJF 300-hPa Zonal Wind

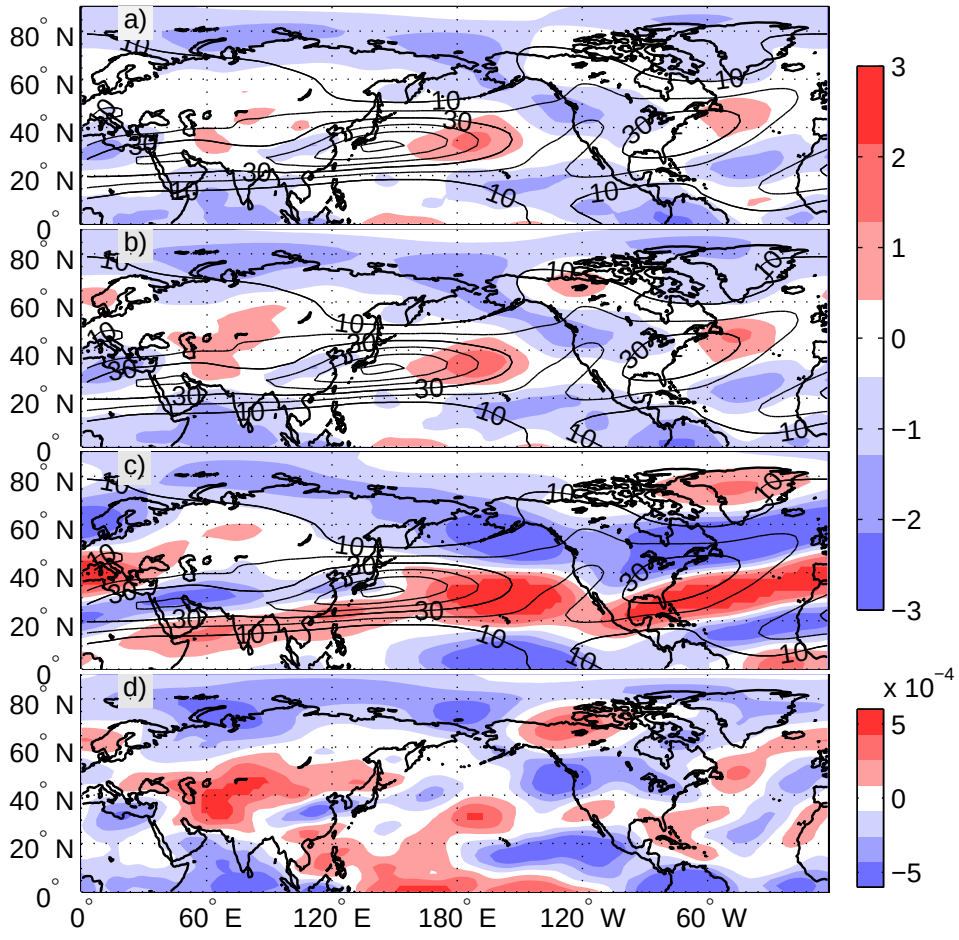


Figure: (a) The point-by-point trend, (b) trend discriminant pattern, (c) leading EOF, and (d) discriminant projection pattern for the trend discriminant, of DJF 300-hPa zonal wind during 1948-2009

Results for DJF 300-hPa Zonal Wind

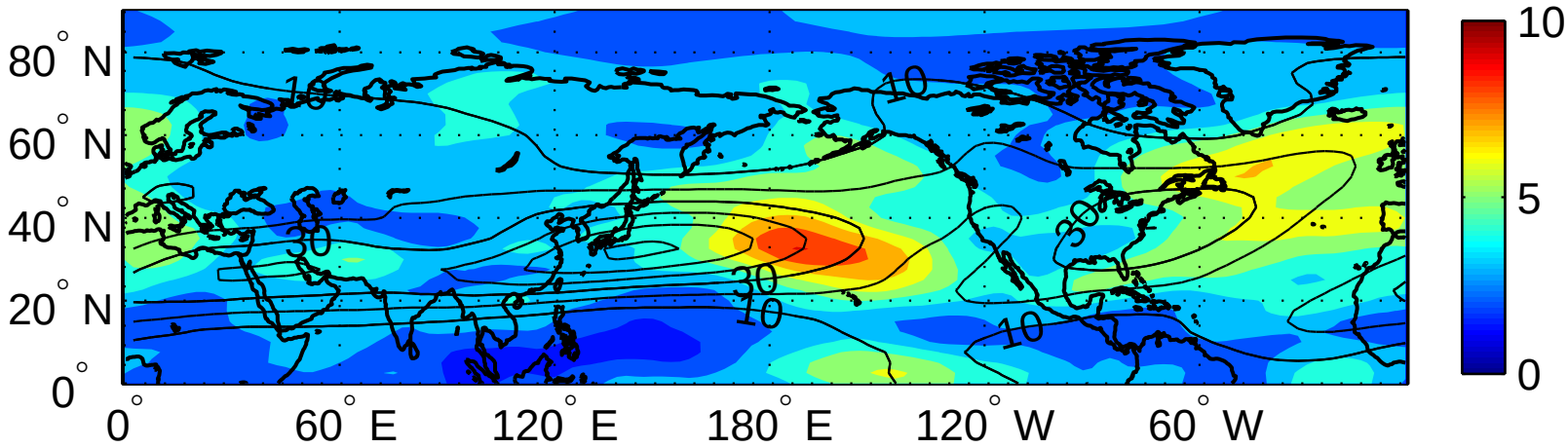


Figure: Standard deviation (shading) and mean (contours) of the December-January-February 300-hPa zonal wind during 1948-2009, in units of ms^{-1}.

Results for DJF 300-hPa Zonal Wind

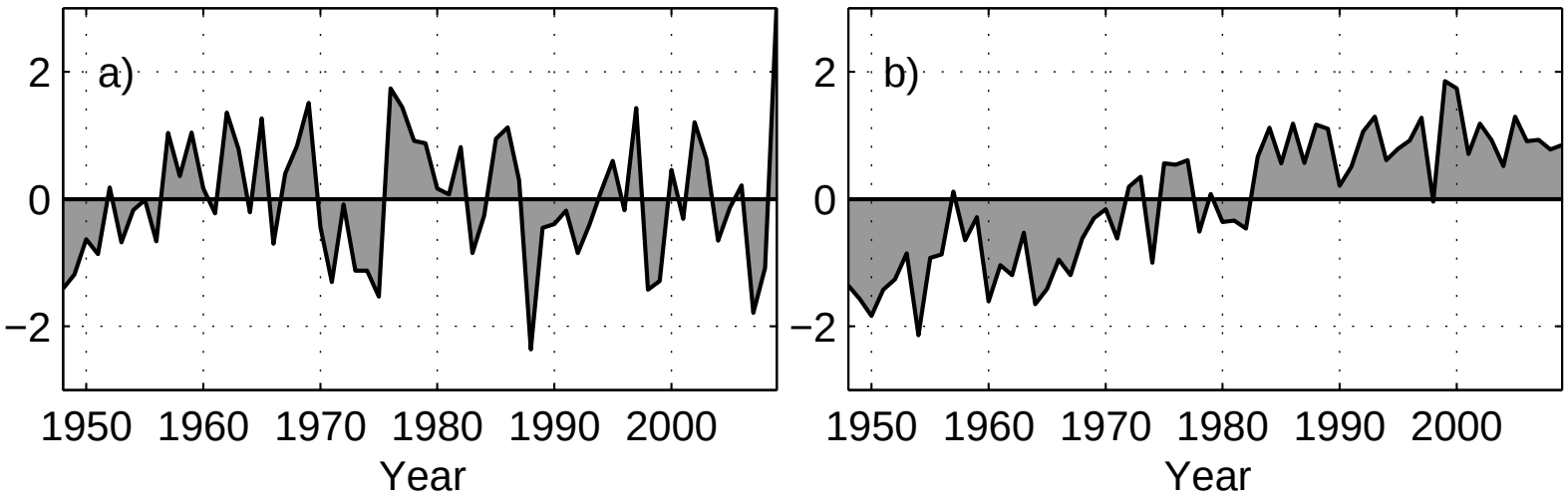
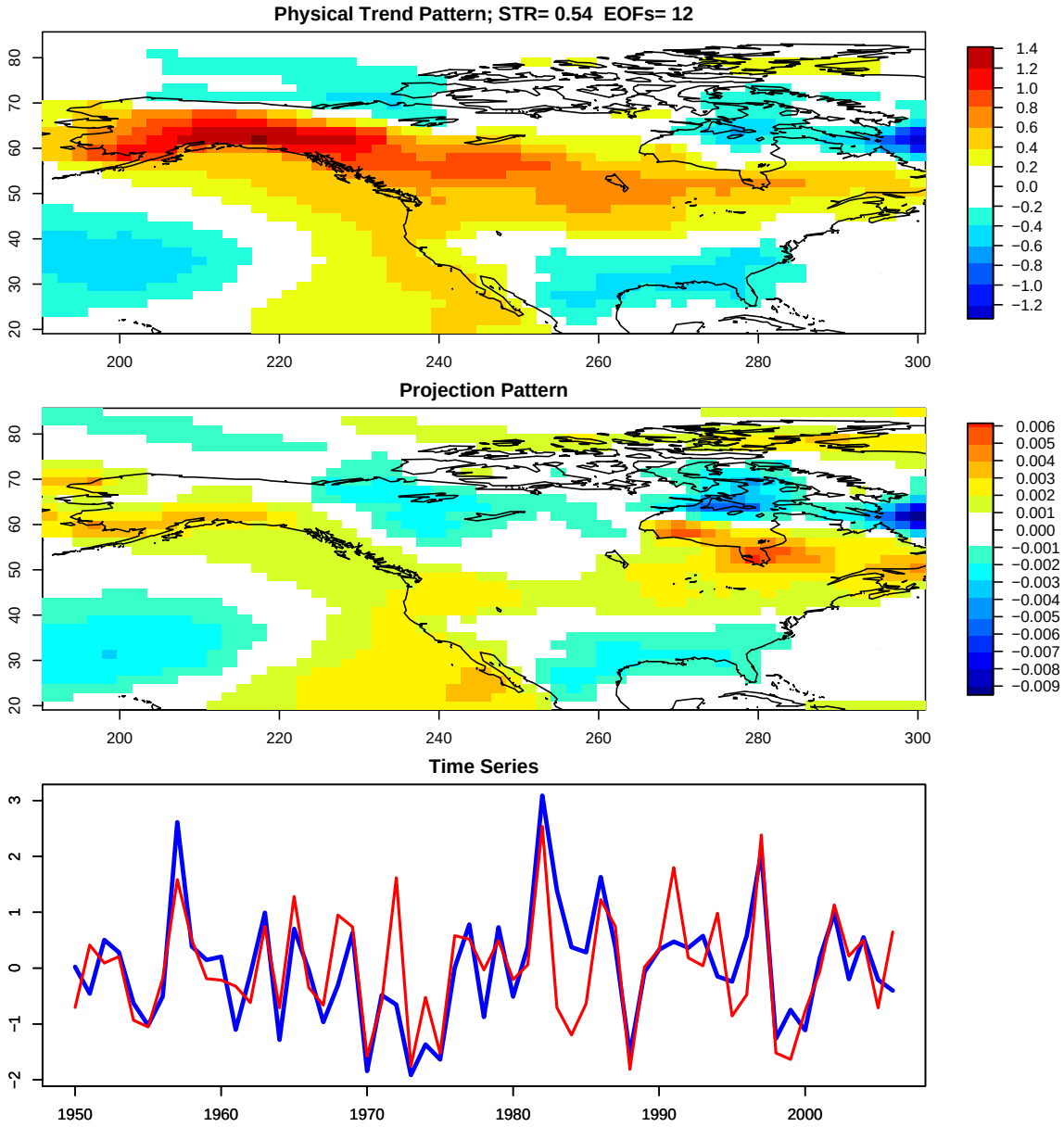
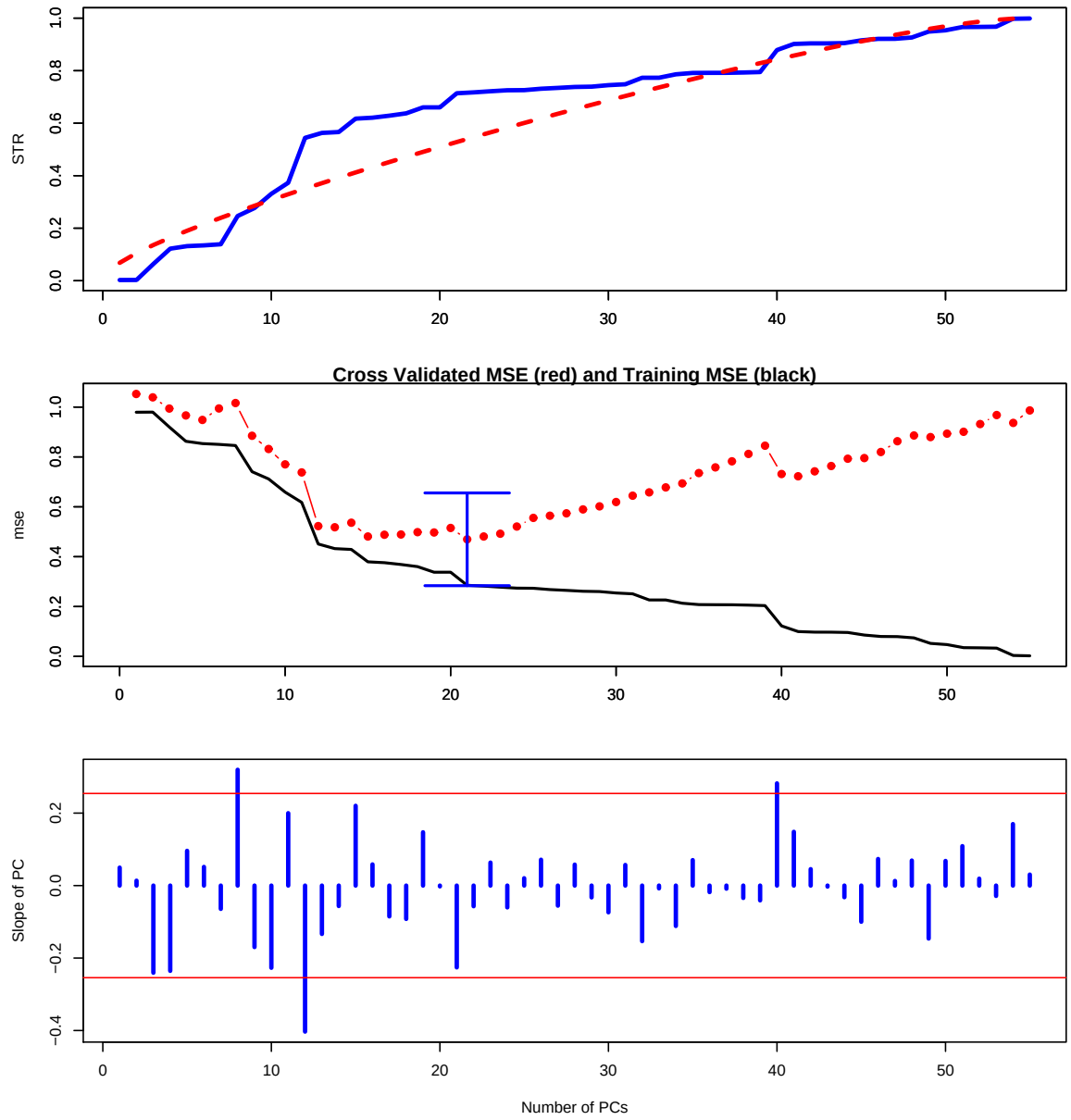


Figure: Time series for the leading principal component (a) and trend discriminant (b) for December-January-February 300-hPa zonal wind during the period 1948-2009.

Results for Regression between DJF T2M and NINO3.4



Results for Regression between DJF T2M and NINO3.4



Summary and Discussion

- Field significance of a regression pattern tests hypothesis that coefficients vanish simultaneously, taking into account interdependence of the tests.
- Field significance can be tested using multivariate regression and discriminant analysis, but test is ill posed in typical climate studies.
- Proposal: project data onto a few principal components, perform field significance in reduced space, then project back to data space.
- The number of PCs is determined from cross validation experiments.
- Application to annual mean SST easily detects trend, since trend dominates first 3 PCs.
- Application to DJF 300-hPa zonal wind detects trend, even though the leading PCs have little-to-no trend.
- Application to regression between DJF T2m and NINO3.4 also detects significant regression pattern, even though individual PCs have no significant regression coefficient.
- Discriminant projection patterns allows regression pattern amplitude to be monitored in real-time.