Statistics for Spatio-Temporal Data

Introduction, Visualization, Descriptive Methods

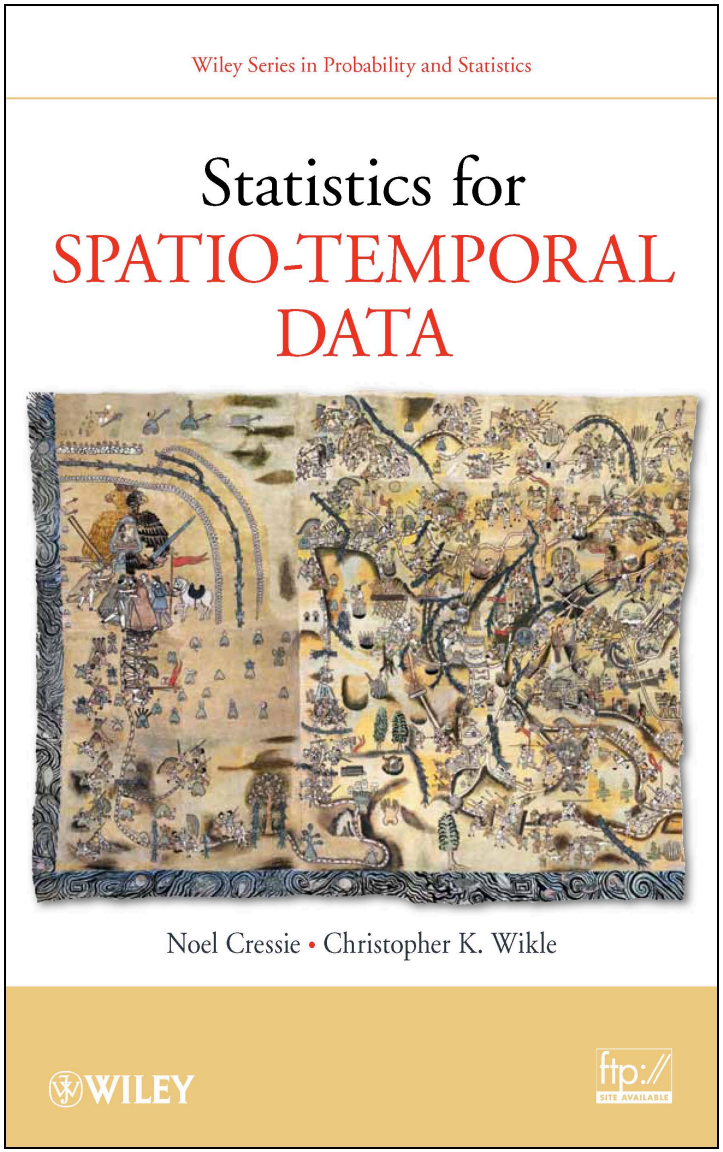
Christopher K. Wikle

University of Missouri
Department of Statistics

May 2012

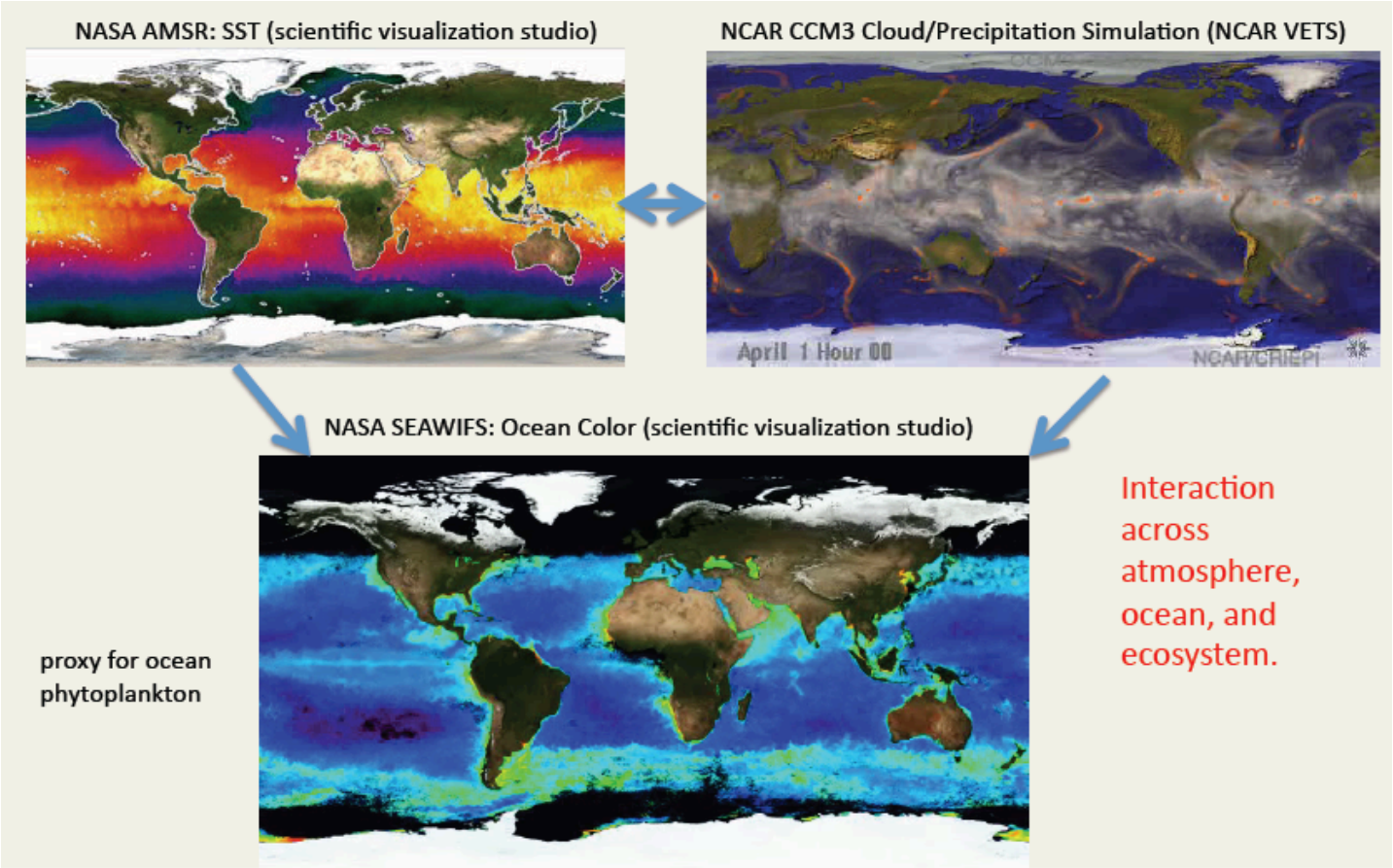
Statistics for Spatio-Temporal Data

This series of lectures is based loosely on the 2011 John Wiley & Sons book by Noel Cressie and Chris Wikle:



Spatio-Temporal Processes and Data

Data from spatio-temporal processes are common in the real-world, representing variety of interactions across processes and scales of variability.

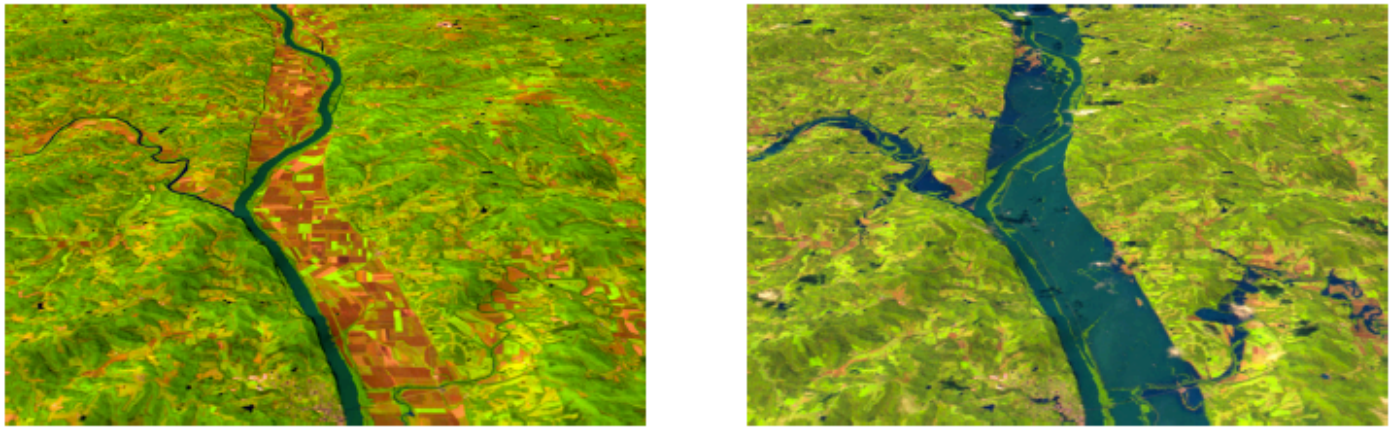


Spatio-Temporal Processes and Data

Spatio-temporal data are not new. Consider the digitally restored Lienzo de Quauhquechollan from the indigenous people of Guatemala who documented the spatio-temporal history of the Spanish conquest from 1527 to 1530.

Spatio-Temporal Processes and Data

Although it may be informative to see snapshots of spatial events in time (see the Missouri River scene below), to understand the process, we must know something about the behavior from one time-period to the next.



Images from NASA's Landsat Thematic Mapper. Each image shows a segment of the Missouri River near Hermann, MO (mile 96.5, at the bottom of the scene), and Gasconade, MO (mile 104.4, in the "V" in the middle of the scene). The river flows from west (top of the scene) to east (bottom of the scene). Left panel: September 1992, before a major flood event. Right panel: September 1993, after a record-breaking flood event in July 1993.

Goals of Spatio-Temporal Data Analysis

In statistical spatio-temporal data analysis, we seek to characterize the process in the presence of uncertain and (often) incomplete observations and system knowledge for the purposes of:

- Prediction in space (interpolation)
- Prediction in time (forecasting)
- Assimilation of observations and mechanistic models
- Inference on controlling process parameters

Spatio-Temporal Statistical Modeling

From a statistician's perspective, what makes it "statistical"?

- Uncertainty in data, model, and the associated parameters
- Estimation of parameters and prediction of processes

We also often make a distinction between "stochastic" and "statistical"

- The former concerns random structures in models
- The latter concerns estimation and prediction given data

Spatio-Temporal Statistical Modeling

Traditionally, there are two primary approaches to spatio-temporal modeling:

- Descriptive (marginal): Characterize the first and second moment behavior of the process
  - Several different processes could imply the same marginal structure; problematic if non-Gaussian
  - Most useful when process knowledge is limited
- Dynamic (conditional): Current values of the process at a location evolve from past values of the process at various locations
  - Closer to the etiology of the phenomenon under study
  - Most useful if there is *a priori* knowledge available concerning process behavior

Outline of this Short Course

- Introduction and Exploratory/Descriptive Methods for Spatio-Temporal Data
- Essential Time Series and Spatial Statistics Concepts
- Marginal Approaches for Spatio-Temporal Modeling
- Hierarchical Modeling
- Dynamic Spatio-Temporal Models
- Examples

Outline of this talk

- Notation
- Exploratory and Descriptive Methods for Spatio-Temporal Data
- Motivation for Spatio-Temporal Statistical Modeling

Notation: Spatio-Temporal Processes

Let $\{Y(\mathbf{s}; t) : \mathbf{s} \in D_s \subset \mathbb{R}^d, t \in D_t \subset \mathbb{R}\}$ denote a spatio-temporal random process, where D_s is the spatial domain of interest, D_t the temporal domain of interest, $\mathbf{s}$ is a spatial location and t a time.

When we refer to discrete time, we will typically write $Y_t(\mathbf{s})$ (i.e., a subscript t)

A purely spatial process is then: $Y(\mathbf{s})$ and a time series is either $Y(t)$ (continuous time) or Y_t (discrete time).

Notation: Distributions and Vectors/Matrices

It has become customary in hierarchical modeling to denote "distributions" using bracket notation. E.g.,

- $[Z]$ - a continuous or discrete distribution for random variable Z
- $[Z, Y]$ - a joint distribution of random variable Z and Y
- $[Z|Y]$ - a conditional distribution of Z given $Y = y$

We also typically denote vectors and matrices by a bold font: e.g., $\mathbf{Y}$, $\boldsymbol{\beta}$.

We use a prime notation to represent a vector or matrix transpose: e.g., $\mathbf{Y}'$.

Exploratory Methods for Spatio-Temporal Data

Reference: Chapter 5 of *Statistics for Spatio-Temporal Data*

For purposes of illustration, we consider primarily two datasets:

- Sea Surface Temperature (SST): Primarily a data set of 2 degree by 2 degree gridded dataset of monthly anomalies (differences from long-term averages) for the period January 1970 - December 2003.
- Eurasian Collared Dove (*Streptopelia decaocto*): Yearly counts from the North American Breeding Bird Survey from 1986 - 2003.

Exploratory Methods: Visualization

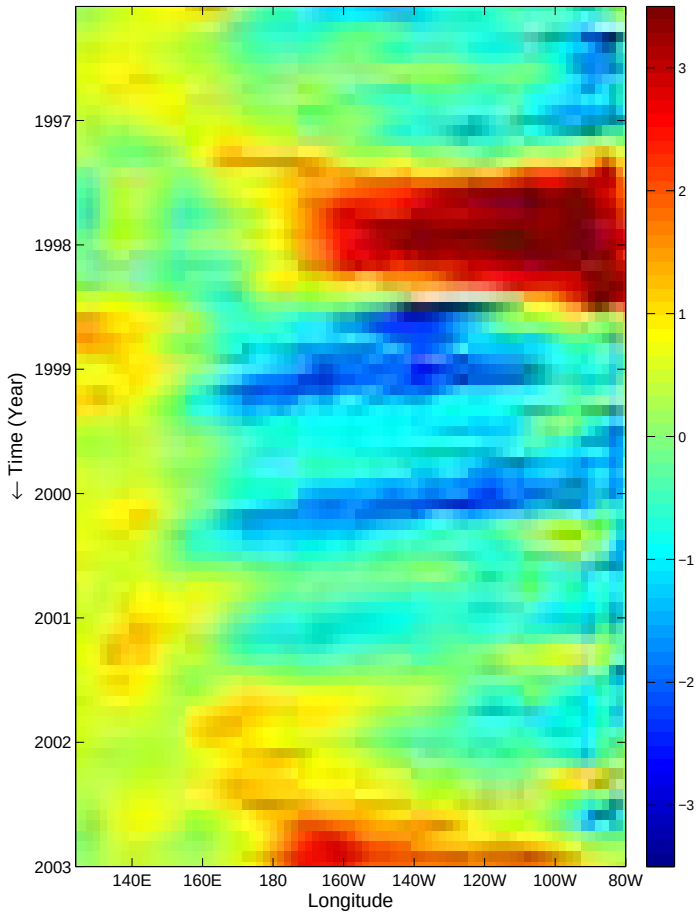
For spatio-temporal data for $D_s \subset \mathbb{R}^2$ and $D_t \subset \mathbb{R}$, we would ideally be interested in examining the evolution of the spatial data through time. An animation (or movie) can be quite useful, especially for dynamical features in the data (e.g., waves).

(sst movie)

Visualization: Marginal and Conditional Plots

Space (1-D)/Time Plots:

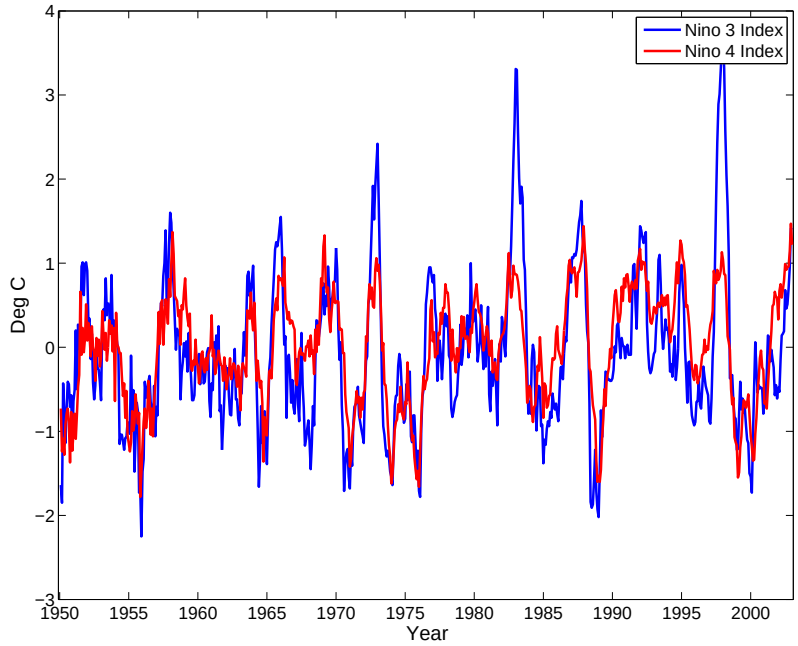
Consider the *Hovmöller diagram*, which presents a 2-D plot, but one dimension represents 1-D space and the other dimension represents time. Consider the SST anomaly data averaged between 1° S and 1° N and plotted from 130E to 80W longitude for the years 1996 - 2003.



Visualization: Marginal and Conditional Plots

Time-Series Plots:

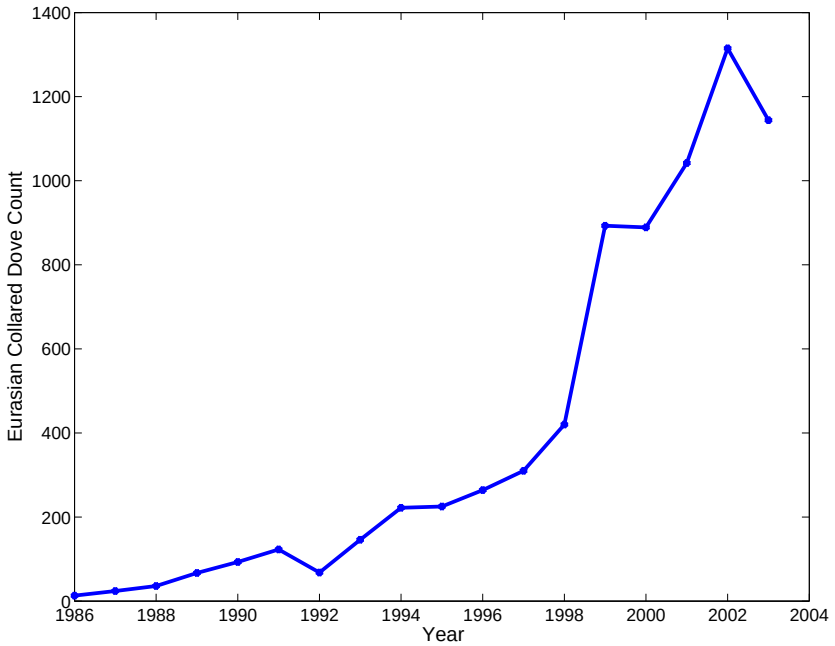
It is typically helpful to plot time series associated with various spatial locations or regions of space. Consider the SST anomaly averages over regions of the Pacific associated with the El Niño and La Niña phenomena.



Visualization: Marginal and Conditional Plots

Time-Series Plots:

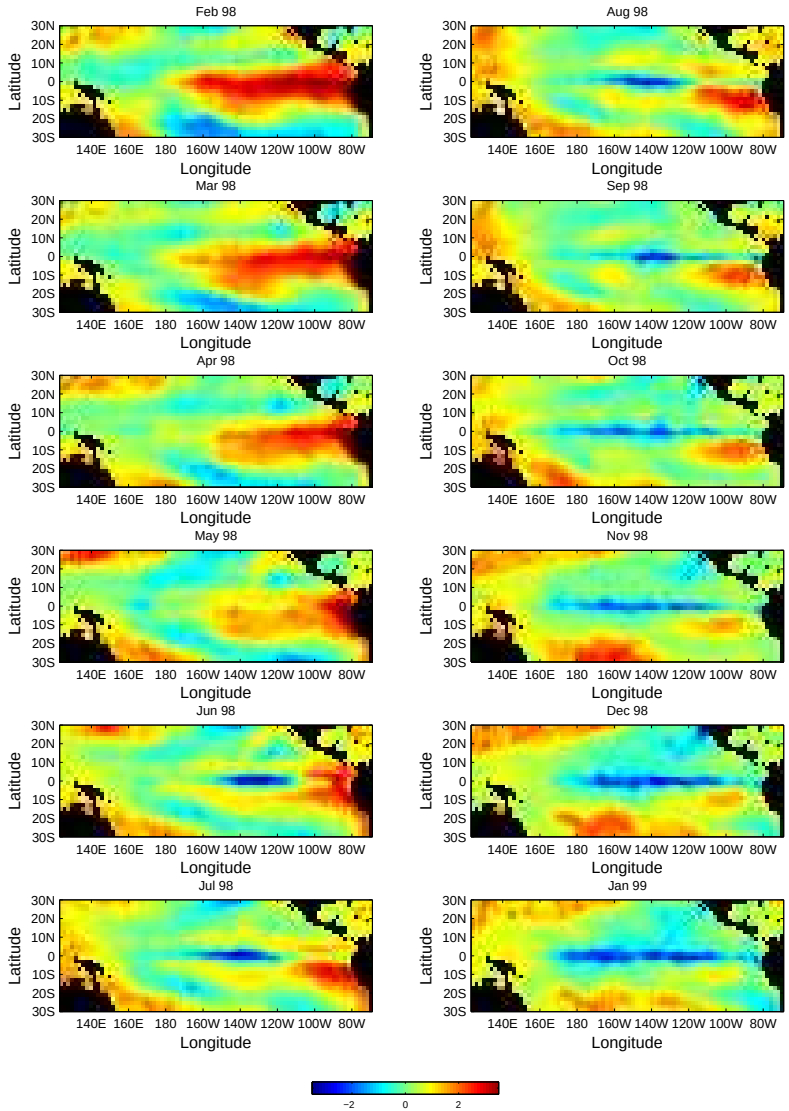
As another example, consider a plot of the invasive Eurasian Collared Dove (ECD) counts aggregated across all spatial locations and plotted for each year from 1986 - 2003.



Visualization: Marginal and Conditional Plots

Spatial Maps:

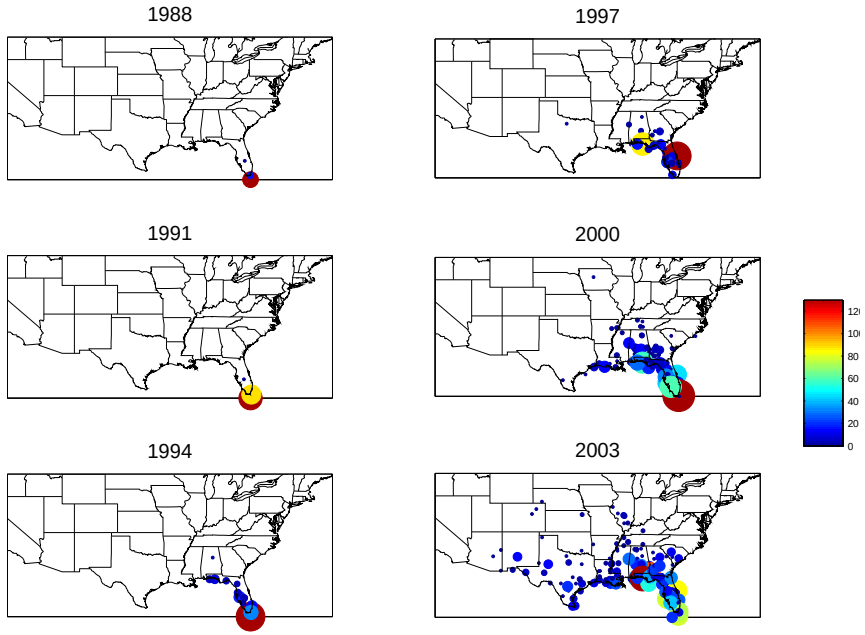
Just as it is useful to look at time series for individual spatial locations or for spatial regions, it is also useful to look at spatial maps for given times, sequences of times, or aggregates over time. Consider the monthly SST anomaly maps for February 1998 through January 1999. Each map shows 2,261 ocean pixels at $2^\circ \times 2^\circ$ resolution.



Visualization: Marginal and Conditional Plots

Spatial Maps:

It is also useful to look at a sequence of maps when the data are not on a regular grid, such as for the Eurasian Collared Dove data. Consider the 3-yearly sequence of yearly BBS sampling-route counts from 1988-2003. The ECD relative abundance is represented by both the size and color of the circles.



Empirical Covariance/Correlation

The previous visualization was concerned with the data directly. Since a key component of spatio-temporal data is the dependence between observations (in space and/or time), it can be useful to plot summaries of this dependence. Plots of empirical spatio-temporal covariance (or correlation) matrices can be informative. First, we define the empirical covariance and correlation.

Assume we have observations:

$$\mathbf{Z}_t = (Z(\mathbf{s}_1; t), \ldots, Z(\mathbf{s}_m; t))',$$

for $t = 1, \ldots, T$.

Empirical Covariance/Correlation (cont.)

An $m \times m$ empirical (averaged over time) *lag-τ spatial covariance matrix* is given by:

$$\hat{\mathbf{C}}_Z^{(\tau)} = \frac{1}{T-\tau} \sum_{t=\tau+1}^{T} (\mathbf{Z}_t - \hat{\boldsymbol{\mu}}_Z)(\mathbf{Z}_{t-\tau} - \hat{\boldsymbol{\mu}}_Z)', \ \tau = 0, 1, \ldots, T-1,$$

where the *empirical spatial mean*, $\hat{\boldsymbol{\mu}}_Z$, is given by:

$$\hat{\boldsymbol{\mu}}_Z = \frac{1}{T} \sum_{t=1}^{T} \mathbf{Z}_t.$$

The *empirical lag-τ spatial correlation matrix* is given by:

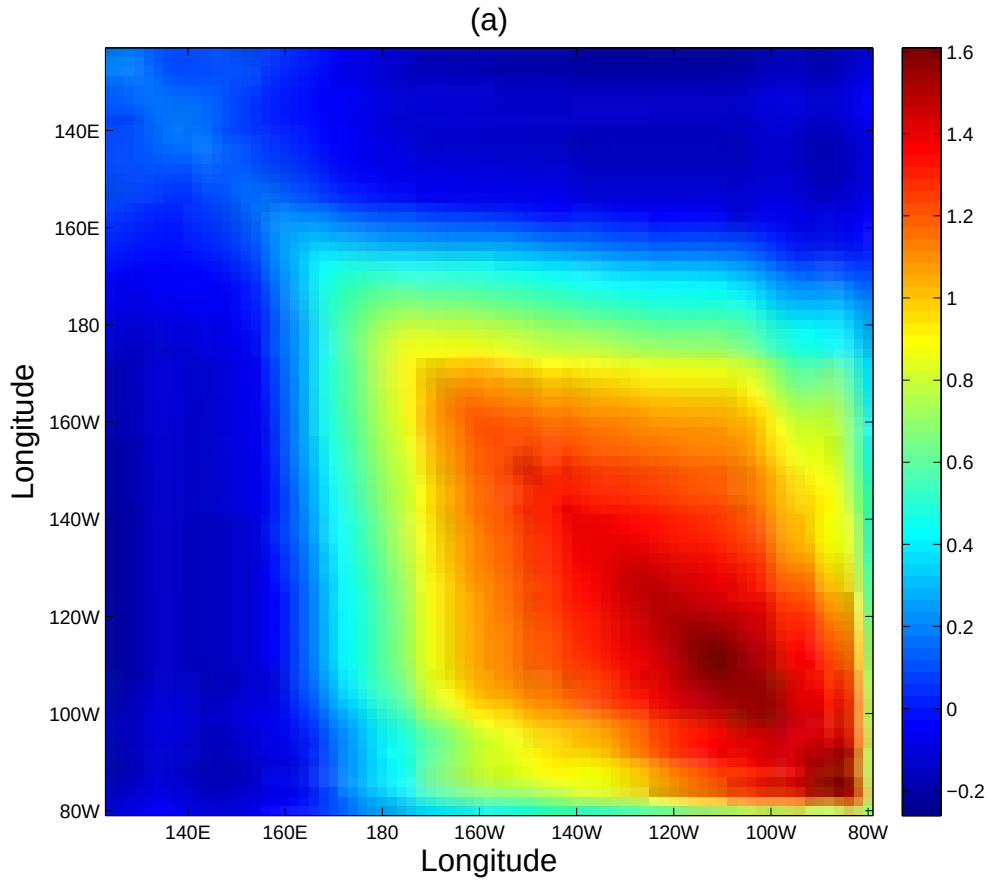
$$\hat{\mathbf{R}}_Z^{(\tau)} = \hat{\mathbf{D}}_Z^{-1/2} \hat{\mathbf{C}}_Z^{(\tau)} \hat{\mathbf{D}}_Z^{-1/2},$$

where $\hat{\mathbf{D}}_Z \equiv \text{diag}(\hat{\mathbf{C}}_Z^{(0)})$ is a diagonal matrix with the spatially indexed empirical variances on the main diagonal.

Visualization: Empirical Covariance/Correlation (cont.)

Covariance Matrix Image Plot:

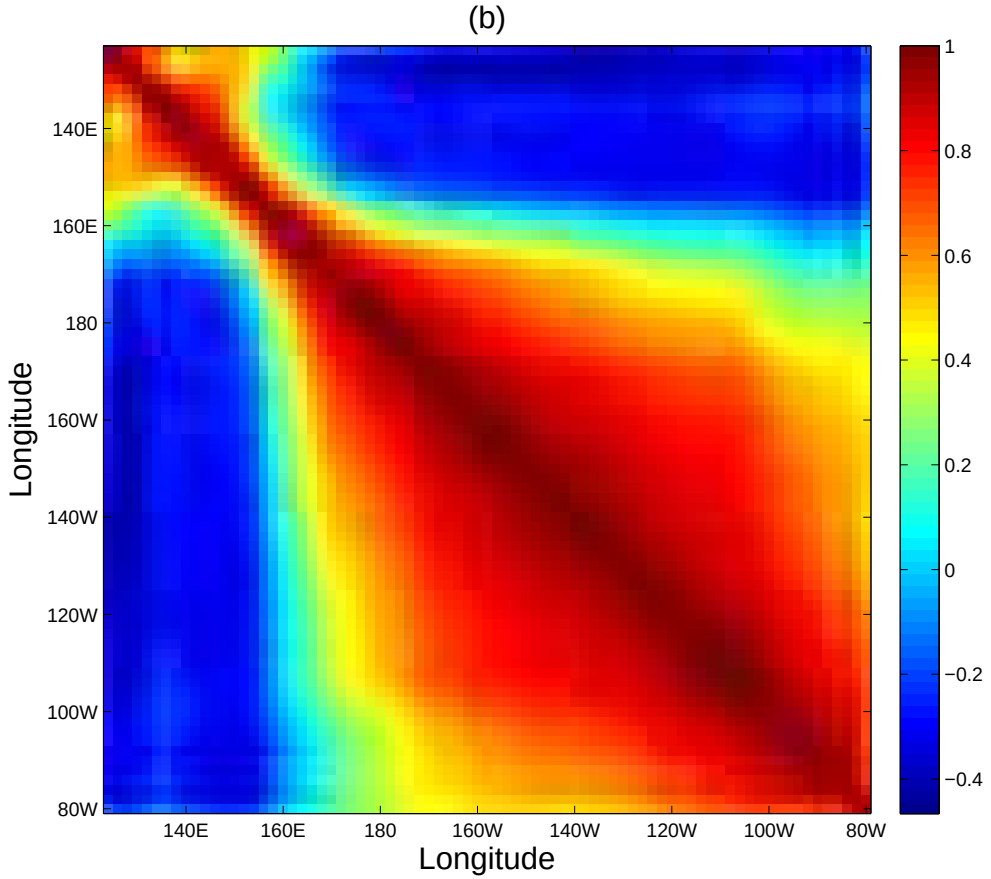
Consider the empirical lag-0 covariance matrix for SST anomalies for locations in a domain *along the equator* in the western Pacific Ocean as presented as an image plot.



Visualization: Empirical Covariance/Correlation (cont.)

Correlation Matrix Image Plot:

Now, consider an image plot of the empirical lag-0 correlation matrix for SST anomalies for locations in a domain *along the equator* in the western Pacific Ocean.



Empirical Cross-Covariance/Correlation Matrices

We may also be interested in the empirical cross-covariance or cross-correlation matrices between two data sets, $\mathbf{Z}_t$ and $\mathbf{X}_t = (X(\mathbf{x}_1; t), \ldots, X(\mathbf{x}_l : t))'$, for $t = 1, \ldots, T$, where the locations in the two data sets need not coincide (and, thus, may not be of the same dimension). An $m \times l$ *empirical lag-τ cross-covariance matrix* is given by

$$\hat{\mathbf{C}}_{Z,X}^{(\tau)} = \frac{1}{T-\tau} \sum_{t=\tau+1}^{T} (\mathbf{Z}_t - \hat{\boldsymbol{\mu}}_Z)(\mathbf{X}_{t-\tau} - \hat{\boldsymbol{\mu}}_X)',$$

where $\hat{\boldsymbol{\mu}}_X$ is defined analogously as $\hat{\boldsymbol{\mu}}_Z$. Similarly, the empirical *lag-τ cross-correlation matrix* is given by

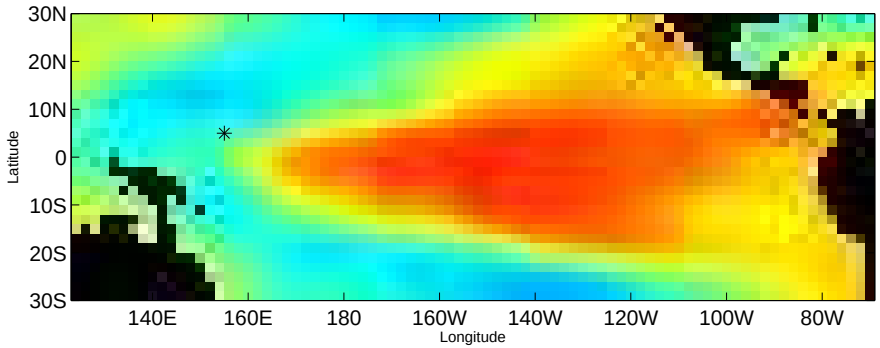
$$\hat{\mathbf{R}}_{Z,X}^{(\tau)} = \hat{\mathbf{D}}_Z^{-1/2} \hat{\mathbf{C}}_{Z,X}^{(\tau)} \hat{\mathbf{D}}_X^{-1/2},$$

where $\mathbf{D}_X = \text{diag}(\hat{\mathbf{C}}_X^{(0)})$.

Visualization: Empirical Cross-Correlation Matrix

Cross-Correlation Matrix Image Plot:

Consider the lag-6 cross-correlation for the special case of $l = 1$, where the X-variable is the near-surface zonal (i.e., east-west) wind component at 155°E, 5°N and the Z-variable is the monthly anomaly for each of the pixels in the tropical Pacific region 6 months later. In this case, since X corresponds to one location, we can plot the cross-correlation matrix as a spatial map.



Empirical Spatio-Temporal Covariance Matrix

As we will see later, the joint spatio-temporal covariance structure is critical for optimal prediction. Thus, it is important to examine the empirical covariance function at various space and time lags. Assuming that the first moment depends on space but not on time, and the second moment depends only on the spatial and temporal *lag differences*, the estimated spatio-temporal covariance at spatial lag $\mathbf{h}$ and time lag τ is:

$$\hat{C}_Z(\mathbf{h};\tau) \equiv \frac{1}{|N_s(\mathbf{h})|}\frac{1}{|N_t(\tau)|}\sum_{\mathbf{s}_i,\mathbf{s}_j \in N_s(\mathbf{h})}\sum_{t,r \in N_t(\tau)}(Z(\mathbf{s}_i;t)-\hat{\mu}_Z(\mathbf{s}_i))(Z(\mathbf{s}_j;r)-\hat{\mu}_Z(\mathbf{s}_j)),$$

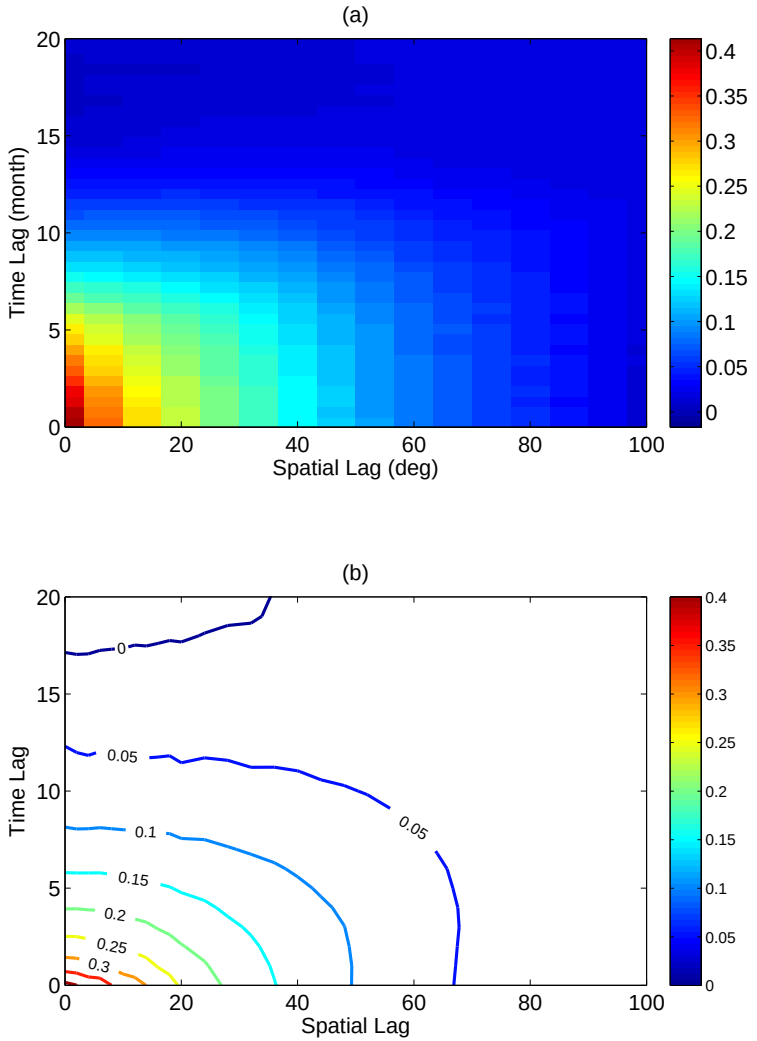
where $\hat{\mu}_Z(\mathbf{s}_i) \equiv (1/T)\sum_{t=1}^{T} Z(\mathbf{s}_i;t)$, $N_s(\mathbf{h})$ refers to the pairs of spatial locations with spatial lag within some tolerance of of $\mathbf{h}$, $N_t(\tau)$ refers to the pairs of time points with time lag within some tolerance of τ, and $|N(\cdot)|$ refers to the number of elements (cardinality) of the set $N(\cdot)$.

From this formula, we can construct the lag-τ empirical covariance matrices $\hat{\mathbf{C}}^{(\tau)}$, $\tau = 0, 1, 2, \ldots$.

Visualization: Empirical Spatio-Temporal Covariance

Spatio-Temporal Covariance Matrix Image Plot:

Consider the SST anomaly data set with spatial lags ($\mathbf{h}$) in degrees and temporal lags (τ) in months. In this case, we show both an image plot and a contour plot - both are equally informative but have individual strengths and weaknesses with respect to visualization.



Empirical Orthogonal Functions (EOFs)

As will be discussed later, dimension reduction is an integral part of spatio-temporal modeling. As in classical multivariate analysis, one of the most effective methods of dimension reduction comes from a spectral decomposition of the empirical variance/covariance matrix (i.e., principal components). Although there are analogous decompositions for continuous space/time (e.g., the Karhunen-Loéve decomposition), we focus on the discrete case here.

Specifically, if $\mathbf{Z}_t = (Z_t(\mathbf{s}_1), \ldots, Z_t(\mathbf{s}_m))'$ for $t = 1, \ldots, T$, consider the spectral decomposition of the empirical lag-0 covariance matrix:

$$\hat{\mathbf{C}}_Z^{(0)} = \boldsymbol{\Psi}\boldsymbol{\Lambda}\boldsymbol{\Psi}',$$

where $\boldsymbol{\Psi}$ is the matrix of eigenvectors and $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues.

Empirical Orthogonal Functions (cont.)

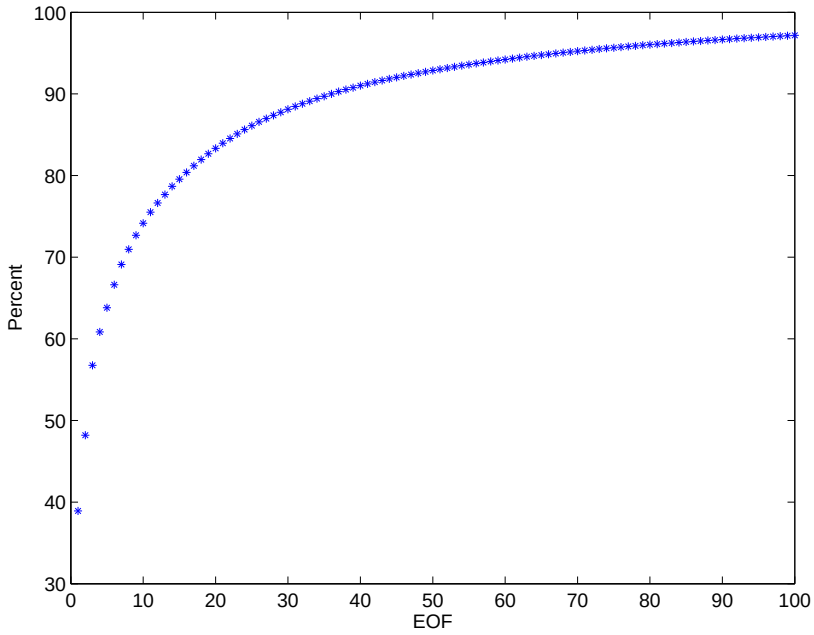
Each column of $\mathbf{\Psi}$, say $\boldsymbol{\psi}_k$, corresponds to a spatial map $\boldsymbol{\psi}_k = (\psi_k(\mathbf{s}_1), \ldots, \psi_k(\mathbf{s}_m))'$. This map, called the k-th empirical orthogonal function (EOF), is analogous to the "loadings" in traditional principal component analysis. That is, we can define new variables $a_t(k) = \boldsymbol{\psi}_k' \mathbf{Z}_t$, for $k = 1, \ldots, m$. The time series, $a_t(k)$ are then the principal component time series.

In this case, the EOF eigenvectors are orthogonal, $\mathbf{\Psi}'\mathbf{\Psi} = \mathbf{I}$ and $\boldsymbol{\psi}_1$ is the vector that allows $\text{var}(a_t(1))$ to be maximized, $\boldsymbol{\psi}_2$ is the vector that maximizes $\text{var}(a_t(2))$ subject to the orthogonality constraint, etc.

Thus, as with traditional principal component analysis, $\text{var}(a_t(k)) = \lambda_k$, $k = 1, \ldots, m$. In practice, if there is substantial spatial dependence, most of the variability in the $\mathbf{Z}_t$ data set can be expressed in terms of just a few of the principal component time series.

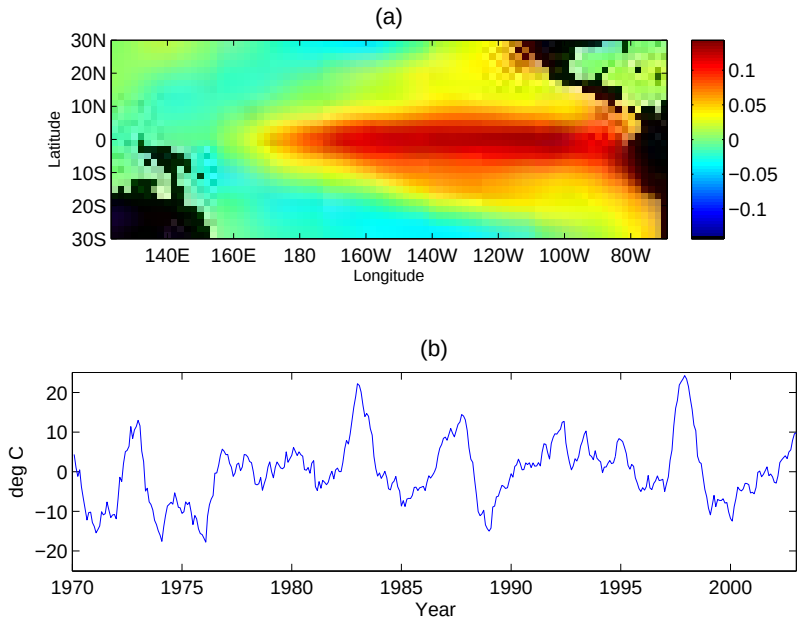
EOF Analysis

Consider a plot of the cumulative variance accounted for by the first 100 EOFs for the monthly Pacific SST anomaly data from January 1970 through December 2002.

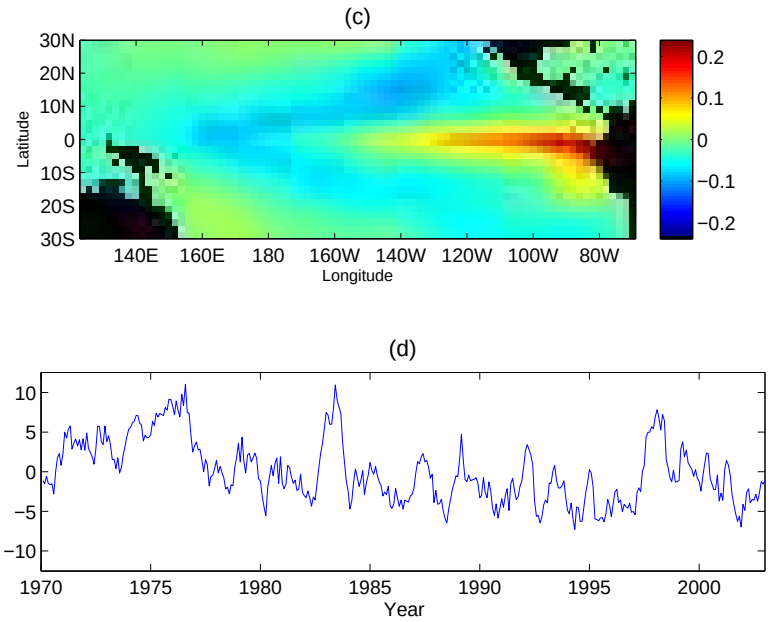


Visualization: EOFs

First EOF and PC time series for SST anomaly data (38.8% of variation):

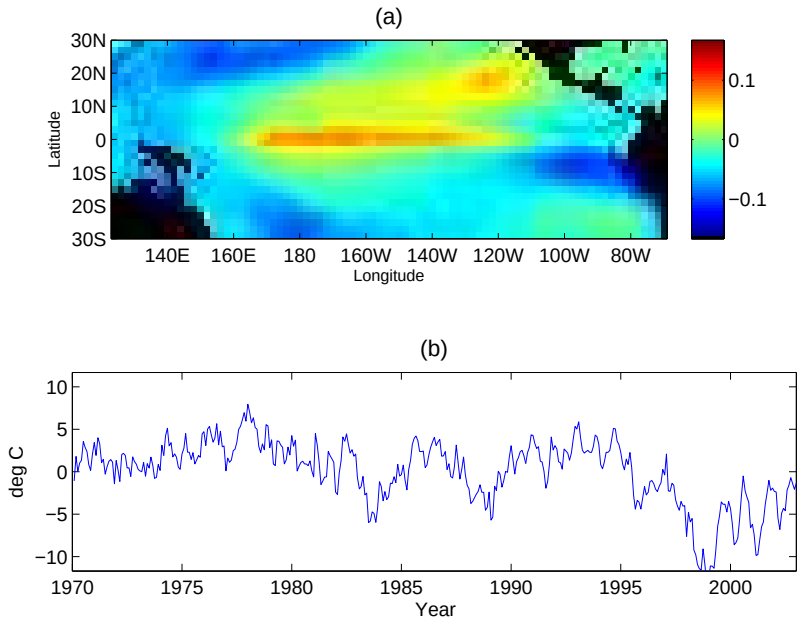


Second EOF and PC time series for SST anomaly data (9.3% of variation):

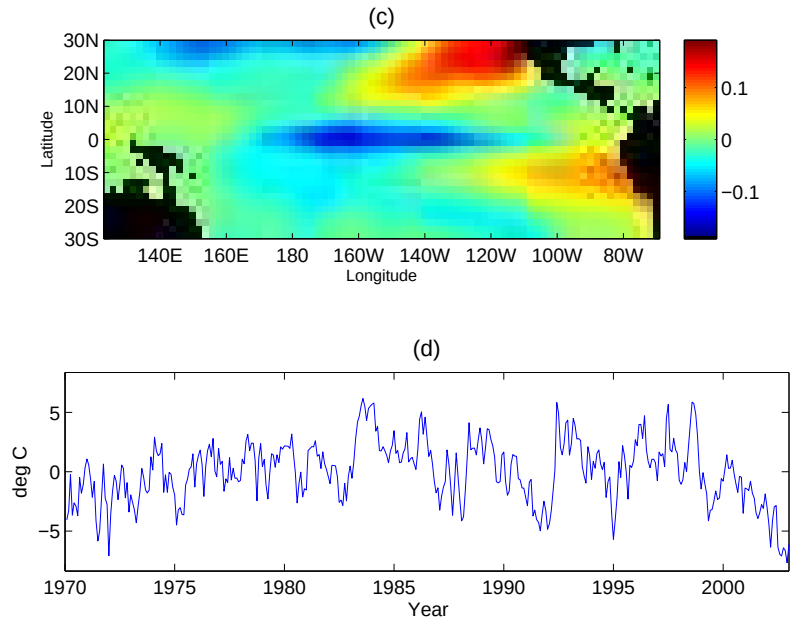


Visualization: EOFs

Third EOF and PC time series for SST anomaly data (8.5% of variation):



Fourth EOF and PC time series for SST anomaly data (4.2% of variation):



EOFs: Further Discussion

There are several additional points to make concerning EOF analysis.

- One can just as easily consider eigenvector/eigenvalue decomposition of the empirical temporal covariance matrix. In this case, the eigenvectors represent time series and the projections of the data onto the eigenvectors correspond to spatial random fields.
- Calculation of spatial EOFs when $m > T$: When the number of spatial locations exceeds the number of time replicates then the empirical spatial covariance matrix is not positive-definite. However, one can still obtain EOFs up to $T - 1$ by considering the spectral decomposition of the temporal covariance matrix and applying some simple transformations, or via a singular value decomposition on the original data matrix (see Section 5.3.4 in C&W 2011).

EOFs: Further Discussion (cont.)

- EOF analysis implicitly assumes that the spatial areas of influence are the same for each data point. If that is not the case, one should consider the Karhunen-Loéve integral equation representation of EOFs (see Section 5.3.1 in C&W 2011).
- We can also consider: multivariate EOFs, extended EOFs, complex EOFs, Hilbert EOFs, etc. (see Section 5.5 in C&W 2011).

Spatio-Temporal LISAs

Local Indicators of Spatial Association (LISAs) (Anselin, 1995) decompose global statistics into components such that:

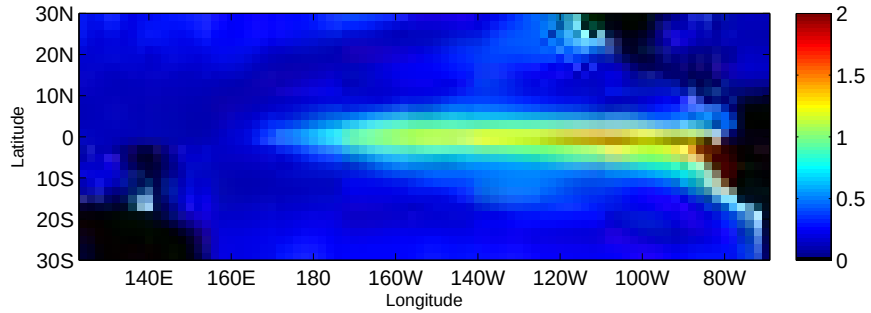
- each component is associated with a spatial coordinate
- an average of the local components yields the global component.

In a purely spatial setting, LISA applied to an empirical covariance function can show unusual components at various locations, allowing a visualization of spatial outliers or departures from stationarity.

Similar visualizations can be used for LISAs in the spatio-temporal setting. Consider the following examples.

Spatio-Temporal LISAs: Example 1

Consider the *temporal* empirical covariances for the monthly SST anomaly data from January 1970 to December 2002. The figure shows a spatial plot of these temporal empirical autocovariances at lag $\tau = 1$ (month). The El Niño region is prominent.



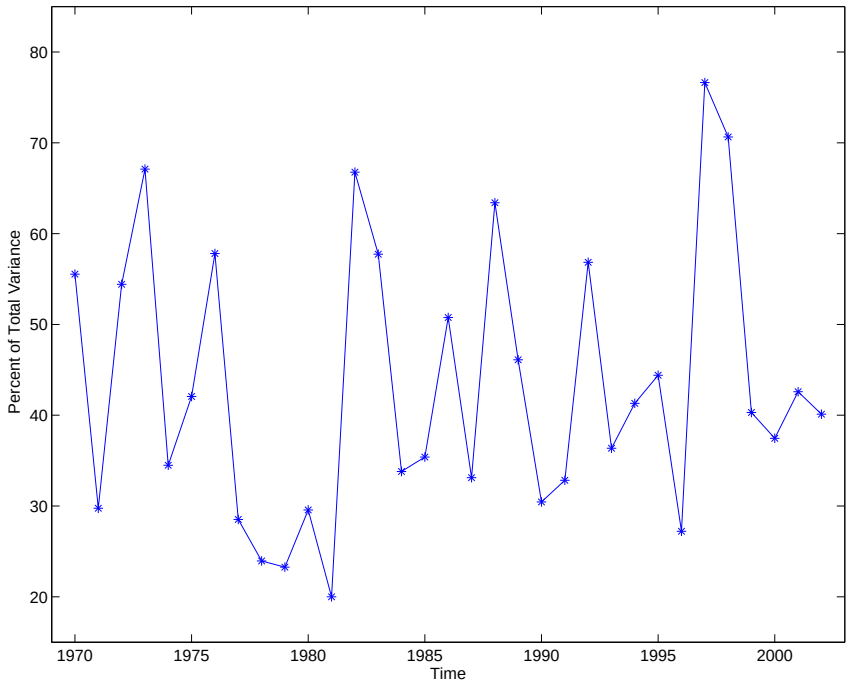
Spatio-Temporal LISAs: Example 2

Recall that the empirical (averaged over time) lag-τ spatial covariance matrix is based on the sum of component matrices $(\mathbf{Z}_t - \hat{\boldsymbol{\mu}}_z)(\mathbf{Z}_{t-\tau} - \hat{\boldsymbol{\mu}}_z)'$. Here, we consider the first EOF from the sum of these component matrices over the 12 months of each year for the period January 1970 - December 2002, which gives 33 different covariance matrices. The eigenvalue and associated eigenvector are LISAs because they approximate the empirical spatial covariance matrix from that year, and the overall empirical covariance matrix is constructed from a linear combination of these submatrices.

We consider plots of the leading eigenvalue for each year and the EOF maps for two example years.

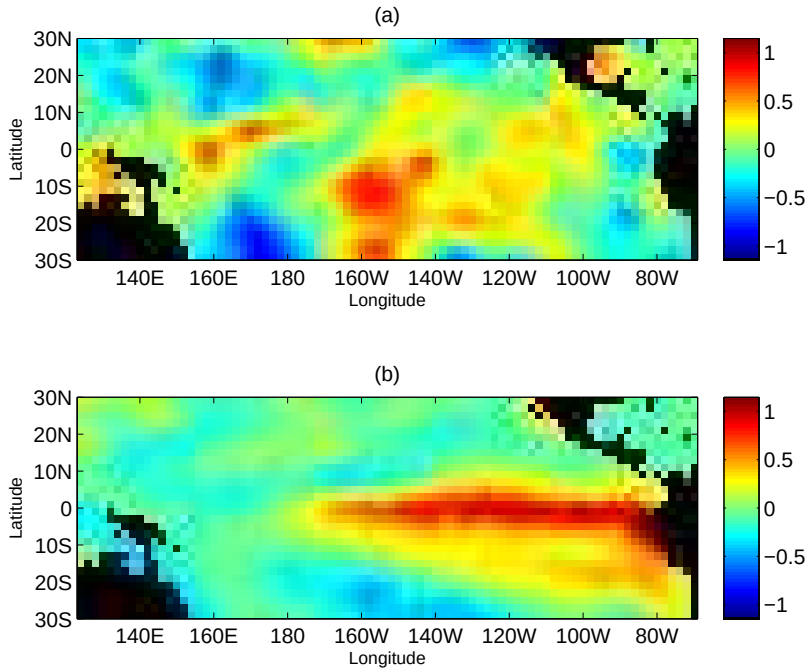
Spatio-Temporal LISAs: Example 2

LISA plot showing the eigenvalue of the leading EOF for the monthly SST anomaly data for consecutive 12-month periods starting January 1970. For presentation, the leading eigenvalue is normalized by the total variance to show the percent variance accounted for by the leading EOF for each year.



Spatio-Temporal LISAs: Example 2

The spatial map of the leading EOF of the monthly SST anomaly data associated with the yearly LISA analysis for (a) 1981 and (b) 1982. Note, 1982 was a significant El Niño year and 1981 was not.



Spatio-Temporal Parallel Coordinate Plots

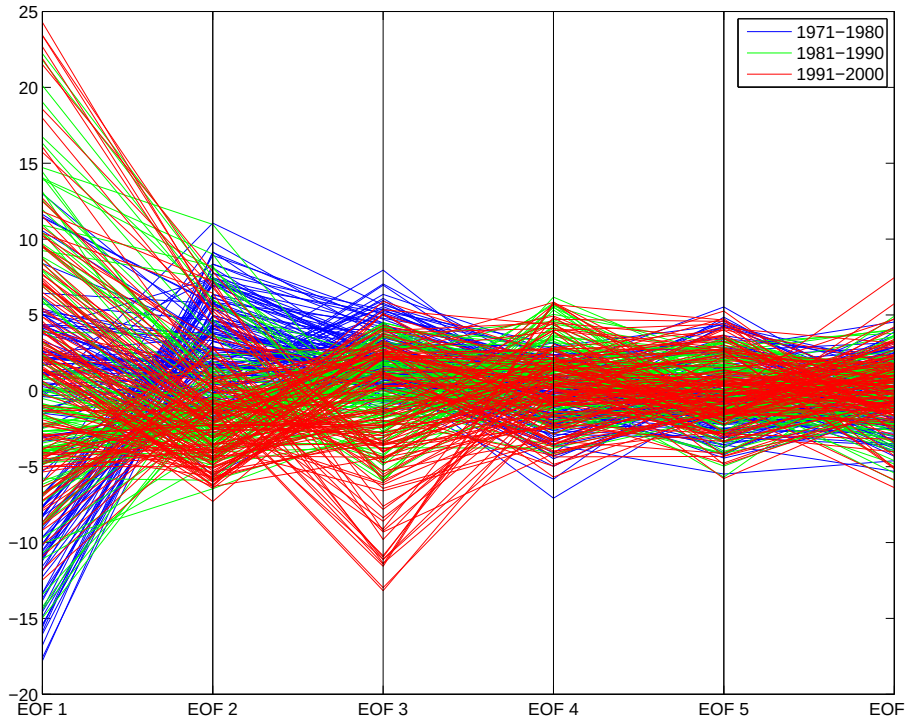
The *Parallel Coordinates* plot is a way to visualize multivariate data (e.g., Inselberg, 1985). Visually,

- the vertical axes are placed in parallel according to some predetermined order
- each component of a given multivariate observation is plotted on its respective axis
- a piecewise straight line is drawn between corresponding values on each axis
- color can be a useful way to separate groups

Note, in high-dimensional spatio-temporal data, these plots can be so "busy" that they are difficult to interpret due to the "visual clutter." One way to reduce this clutter is to collapse the m spatial locations into fewer components - such as EOFs. In this case, the vertical axis then represents a particular EOF rather than a location in physical space. Consider an example with the SST anomaly data.

Spatio-Temporal Parallel Coordinate Plot: Example

Parallel coordinate plot for the monthly Pacific SST anomalies. The data were projected onto the first six EOFs and the associated principal components represent the x-axis coordinates. Each connected line represents the values of these EOF variables for a given month (from January 1970–December 2000). Lines associated with different decades are indicated by different colors.



Spatio-Temporal Canonical Correlation Analysis (ST-CCA)

Canonical correlation analysis (CCA) is a long-standing multivariate statistical method that obtains linear combinations of two sets of data such that the correlations are maximal. This can be extended to the spatio-temporal context.

Consider two spatio-temporal datasets:

$$\mathbf{Z}_t = (Z_t(\mathbf{s}_1), \ldots, Z_t(\mathbf{s}_m))' : t-1, \ldots, T,$$

$$\mathbf{X}_t = (X_t(\mathbf{x}_1), \ldots, X_t(\mathbf{x}_l))' : t-1, \ldots, T,$$

where the spatial domain may be different, but we assume that the temporal domain is the same for both datasets.

We seek a linear combination of the data vectors $a_t(k) = \boldsymbol{\xi}_k' \mathbf{Z}_t$ and $b_t(k) = \boldsymbol{\psi}_k' \mathbf{X}_t$, where $\boldsymbol{\xi}_k = (\xi_k(1), \ldots, \xi_k(m))'$ and $\boldsymbol{\psi}_k = (\psi_k(1), \ldots, \psi_k(l))'$ are both spatial maps such that the correlation (i.e., the k-th canonical correlation, r_k^2) between $a_t(k)$ and $b_t(k)$ is maximal.

ST-CCA (cont.)

Specifically, we do this such that $a_t(1)$ and $b_t(1)$ give the maximum correlation, and then find the next set $a_t(2)$, $b_t(2)$ such that the canonical correlation is maximized subject to being uncorrelated with $a_t(1), b_t(1)$, etc.

Define $\hat{\mathbf{C}}_Z^{(0)}$, $\hat{\mathbf{C}}_X^{(0)}$ as the empirical lag-0 spatial covariance matrices for Z and X, respectively, and $\hat{\mathbf{C}}_{Z,X}^{(0)}$ as the empirical lag-0 spatial cross-covariance matrix between Z and X. Then, let $\tilde{\boldsymbol{\xi}}_k$ and $\tilde{\boldsymbol{\psi}}_k$ be the left and right singular vectors, respectively, associated with the k-th singular vector r_k^2 from the singular value decomposition of

$$(\hat{\mathbf{C}}_Z^{(0)})^{-1/2}\hat{\mathbf{C}}_{Z,X}^{(0)}(\hat{\mathbf{C}}_X^{(0)})^{-1/2}.$$

Then, it can be shown that $\hat{\boldsymbol{\xi}}_k = (\hat{\mathbf{C}}_Z^{(0)})^{-1/2}\tilde{\boldsymbol{\xi}}_k$, $\hat{\boldsymbol{\psi}}_k = (\hat{\mathbf{C}}_X^{(0)})^{-1/2}\tilde{\boldsymbol{\psi}}_k$, and $a_t(k) = \boldsymbol{\xi}_k'\mathbf{Z}_t$, $b_t(k) = \boldsymbol{\psi}_k'\mathbf{X}_t$.

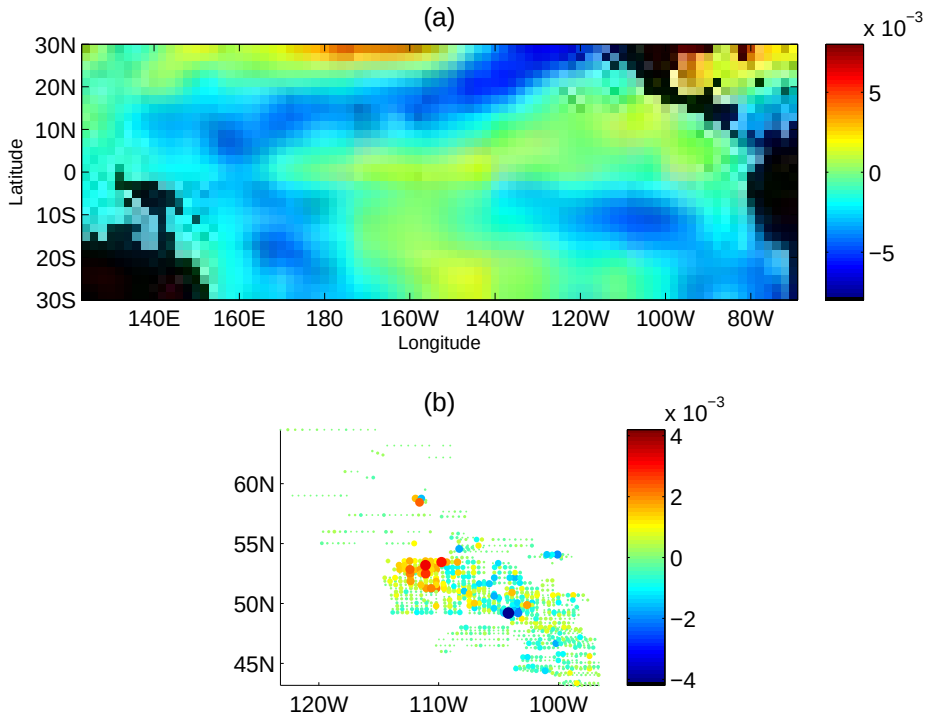
ST-CCA: Example

To illustrate the exploratory application of ST-CCA, consider the tropical Pacific SST anomalies averaged over January through March for each year form 1970 through 1999.

In addition, consider a dataset from the U.S. Fish and Wildlife Service Breeding Population Survey (BPS) for the same period. This survey has been conducted each year (since 1955) and consists of 18-mile linear segments (1/4 mile wide) over which an aircraft pilot and an observer count and speciate the waterfowl population. We are interested in Mallard duck counts. The breeding habitat is sensitive to precipitation and there is a strong link between the tropical Pacific SST and precipitation in North America. Thus, it is hypothesized that there is a possibly strong relationship between the SST anomalies and the breeding population count.

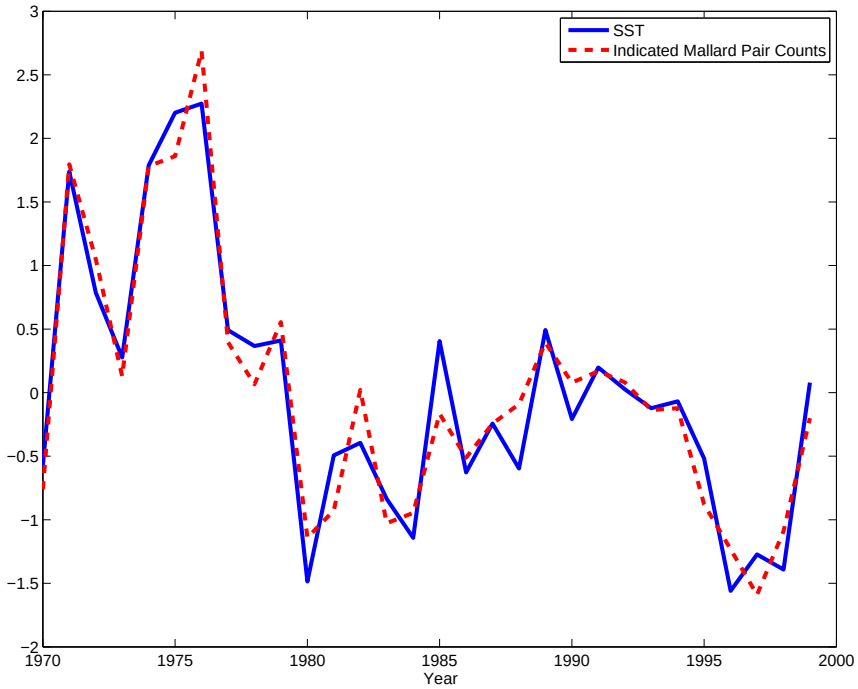
ST-CCA Example (cont.)

(a) First CCA pattern for the SST anomalies (ξ_1). (b) First CCA pattern for the Mallard pair counts (ψ_1).



ST-CCA Example (cont.)

Time series of first canonical variables for yearly SST anomalies ($a_t(1)$, blue solid line) and for yearly counts of breeding Mallard pairs ($b_t(1)$, red dashed line). Note, the correlation between these two time series is 0.96.



Other Visualizations

Note, there are many other visualization and exploratory methods for spatio-temporal data that could be considered. For example, Cressie and Wikle (2011) discuss some of these in Chapter 5:

- Spatial and Spatio-Temporal Spectral Analysis
- Spatial and Spatio-Temporal Cross-Spectral Analysis
- Principal Oscillation Pattern (POP) Analysis
- Spatio-Temporal Field Comparison